Technische Universität München – Associate Professorship of Travel Behavior
Prof. Dr.-Ing. Rolf Moeckel
Arcisstraße 21 80333 München

MASTER'S THESIS

Bicycle Crash Analysis for the City of Munich

Author:

Pawan Shambhu Singh

Mentoring:

Dr. Carlos Llorca Garcia (TUM)

MASTER'S THESIS

of

Pawan Shambhu Singh

Date of Issue: 2021-11-15

Date of Submission: 2022-05-15

Acknowledgement

First of all, I would like to thank my supervisor Dr. Carlos Llorca Garcia for his invaluable time and feedback on the work done for the thesis. Through very interesting and information rich discussions about the concerns in road safey, he helped me narrow down the topic. With consistent feedback and hints regarding the analysis process, he has been incredibly helpful.

I would also like to thank my mom and dad for their constant emotional and financial support, which helped me focus on the studies. Also my freinds Lukas, Taras, Karma and Vald for their consistent support and for engaging me in interesting discussions regarding the thesis and helped me stay motivated and cheerful.

Abstract

The thesis aims to analyze the bicycle crashes in the city of Munich, based on the data collected from opensource platforms. This included the data obtained from Untfallatlas and Geofabrik portal. Some data was also acquired from traffic data provider TomTom, which included the average travel speed data for limited sections of street for the city of Munich. The data obtained included the information regarding the crash locations, the location of infrastructure associated with road like traffic signals, road crossings, location of educational buildings etc. The dataset was constructed using QGIS, a free GIS and mapping software. Through the literature review, it was established that when analyzing accident data and count data, certain models perform better than others. This included Poisson model, Negative Binomial model and Zero-Inflated Negative Binomial (ZINB) models. Since typical count data used in crash analysis consists of significant number of zeroes, the ZINB model was suggested to work better. This was verified first by investigating the data for observing significantly high number of zeroes and then by analyzing all four models. Based on statistical tests and predicted values, it was concluded that the ZINB model does indeed perform better as compared to the other three models, due to it's ability to model the count data and zeroes independently. Two spatial models were also explored on a trial basis. The spatial lag model and spatial error models, both perform based on the spatial weights established based on understanding on the data. Both models performed good as compared to an ordinary least square model, but they did not predict the crashes accurately as compared to the ZINB model. The model suggested that streets with mixed right of way are more susceptible to bicycle crashes. Based on the predicted and analyzed values from the ZINB model, the sections of city more prone to bicycle crashes were identified as being located close to the city center and primarily include the districts of Altstadt-Lehel, and Ludwigsvorstadt-Isarvorstadt. The solution proposed for these central districts was a car free zone, which was modeled and observed to be successful in reducing crashes. Although there were some limitations to the models and the data used in analysis, they have been mentioned in respective sections.

Table of Contents

1 Introduction

With the Rapid climate change happening around the world, several governments have called for use of more environmentally friendly and sustainable modes of transport. This has led to a huge change in the use of private vehicles, Public Transport and Bicycles etc. In Germany, the use of Bicycles has seen a positive trend over the last few years. According to the Mobility in Germany study published by BMVI (Bundesministerium für Digitales und Verkehr informiert) which suggests that the usage of bicycles has seen an upward trend, meaning that the distance travelled by bicycles has increased significantly over the last few years (Follmer & Gruschwitz, 2019). Accident statistics published by the Statistisches Bundesamt (Destatis) show that although the bicycle crashes have not shown a significant increase, they have not decreased either. While there are many reasons for crashes happening and the reasons could range from a rider's lapse of judgement to something related to the environmental condition at the time of the crash, insufficient infrastructure or external influence. The investigation of these incidents could lead to an increase in safety and general satisfaction associated with the use of a bicycle.

Increasing the safety of road users is always of paramount concern for Governments and city administrations. The Nationaler Radverkehrsplan 3.0, which was adopted by the Federal Government of Germany in 2021 also focuses on the importance of Cycling. It presents various initiatives which should prioritise Cycling as a mode of Transport throughout Germany, as it helps tackle the Climate Change and plans to increase the Bicycle infrastructure to help promote these goals (Bundesministerium für Digitales und Verkehr (BMDV), 2022). Taking this into account, the work done as part of this thesis takes a look at the bicycle crashes that have happened around the city of Munich over the last five years, beginning from 2016 to 2020.

Crashes are analyzed by many researchers with different scopes, based on one's field of study or focus. While there are many ways to perform this analysis, one of the popular ways to do this is by using statistical models. Quite a lot of research has been conducted around the use of linear regression models for analyzing and predicting crashes. While it can provide a good platform for analysis in many cases, more often it fails to fully establish the relationship of crash data due to its nature. The work done as part of this thesis explores different models to analyze crashes and identify the regions with a high number of crashes. The motivation behind this work stems from the use of bicycles to travel around the city of Munich frequently. While one may draw a conclusion based on personal experience, exploring crash data can help provide differentiation between perceived and actually dangerous locations. While both can be high-risk locations, it is important to understand the reasoning behind them to provide meaningful solutions for both.

The study area comprises the city of Munich is located in the southern part of Germany and is the capital of the state of Bavaria. It has a population of 1.56 million and is the third-largest city in Germany (Landeshauptstadt München, 2022a). There are a total of

75 districts within the city boundaries. The city is popular for it's various festivals throughout the year, which attract a lot of tourists and induces a significant travel demand. There are many international universities including the Technical University of Munich and the Ludwigs Maximilian University located in the district of Maxvorstadt. While the central station and the city centre are located in the districts adjacent to this district. This means that regularly, many people travel within and across this district for various purposes. The city of Munich also has 8 subway trains and 8 suburban trains, which cater to the city and the it's surroundings (Landeshauptstadt München, 2022b).

The thesis is structured as follows:

1. Introduction: Introducing the Thesis and the study area
2. Literature Review: Here, the base for research was established. Various research papers and their conclusion is summarised in this section which was used as a basis for analysis.
3. Data Collection: The data used in the analysis are discussed here, the collection and their interpretation and how the finalized data was used to generate a dataset for analysis.
4. Analysis: This section details the work that was done as part of the data exploration and data analysis. This includes different statistical and spatial models along with validation of results.
5. Applications: Some solutions for identified regions are discussed here. A possible solution has also been modelled as part of this section.
6. Conclusion: This details the conclusion of the work done for the thesis, followed by the limitations and the future research scope.

2 Literature Review

Crash analysis and prediction are typically used to improve the safety of road users and improve the quality of driving. Several factors contribute to crashes, this could be due to driver error, insufficient or faulty infrastructure, environmental reasons, traffic density or volume etc. While no one reason can be blamed in particular for a crash happening, there is a possibility to understand the influence of all of these factors on the reasons behind a crash. This can be typically done using various methods, like deep or self-learning models or statistical models. While the field of machine learning is relatively new, statistical models have been used for a long time for crash analysis. But the analysis starts at the beginning where the data is gathered, so it is important to understand this process.

2.1 The data collection for crashes

The approach to crash analysis starts with the process of data collection. The first important thing is to classify the data for the analysis. Crash data is primarily collected by the police as incident reports, which are then used by institutions to research and propose improvements. These reports generally arise when a crash is reported, either by an observer or a party involved in the crash. The Police then document the details of the crash and this is later used for analysis. (Imprialou & Quddus, 2019) in their paper about crash data quality state that typically, the crash data collected is categorized by aspects, these are crash location, crash severity, crash time, users and vehicles involved in a crash, and crash contributing factors. Depending on the analysis, the factors can be chosen for analysis but not all five of these factors are considered equally important. The crash location is considered to be one of the most important factors for crash analysis. Depending on the chosen form of analysis, the depth of data to be used may vary which can also result from the data collection or generation process. They also talk about underreporting or misreporting of reasons for crashes during reporting, due to the complexity of the issue. These issues could lead to a lack of depth in data, meaning a lack of details associated with the crash. So, the selection of attributes or factors affecting crashes should be carefully decided for the analysis, depending on the general knowledge about the location of the analysis and the general factors that result in crashes.

(Miler et al., 2016) in their paper discuss the process of crash data collection and the accuracy of the location reported for crashes. Traffic locations need to be evaluated to better implement the resources to counteract the crashes. Their research confirmed that many inaccuracies can be introduced in the data when reporting the location of crashes like incorrect latitude and longitude. This inaccurate data cannot be discarded or used entirely reliably for the purpose of analysis. While there are several ways to deal with such inaccuracies, one effective way is to smooth the data over a small area. While the location of reported crashes is important, it is also crucial to look at the kind of crashes that are reported, especially when talking about bicycle crashes. (Shinar et al., 2018) talk about this issue in great detail. Along with a group of researchers from 17 countries

around the world, they surveyed the reporting of crashes and reveal some surprising findings. They state that the reporting of bicycle crashes is severely biased against the less severe crashes or crashes that do not involve a motorized vehicle. While they argued the definition of what constitutes a crash in different countries, Germany is one of the countries where a moving vehicle has to be involved to be reported as a motor vehicle crash. Their study concluded that in most cases, the reporting of a bicycle crash depends on the severity of the injury i: e needs to go to a hospital or first aid etc. Another factor was the type of crash i: e collision with a vehicle or falling down a bike. While falling off the bike was the most common crash type, it was the most underreported type of crash. If reported, these types of crashes and the reasons behind them could help explore the possibility of crashes happening due to infrastructure or lack thereof. Along with this, the non-severity of crashes was also one of the biggest reasons for underreporting. Such underreporting could be the reason for the existence of additional zeroes in the dataset and can be called locations with no crash “reported”, which gets mixed in with locations where no crashes were “observed”. (Medury et al., 2019) conducted a study regarding reporting bicycle crashes involving other bicycles and pedestrians. This was done based on open-source data gathered from surveying people in the study area and the official reported crash record. Their study revealed that a significant number of crashes went underreported which involved bicycles and pedestrians. Although in most cases there were no physical injuries reported, such underreporting hinders the improvement of safety on streets for the vulnerable users. This study confirms the existence of bias in reporting crashes. While the existence of zeroes in a crash analysis for larger areas like a city or district will result in most locations having no observed crashes for a time certain period, such underreporting can increase this value in generating a false sense of safety in the study area.

2.2 Use of custom grid in crash analysis

The smoothening of data can be done using several methods, one of these is to use a spatial grid of fixed cell size. Although there exists a notion that the relationship between the number of crashes and a grid is difficult to model, the study by (Kim et al., 2006) presents some interesting findings regarding crash analysis using grids. They conducted a study of crashes over the county of Honolulu, Hawaii wherein they used grids to rank zones based on the density of crashes. A grid analysis can help establish zones which can be useful for more focused planning and measures. (Cai et al., 2017) also conducted a similar analysis where they proposed using grids of different sizes for crash analysis and testing different sized grids for analysis. Their analysis suggested that the smaller grid zones also known as Traffic analysis districts (TAD) perform better as compared to bigger statewide traffic analysis zones. This also helps reign in the inconsistency between different attributes used for analysis. Their study concluded that the TADs offered best fit types when comparing to other methodologies for zone crash analysis. Such a grid structure is also useful when constructing count model datasets, as they provide the flexibility to count attributes in a study area with same consistency.

2.3 Models for crash analysis

When it comes to statistical analysis, several models can be employed in data analysis. One of the most common models for such is the linear regression model. While a regression model can be ideal to plot linear relationships between independent variables in a model, it can often lead to inconclusive results when the data does not have a linear relationship. While there exists a possibility to manipulate the data before employing a linear regression model (Abdel-Salam et al., 2008), it may change the relationship between the variables which is crucial for crash data analysis. Several other statistical models can be used for the analysis of crashes, typically for crash data Poisson or Negative binomial models are used due to the presence of discrete and non-negative values in the crash data (Shankar et al., 1997) (Hadi et al., 1995). Poisson models, while useful operate under an assumption that the variance should be equal to the mean. But if the data does not hold this condition, there must be an existence of under-dispersion or over-dispersion. This can result in a varied standard error and generate senseless output (Chiou & Fu, 2013). One of the drawbacks of a standard Poisson model is that it does not provide any flexibility to accommodate this over or under-dispersion that is observed in the data. (Park & Lord, 2009) in their study and analysis, different mixtures of models were used to analyse crashes and predict them. They suggest that a mixture of Poisson or Negative binomial model are more suitable to use in analysis when the data is generated from a heterogenous set. Since crash data typically consist of over dispersion, a Negative binomial (NB) model is more suited for analyzing crashes, since the NB model can assess the different crash processes and generate more reliable results. Their analysis suggested that for such data with dispersion, a mix of the model must be considered to overcome this problem.

The reason for the existence of this dispersion is generally related to the existence of extra zeroes, that result from the data generation process for count models for crash analysis. (Lambert, 1992) suggests that this excess number of zeroes can be addressed by a two-step regression model like a Zero-Inflated regression model. They explain how the zero-inflated model works under the assumption that the existence of zeros is due to two different reasons. (Garay et al., 2011) in their paper, take a look at the Zero-inflated Poisson (ZIP) and Zero-inflated Negative binomial models (ZINB), and compare them based on model estimates like AIC. They conclude that for data with an excess number of zeroes, a ZINB model shows a better fit than a ZIP model if there are zeros in the data generated through two different processes. (Hauer, 2001) also suggest modelling crash count data using the ZINB model, as they address the distribution of zeroes much better. While there are people, who suggest that ZINB models can lead to incorrect model outputs in the sense that the model assumes that there are places which can result in a crash always or never having a crash (Lord et al., 2005). They further suggest that a good fit should not be the prime factor in selecting one model over the other and clarify that the models may not be best suited for highway entities (Lord et al., 2007).

On the other hand, (Pew et al., 2020) suggest that dismissal of the ZINB model solely based on theory is not the best decision. Since the Negative Binomial model is not restricted by the condition of the mean being equal to variance, the Zero-inflated negative

binomial model addresses the overdispersion in data which is not taken into account in the case of a Zero-inflated Poisson model. The authors compared three different models for the same crash data, and based on different statistical tests confirmed that the ZINB model can help in identifying the crash hotspots better as compared to the ZIP. At the end of their research, they conclude that the probability function of a zero-inflated random variable, which assumes a nonzero probability for positive integers, so the model cannot assign or suggest a location to be inherently safe. So, the ZINB models can be reliably used for crash analysis and prediction, given that they are evaluated against other models.

(Washington et al., 2020) in their book also give extensive details regarding the ZINB model and its workings. They also talk about the probability of an event not happening based on two conditions, where an event was not observed or the inability for the event to occur. They state that the biggest disposition for a zero state is the preponderance of zeroes in the data, which are typically unexpected in a poison model. While there is always a possibility that overdispersion will include excess zeros, it must be determined whether excess zeroes arise from true over dispersion or from an underlying process. One of the ways to address this issue is to model the data in both NB and ZINB states and run a statistical test. This will ensure that the selection of the model is based on correct parameters like the existence of two processes in data generation. They also state the relevance of using Vuong's test for selecting the better model between a Negative Binomial model or a Zero-inflated Negative Binomial model. (Vuong, 1989) suggested using a likelihood ratio-based test for selecting the optimum model, especially for zero-inflated models, along with another criterion like AIC. Although, there has been some criticism regarding the use of the Vuong test for testing the Zero-inflated models, (Wilson, 2015) states that the Vuong test is applicable for the nested models only and states that the test cannot be applied to models like ZINB, as they are fitted using a link function, which does not fit with the assumptions of Vuong test. So, using a Vuong test to verify the model is not the best solution. One way to test if using ZINB is better than the NB model is to calculate a chi-squared test statistic. (McHugh, 2013) states that for large datasets, this test can be significantly useful and can be performed when the sample sizes are unequal and more importantly the distribution of data shows a certain level of skew. This test statistic can be used to reject the null hypothesis, in this case, the NB model (Algeri et al., 2020). This can be done manually using the likelihood ratios obtained from the model summary and the degrees of freedom, usually depending on the independent parameters used in the analysis.

2.4 Spatial Regression

While crash prediction is a complicated subject in the sense that modelling all the factors responsible for a crash at the same time would result in a very large and complicated model. One of the interesting models in crash prediction is using a spatial regression

model. Spatial regression can allow an understanding of the relationship between neighbours or different observations in a specific region (Anselin, 1988) (LeSage, 2005). More commonly it is observed that when data is collected in a definite study region, the points are found to be spatially dependent, this suggests that the observations close to each other show some sort of similarity. (LeSage, 2008) explains the process of spatial regression in great detail and suggested that the traditional assumption of independent observations has some sort of spatial relationship with the elements of the dataset. In their study, they establish the relationship between the commuting times and the effect on them based on the neighbouring counties. To establish this relationship, they suggest using spatial weights, which can account for the relationship between an element and its neighbours, based on the individual parameters or properties. They discuss two kinds of sptial models, namely the Spatial autoregressive (SAR) model and the Spatial error model (SEM). They conclude by suggesting the superiority of the Spatial regression model over the traditional regression model, owing to the use of spatial dependence between observations.

(Rhee et al., 2016) analyzed the traffic crashes in the city of Seoul using Spatial regression models. While, different kinds of data have been used for spatial regression, which includes area characteristics or data describing driver behaviour. Their study involved the use of traffic analysis zones (TAZ), but opted to use a nonstandard TAZ and defined their own, which they concluded yielded better results for safety analysis. The data used in the analysis involved demographic data, income, age, gender, socioeconomic and road data. They concluded that while the use of TAZ is ideal for road safety analysis, it adds a certain level of complexity when it comes to analyzing mixed land use and it is difficult to segregate commercial and residential zones in such cases. While, the selection of TAZ cannot be uniform for all models or analyses, as the study area and the depth of data change, the TAZ should be adjusted accordingly. They also compared the results between the OLS and the spatial models, concluding that the Ordinary Least Square model fails to account for the over or under-dispersion in data, which is balanced by spatial regression models. Their tests revealed that the spatial error model performs better when compared to the spatial lag model. While the performance of the spatial model highly depends on the attributes provided to the model, which should be rich and diverse in their nature and explain more details about the study area. On the other hand, (AL-Hasani et al., 2019) compared the SAR and SEM in their study of crashes in Oman. They compared the results for both of the spatial models with the Ordinary least square model (OLS). They concluded that the Spatial lag model performs better than the Spatial Error model when tested in various statistics like AIC and loglikelihood.

(Jia et al., 2018) conducted spatial regression analysis for an administrative country in China. They also compared the results from OLS to a spatial lag and spatial error model. While their study revealed that the spatial error model performs better than the lag model, both models suggest the existence of spatial correlation. They also identify the lack of detailed data in analysis for complications in spatial crash analysis. (Gao et al., 2006) in their analysis of spatial models explain the importance of having more data should improve the quality of the model. They explain that if a spatial model performs significantly well against a regression model, it confirms the existence of spatial relationships within

the data. (Anselin, 2002) also explained the lack of clarity while selecting the specification for spatial weights and suggests the selection based on the better model output.

3 Data Collection

All the data used for the analysis was gathered from open-source platforms. Some of the data gathered includes:

1. Analysis Grid
2. Crash data for the city of Munich for the years 2016-2020
3. Traffic Network and travel speed data for the city of Munich
4. OpenStreetMap data for the City of Munich

All the data was gathered from open-source platforms and processed using QGIS and R-studio. Their information is explained below.

3.1 Crash Data

The crash data was obtained from Untfallatlas for the years 2016 through to 2020 (Statistische Ämter des Bundes und der Länder, 2022). The crash data obtained for each year contained shape files and database files along with all the information needed to project the crashes in mapping software, namely QGIS. The shape file contained information on individual crashes all across Germany, their date, time, the kind of crash, the vehicles involved in crashes and the X-Y coordinates. Since the crash data included data for all of Germany, the first steps included sorting out the Crashes for the city of Munich. The crashes were first sorted for the individual states with the categorical attribute “ULAND”, which for the state of Bavaria was “09”. This sorts all the crashes for the state of Bavaria out of the whole crash dataset for the whole country. The next step involves sorting the crashes for the government districts, with the category “UREGBEZ”, meaning “Regierungsbezirk” which means the Government districts. For the region of Munich, the code used was “01”. Finally, the crashes for the district of Munich are sorted, using the categorical attribute “UKREIS”. The district of Munich uses the code “62”. After processing the data through these steps, the only crashes left are for the district of Munich, which is the area of interest for this thesis. The next step includes sorting out all the crashes that involve a bicycle. This was done by using the attribute “IstRad” which signifies that there was a crash that involved a bicycle. The “IstRad” column consists of “1” and “0”, where 1 signifies that a crash happened involving a bicycle and 0 the crash which did not involve a bicycle. Using the Attribute selector, for the column “IstRad”, once all of these crashes are sorted for only bicycles for one year the same process was applied to the dataset for all the years, from 2016 to 2020. All the annotations and the metadata can be obtained from the Geofabrik portal (Statistische Ämter des Bundes und der Länder, 2022). A sample of the sorted data for the year 2016 is shown in the table below:

Attribute	Value				
OBJECTID	60859	61160	61566	62079	62342
ULAND	9	9	9	9	9
UREGBEZ	1	1	1	1	1
UKREIS	62	62	62	62	62
UGEMEINDE	0	0	0	0	0
UJAHR	2016	2016	2016	2016	2016
UMONAT	1	1	1	1	1
USTUNDE	15	14	5	15	7
UWOCHENTAG	1	5	4	4	6
UKATEGORIE	2	2	3	3	3
UART	3	5	5	5	5
UTYP1	6	2	2	2	3
ULICHTVERH	0	0	2	0	1
IstStrasse	1	0	1	0	1
IstRad	1	1	1	1	1
IstPKW	0	1	1	1	1
IstFuss	0	0	0	0	0
IstKrad	0	0	0	0	0
IstGkfz	0	0	0	0	0
IstSonstig	0	0	0	0	0
LINREFX	688689.2	684357.2	691762.5	683848.2	681146.4
LINREFY	5335742	5332118	5331942	5335423	5335922
XGCSWGS84	11.5368	11.47705	11.57636	11.47164	11.43557
YGCSWGS84	48.14692	48.11561	48.11184	48.14547	48.15073

Table 3.1 Attribute table for original crash data

3.2 Analysis Grid

To analyze the data, instead of using the predefined city zones, a grid covering the city of Munich was used. The district zones of the city of Munich are too big to be critical and they event out the crash data, preventing having to look at more critical locations and understanding whether the whole district has a high number of crashes or just a small region in the district. It does not help in addressing the crash hotspots. The size selection of the grid blocks was done on a trial-and-error basis. Three grid cell sizes that were investigated are 250 m X 250 m, 500 m X 500 m and 1000 m X 1000 m. While the 250 m X 250 m cell-sized grid block resulted in a very high computation time in QGIS for different functions. For example, for the function “join attributes by location”, the computation time was around 3504.8 seconds, which is approximately 60 minutes and for the other two, it was relatively low. This high computation time is due to the several attributes being closely located to one another, which increases the processing time for several functions of the GIS software. A 1000 m X 1000 m cell-sized grid smoothed the crash data and resulted in a very sparse grid, an example seen below in Figure 3.1. Both Karsplatz and Sendling Tor become part of the same grid cell, but since they are both

locations with a high number of crashes, it smoothed out the data. So, to be more concise, a grid was used with each grid cell of a fixed size of 500 m x 500 m, which addresses both the issues face by a 250 m X 250 m and 1000 m X 1000 m grid. An example of the grid cell of size 1000 m X 1000 m is shown below, showing the cell covering a significantly big area.

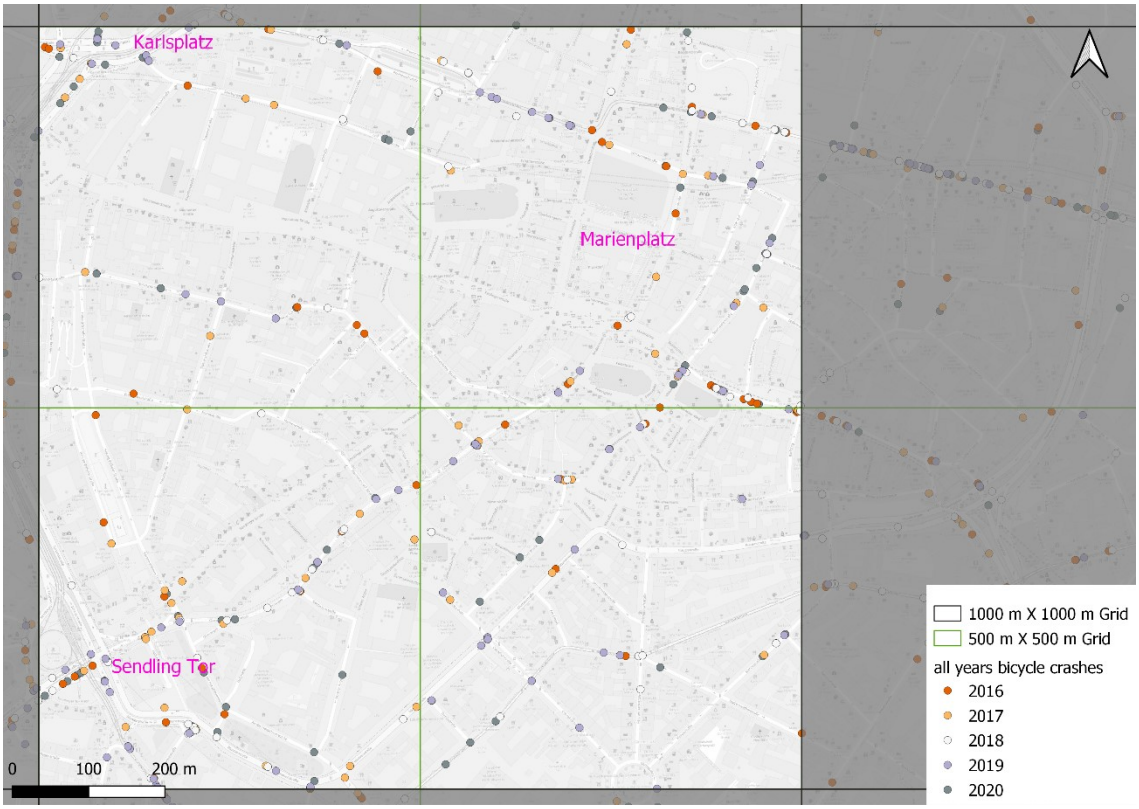


Figure 3.1 One block of 1000 X 1000 m grid

The map below shows the crash data together for all five years.

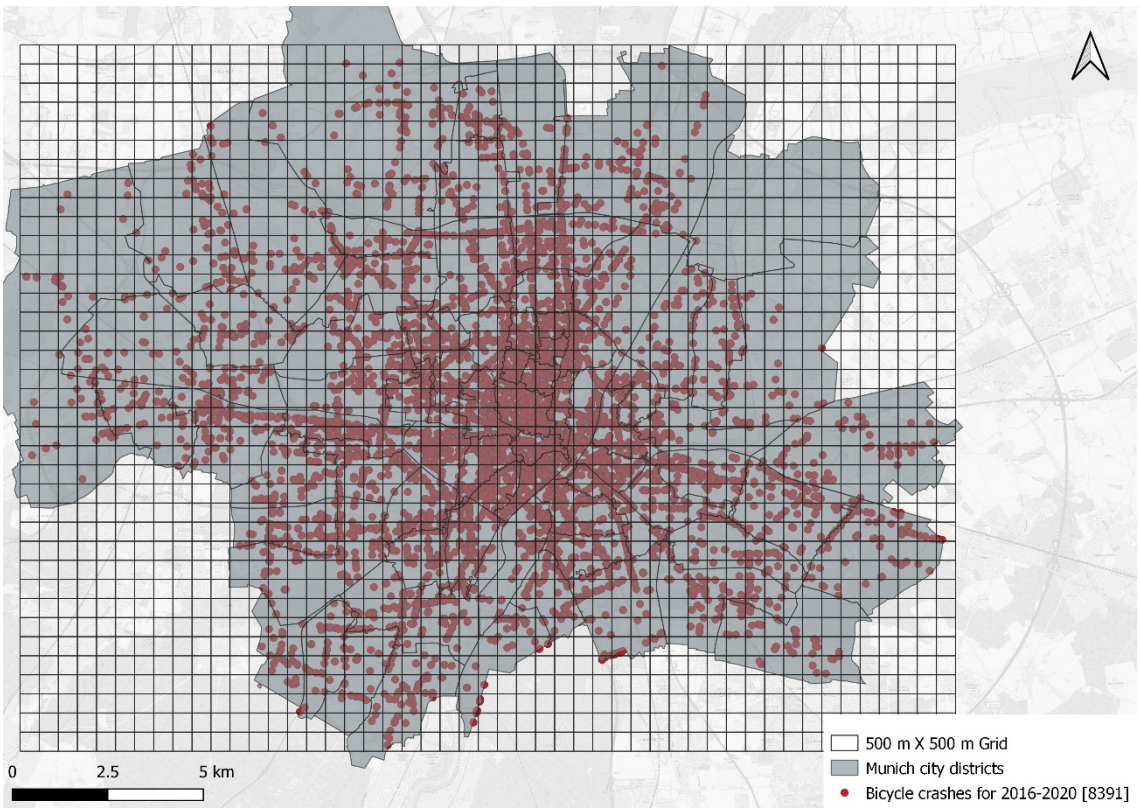


Figure 3.2 Overview of Grid and crash locations from 2016-2020

As seen in the map above, the grid covers the entire district of Munich. The grid size is selected in a way that all the crash points are covered by the grid. The bicycle crash points from the years 2016 to 2020 are plotted together and the grid size is fixed in consideration of the same. This ensures that the grid selected does not exclude any crashes. Attempts were made to mask the grid to fit the size of the Munich district map's borders, but this was unsuccessful. As both the layers are not the same, i:e the Munich district map is a raster layer and the grid is a vector layer. The grid is placed in such a way that the approximate centres of the grid and the city match together, which will make for easy interpretation of the grid in later sections.

3.3 Open Street Map data for the city of Munich

The Open Street Map data was obtained from the open-source platform Geofabrik (GEOFABRIK, 2022). The shapefiles provided by Geofabrik consist of OpenStreetMap data for the region of Oberbayern. This data is categorized into three types, point features, line features and polygon features, but polygon features are not of any interest to this thesis.

3.3.1 Point Features

The point features contain the data with six further categories each represented by its shapefile, which are explained using code, “fclass”, description and OSM tag. The attribute “fclass” was used to classify the different kinds of attributes and the code is the unique ID for each class used to further identify each class. The six categories are as follows:

1. Places: This shape file includes the location of cities, towns and villages etc, and they’re marked at the approximate centre of the cities. The different locations are differentiated using the “fclass” attribute.
2. Points of Interest: This shapefile includes different points of interest which are categorized as Public, Health, Leisure, Catering, Accommodation, Shopping, Money, Tourism and Miscellaneous. It includes information regarding locations like the police station, school, university, cage, hospitals etc. which were identified using the “fclass” attribute.
3. Places of Worship: As the name suggests, the data included in this layer consist of all the different kinds of places of Worship, with categories like Christian_catholic, Christian_baptist, Jewish, Sikh etc which can be identified with the “fclass” attribute.
4. Natural: This shape file consists location of all the natural things like trees, lakes, etc.
5. Traffic: This shape file consists of both point and area features. Traffic signal stops and crossing are some of the point features.
6. Transport: This shapefile also consists of both area and point features. Transport stops including train stops, tram stops, bus stops and other public transportation-related halts are included in this shape file.

The point data used for the analysis include the counts for supermarkets, beer gardens, Educational Institutes (including schools, Colleges and Universities), motorway junctions, rail stops (including Subway stops, S-Bahn stops and other rail stops), crossings, bus stops, tram stops and traffic signals (OpenStreetMap Wiki, 2022). The names of the following attributes used in the analysis are “signal” for traffic signals, “tram” for tram stops, “busstop” for bus stops, “railstop” for all the rail stops including S-Bahn stops, ubahn stops and main rail stops, “crossing” for all the road crossings, “junction” for motorway junctions, “education” for all the education-related buildings, which includes schools, universities etc., “biergarten” for beer gardens, and “supermarkt” for all the supermarket locations.

3.3.2 Line Features

The Line features contain several transport-related line features such as roads and railways. Line features further have three categories. Roads and Paths, Railways, trams and Cable Cars and lastly, waterways. Although for the purpose of this thesis, waterway features are not used.

1. Roads and Railways: The Roads shape file consists of all the road routes, categorized using “fclass” attribute as a motorway, trunk, primary, secondary and tertiary. There are also minor roads such as residential, pedestrian, living streets etc. There are further routes from non-motor vehicle streets like cycleways, footpaths and bridleways also available.
2. Railways: The railway shapefile consists of information regarding all the rail routes. These include light rail, regular rail tracks, subways and trams etc.

The Line features used in the analysis include the length of bicycle paths, primary streets, secondary streets, tertiary streets, residential streets, footways and unspecified paths. The Primary streets are mostly the roads that are classified as national roads, the secondary roads are the regional roads, the tertiary streets are the roads local to the region, residential streets are the streets in the residential areas, and bicycle paths are the paths designated for cycling, footway is the footpaths for pedestrians, and the unspecified paths (OpenStreetMap Wiki, 2022). The names of the following attributes used in the analysis are “bicycle” for bicycle paths, “primary” for primary streets, “secondary” for secondary streets, “tertiary” for tertiary streets, “resident” for the residential streets, “footway” for footways and “path” for unspecified paths.

3.4 Traffic Network and travel speed data for the city of Munich

The traffic network and the travel speed data were obtained from TomTom, a mobility and location service developer, that also provides GPS services (TOMTOM, 2022). They also publish travel speed data for different regions across Europe. But the travel speed data was aggregated data from different users of GPS in the city of Munich. While using a trial version of TomTom, a shape file, which includes the transport network for the city of Munich, along with a database file which contained the information on travel speed in Munich was downloaded. In QGIS, the traffic network and the database file were combined using a shared “ID” attribute in both files. The database file contains different variables like Average Travel time, Average travel speed, median Travel speed etc. The Average travel speed was represented for small segments of the network, whose length was mentioned along with a segment-specific identification number, speed limit and a street name if available. The sample of data for speed is shown in the table below:

Id	1	2	3
Segment Id	-1.3E+13	-1.3E+13	-1.3E+13
NewSegId	-00004435-3100-0400-0000-0000005b9142	-00004435-3100-0400-0000-0000005b9163	-00004435-3100-0400-0000-0000005b916a
Length	129.19	77.46	75.96
FRC	1	1	1
SpeedLimit	50	50	50
StreetName	Chiemgaustraße	Chiemgaustraße	Chiemgaustraße
AvgTt	23.21	6.33	6.34
MedTt	22.04	5.62	5.83

ratio	1	1	1
AvgSp	29.6	48.62	47.56
HvgSp	20.04	44.03	43.16
MedSp	21.1	49.6	46.9
SdSp	18.3	10.22	11.8
Hits	68376	66640	69509
P5sp	11	30	29
P10sp	12	35	34
P15sp	12	40	37
P20sp	13	42	39
P25sp	13	44	40
P30sp	14	46	42
P35sp	15	47	43
P40sp	16	48	44
P45sp	18	49	46
P50sp	21	50	47
P55sp	26	51	48
P60sp	31	51	50
P65sp	38	52	51
P70sp	44	53	53
P75sp	48	55	55
P80sp	51	56	57
P85sp	53	57	59
P90sp	56	60	62
P95sp	60	63	67

Table 3.2 Attribute table for travel speed data

The annotations for the speed data attributes are as given in the table below:

Id	The value used for linking the additional DBF files per time set to the Shapefile
AvgTt	The arithmetic average travel time for this time period (seconds)
MedTt	The arithmetic median travel time for this time period (seconds)
ratio	Average travel time of comparison set divided by the base set
AvgSp	The arithmetic average speed for this time period (kph)
HvgSp	The harmonic average speed for this time period (kph)
MedSp	The median speed for this time period (kph)
SdSp	The standard deviation of the speed for this time period
Hits	The number of measurements used for the calculation
P5sp	The 5th percentile speed, 5 percent of speeds are above this value (kph)
P10sp	The 10th percentile speed, 10 percent of speeds are above this value (kph)
P...sp	Percentile speeds are given in steps of 5 centiles.
P90sp	The 90th percentile speed, 90 percent of speeds are above this value (kph)
P95sp	The 95th percentile speed, 95 percent of speeds are above this value (kph)

Table 3.3 Data Annotations for TomTom traffic data

The next step was to identify the road segments that corresponded to each individual grid cell. This was done using the QGIS software and, the “join attributes by location” command. The modified network and the grid were combined together. The output resulted in a new network, with repeated identification numbers for grid cells. A sample of this output table can be seen in the table below, here “id” is the unique identification number for each segment of the road and “id_2” is the identification number for the corresponding grid cell number.

Attribute	**Values**		
Id	22480	22929	23188
Segment Id	1.28E+13	1.28E+13	1.28E+13
NewSegId	bc390758-7ebd-48f0-abe9-37a17df4c4d3	c55233fb-5956-4cbb-afbc-57d50a45fdea	ca8a5a48-9033-4774-9e8f-94abd258f53d
Length	20.26	110.3	82.53
FRC	3	3	3
SpeedLimit	50	50	50
StreetName	Lochhausener StraÃƒÅ¸e	Lochhausener StraÃƒÅ¸e	Lochhausener StraÃƒÅ¸e
AvgSp	43.17	46.84	43.23
HvgSp	40.75	42.43	40.11
id_2	13	13	13
left	677883.5	677883.5	677883.5
top	5339139	5339139	5339139
right	678383.5	678383.5	678383.5
bottom	5338639	5338639	5338639
area	250004.6	250004.6	250004.6

Table 3.4 Initial attribute table for combined speed data

It can be seen in the table above that the “id_2” row which represents the unique identification for the grid cells shows repetition. While the row “Id”, which represents the unique identification for Network segments for the city of Munich has unique values. The same cell number 13 is matched with a unique “Id” from the road network. These are the segments which overlap with the grid cell number 13. A similar matching and pairing of both IDs were present for all the other grid cell identification numbers from 1 to 1813, for which data was available.

The next stage was to combine all the unique network segments with speed data together for one of the grid cells. The speed data was averaged for individual grid cells together based on the weight for the length segment for each speed value. So, for the row “id_2”, all the corresponding average speed values were averaged based on the corresponding length of the segments. This was done through a simple code in R. The resulting output was an excel sheet with all the Grid IDs followed by columns for attributes and lastly with a weighted average for all the rows where data was available. A sample table can be seen below:

Id	Weighted Average Speed
13	38.56799
17	53.83994
20	45.41415
21	56.68
22	44.04622
45	54.04837
50	37.85586

Table 3.5 Sample of final attribute table for the speed data

In the table above, the “Id” column shows the identification numbers for individual grid cells and their corresponding travel speeds for a few cells, which were averaged using the segment lengths as the weights. Although the data for speed was processed, there was still a significant amount of grid cells, which had infrastructure but no data for speed. This was confirmed by looking at the Histogram for this newly generated data.

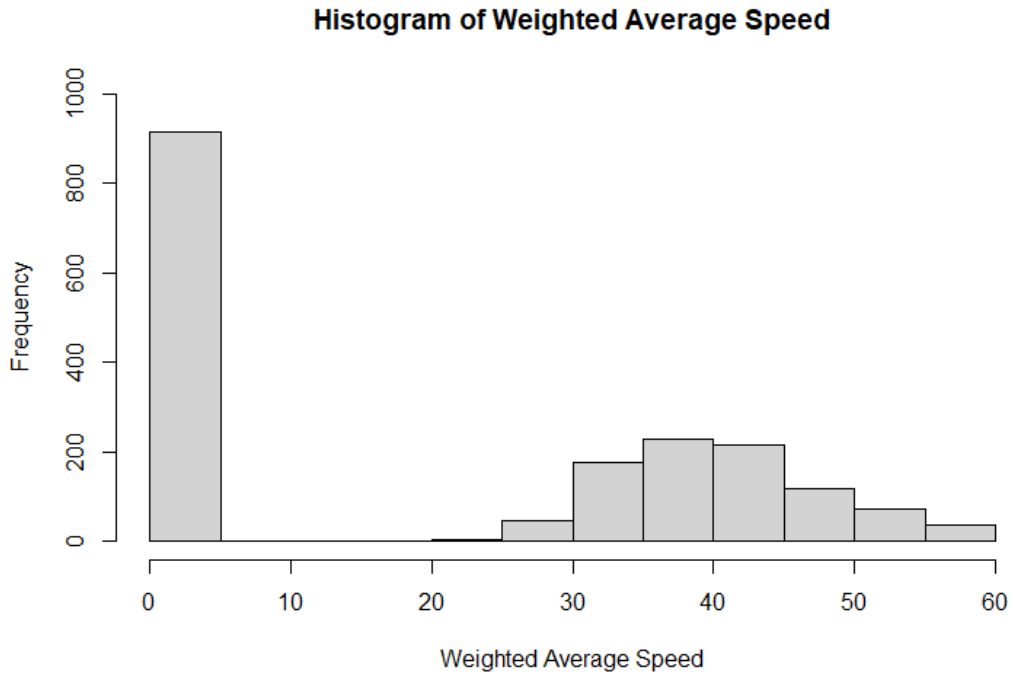


Figure 3.3 Histogram for Weighted Speed Average

As seen above, more than half of the values for speed were missing from the grid cells. Although there were some grids, where the value of speed should remain 0 owing to the lack of streets in these cells, the missing values for locations with a lack of data were addressed using a set of assumptions. For the cells without any infrastructure present, a value of 0 was kept as it is since there is no movement of vehicles in such zones. With the lack of data for speed in many segments of the roads for the city of Munch, it is understood that the city zone and streets are too complicated to individually assign a value for speed. But in the city of Munich, a concept called Tempo-30 has been introduced which reduced the speed limit on many streets within the city to 30 km/h. In a report published by the “Referat für Stadtplanung und Bauordnung, Landeshauptstadt

München", it suggests that almost 85% streets within the city of Munich already follow the speed limit of 30 km/h (Zorn, 2010, p. 8). So, in grid cells with positive integer values for infrastructure attributes and no speed value present, the assumption was made that the travel speed will be a minimum of 30 km/h based on the fact that these were inner residential streets.

While there is a lack of demand data for the analysis, like the total number of cyclists and the total number of car users, this is represented indirectly in the model. An assumption was made that since the travel speed data was generated from streets where there was a movement of cars, that shows the existence of cars in the model. Although this is not specified in terms of the number of cars, just the mere existence of cars. And the grid cells with no speed data can be considered to be the cells with no movement of cars. While this is not a very specific or a model altering assumption, this can be verified by making changes in the dataset while predicting the crashes.

3.5 Final Dataset

The final dataset used for the analysis comprises a grid that covers the entire city of Munich, with each grid cell of size 500 m X 500 m. The count data was generated using the "count points in polygon" and "sum line lengths" commands. Count points in polygon command require two inputs, a polygon file and a points file. Individually for each attribute, the count dataset was generated which consists of the grid identification and the corresponding number of counts for the attributes. For the line attributes, the command "sum line by length" command was used. This command requires a polygon and line attribute as input and the resultant output is a new shape file, with the polygon file as the main attribute and the last column as the length of the line attributes corresponding to each polygon. For the analysis, the inputs were the main grid and the individual line attributes which include, bicycle paths, primary streets, secondary streets, tertiary streets, residential streets, footways and unspecified paths.

Both these processes were done for all the point feature attributes and line features resulting in individual grid shape files for all of them. Lastly, to obtain the final dataset, all the grid shape files with count and length data were merged to obtain the final dataset. This was done using the "join" option from table properties and using "ID" as the target field, since "id" for all the grid remains consistent. A sample of the resulting dataset is shown in the table below:

Actual attribute name	Values				
Year	2016	2016	2016	2016	2016
id	938	939	940	941	942
Primary_id	2016938	2016939	2016940	2016941	2016942
area	249977	249976.6	249976.7	249976.7	249976.9
crashpoint	4	7	6	10	5

signal	6	14	7	18	3
tram	0	2	1	3	0
busstop	2	0	2	3	5
railstop	2	0	0	0	0
crossing	8	23	11	16	12
junction	0	0	0	0	0
education	0	0	0	0	1
biergarten	0	0	0	0	0
supermarkt	1	1	1	3	3
bicycle	2736.157	2501.848	1235.563	2711.306	0
primary	0	0	0	0	0
secondary	1465.619	2135.413	1154.69	1463.527	510.1173
tertiary	0	0	0	871.04	0
resident	2364.203	2142.29	2569.949	2252.223	3745.116
footway	3461.823	4787.104	2522.41	3978.755	5661.775
path	49.03771	417.0872	61.86486	0	0
Weighted Average Speed	44.27952	37.59701	29.71777	29.90535	27.71655

Table 3.6 Final Dataset Attributes

Since the infrastructure and speed data available was only for the year 2020, the same data is used for the different years from 2016-to 2019. So, using the steps mentioned above the dataset for all the years was prepared and saved as an excel file. Finally, the excel sheets for all years were concatenated to give a final combined dataset. This included the year 2016-2019, which was used to analyze using the different statistical models and the data for the year 2020 was used for calibration.

3.5.1 Dataset for Spatial Regression

For spatial regression, the input data has to be shapefile with all the regression attributes. A shapefile is used to store information like the location and physical attributes of geographical features. The dataset for spatial regression consists of all the same attributes as used in the statistical regression models. While this can be easily done for the data for one year using the “count points in polygon” command, the data for different years cannot be merged or concatenated together as easily. This is because the shape file retains its shape of the grid for each individual year, so multiple years cannot be concatenated together but rather merged by taking an average. To overcome this issue, an assumption was made to take an average number of crashes for all four years from 2016-to 2019. To obtain the dataset for spatial regression, first, all the attribute shape files were combined using the join function in QGIS. Then, for the attribute “crashpoint”, the crash locations for all years were simultaneously merged into one shape file. Using the “count points in polygon” command, these attributes were then merged into the new dataset for spatial regression. The resulting dataset has the same attributes as the dataset for regression analysis for the infrastructure, but the number of rows is reduced down to 1813 and the value for “crashpoint” is taken as an average of the count for all four years of data.

4 Analysis

As part of the analysis, four models were evaluated based on the best statistical fit, and accuracy of predicted values. Based on the literature review, the selected models include the Multiple linear regression mode, Poisson regression model, the Negative binomial mode and the Zero-inflated Negative Binomial model. Along with these, two spatial regression models were also used to analyze the data, although they were used on a trial basis, they have been explained in terms of goodness of fit and their limitations in this kind of analysis. The code that was used in the analysis can be found on the "GitHub" repository under the name Bicycle-crash-analysis-code.

4.1 Data Correlation and exploration

Before doing any sort of regression or modelling, a correlation between all the variables was established. This was done by simply using the "cor" command in R (R Core Team, 2021). This command helps calculate the correlation between all the variables. This is generally done using the "Pearsons" method. It is one of the most common methods used to rank variables. The method uses numbers between -1 and 1 to assign a rank, based on how strong or weakly those variables are correlated. While a negative sign indicates a negative correlation and a positive sign indicates a positive correlation. A 0 indicates the weakest correlation, 1 indicates the strongest positive correlation and -1 indicates the strongest negative correlation (Nettleton, 2014). The table below shows the correlation between all the variables used for analysis.

Attributes	crashpoint	signal	tram	busstop	railstop	crossing	junction	education	biergarten	supermarkt	bicycle	primary	secondary	tertiary	resident	footway	path	Weighted Average Speed
crashpoint	1																	
signal	0.64	1																
tram	0.43	0.46	1															
busstop	0.31	0.47	0.13	1														
railstop	0.39	0.34	0.18	0.3	1													
crossing	0.58	0.75	0.45	0.45	0.36	1												
junction	-0.02	0.02	0.03	-0.03	-0.04	0.01	1											
education	0.39	0.31	0.24	0.16	0.19	0.3	-0.04	1										
biergarten	0.18	0.14	0	0.15	0.09	0.15	-0.01	0.07	1									
supermart	0.52	0.49	0.27	0.35	0.39	0.47	-0.03	0.26	0.11	1								
bicycle	0.4	0.57	0.24	0.42	0.23	0.54	0.06	0.14	0.14	0.28	1							
primary	0.15	0.27	0.07	0.17	0.02	0.19	0.03	0.02	0.01	0.07	0.28	1						
secondary	0.5	0.66	0.36	0.4	0.26	0.61	0.03	0.23	0.14	0.35	0.6	-0.01	1					
tertiary	0.13	0.22	0.09	0.29	0.13	0.2	0.02	0.09	0.07	0.11	0.16	-0.02	0.03	1				
resident	0.35	0.34	0.15	0.43	0.19	0.37	-0.07	0.13	0.09	0.34	0.28	0.08	0.24	0.18	1			
footway	0.37	0.4	0.24	0.39	0.27	0.42	-0.01	0.26	0.13	0.36	0.42	0.07	0.34	0.12	0.36	1		
path	-0.08	-0.08	-0.04	-0.07	-0.05	-0.06	-0.02	-0.02	0.03	-0.07	0.01	-0.05	-0.06	-0.07	-0.12	-0.01	1	
Weighted Average Speed	0.16	0.25	0.08	0.3	0.1	0.25	0.14	0.05	0.07	0.12	0.47	0.21	0.29	0.17	0.36	0.35	0.01	1

Figure 4.1 Correlation chart for all attributes

The colour scheme shows the attributes with the strongest positive correlation with shades of green colour, while those with negative correlation in shades of red. As seen in the image above, the attribute "crashpoint" shows a positive correlation with almost all variables, except for the variables "junction" and "path". This suggests that an increase in the number of junctions or length of the path in a grid should result in a decrease in the value of the number of crashes, or a reduction in crashes and a decrease in these attributes results in an increase in crashes. While a positive correlation with other variables indicates that the crashes should decrease with a decrease in the value of those

variables. At the same time, the “Weighted Average Speed” attribute shows the strongest correlation with the attributes “bicycle” and “resident”, this is due to the data manipulation done as part of data processing which included adding speed values corresponding to the residential streets.

While, one of the strongest correlations exists between the pair, “Crossing and Signal” with a value of 0.75, which is appropriate considering that signals mostly exist at almost every road crossing. The highest negative correlation exists between “path and resident” attributes, with a value of -0.12. This suggests that unspecified paths reduce whenever residential streets increase, this is because most of the unspecified paths are located in parks and open grounds.

These correlations can be easily influenced so they should be investigated more to understand the influence of outliers. This was done by visualizing the data and plotting histograms for all the attributes. For the data used in the analysis, most of the data is based on real-life observations from OpenStreetMap, meaning they exist in the real world. So, the existence of outliers could be detected statistically but that will not correspond to the real world. While removing these values may be good for the model itself, it may reduce the real-world representation of the model. So, while looking at the histograms for individual attributes, data should also be checked in QGIS to see if this data truly exists or is merely a counting error. First, we take a look at the attributes “crashpoint” and “Weighted Average Speed”, which represent the number of crashes in a grid cell and weighted average travel speed in a grid cell respectively. This data is most susceptible to errors, as it is based on derived data.

Weighted Average Speed

The histogram for the weighted Average speed as shown below is different as compared from the one shown previously in Figure 3.3. The key difference between these two is the addition of new values in the speed data, which were the additions done for the inner streets of the city of Munich.

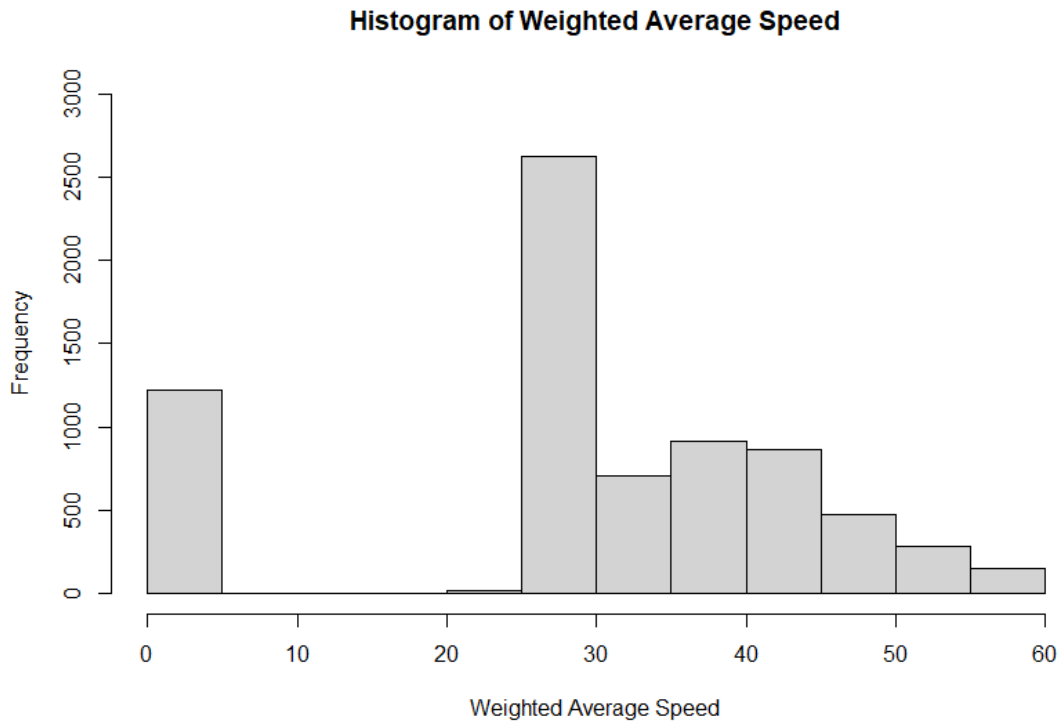


Figure 4.2 Histogram of modified Weighted Average Speed for the period 2016-2019

As seen in the image above, for weighted speed data, values for 0 represent the streets where no cars or other automobiles are driven. It is important to understand the effect of the existence and non-existence of vehicles within the model. The frequency for values between 0 and 20 is missing significantly, as the speed data obtained from the data provider TomTom, only accounted for the data on primary and secondary streets. So when the data was worked on and values were added based on the assumption mentioned in section 3.4, the number of speed data for the value 30 increased. While, the speed data in places with no infrastructure was kept at 0, which explains the loss of values from 0 and the missing values between 0 and 20.

Number of crashes count

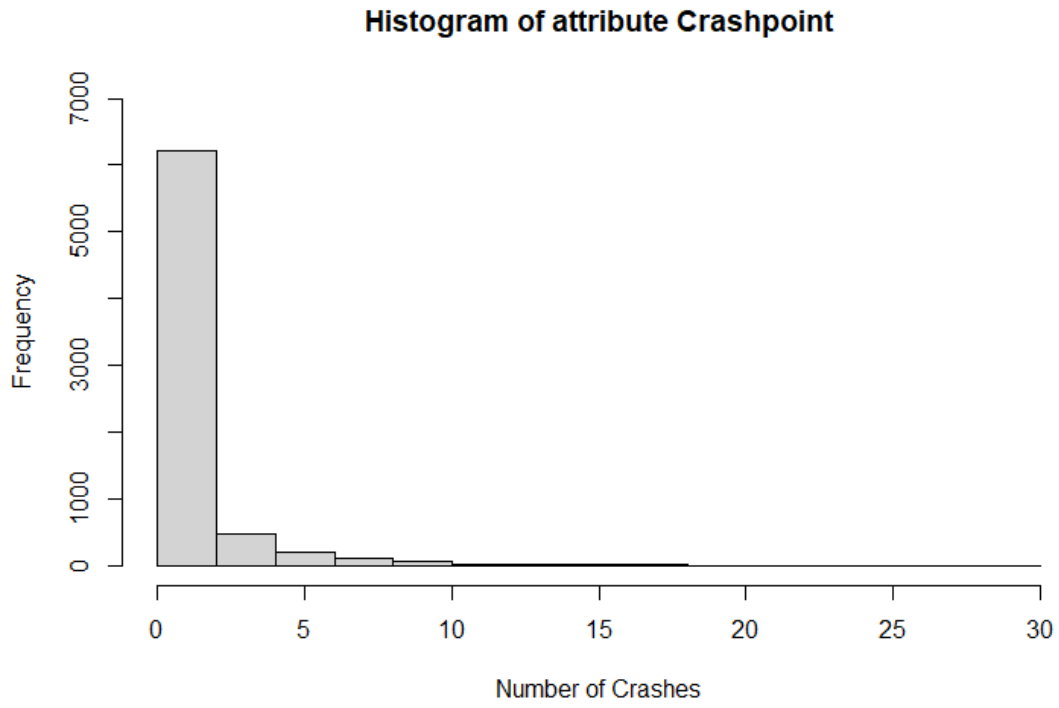


Figure 4.3 Histogram of Crash point counts for the period 2016-2019

For the attribute “crashpoint”, although 30 seems to be an outlier, the methodology by which the crash points were calculated as mentioned in section 3.5, the count 30 is just a summation of crash points over a grid cell. This was verified by looking at the shapefiles for respective grids and verifying the crash locations. First, the grid cell with the maximum value of attribute “crashpoint” was identified, this was done in R, by simply using the “which.max” command. Upon execution, the row with the highest value for attribute “crashpoint” is printed which includes the identification number of the grid cell.

It was observed that the maximum value of the attribute “crashpoint” for the whole dataset is 29 and the corresponding year and grid cell number is 2016 and 1058. So, the dataset for the year 2016 and the grid cell with id 1058 were checked to verify whether 29 crash points truly exist. This was done using QGIS.

Figure 4.4 Grid slot 1058 for the year 2016

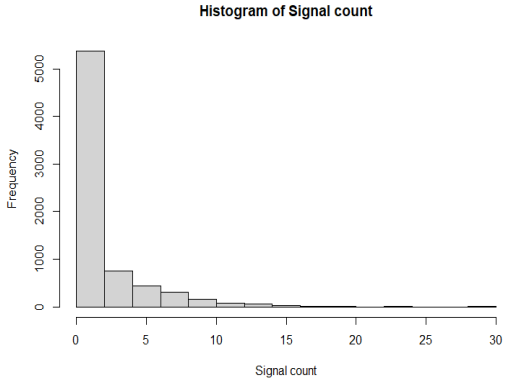
As seen in the image above, the grid cell for 2016 does correspond to the same number of accident counts i:e 29. So, this value was taken into consideration for our model as this grid cell represents the most severe crash zone in the dataset.

For the other attributes, the summary obtained through R was analyzed carefully to find out any substantial outliers. The table below shows the min, median, mean, and maximum values for all the attributes that were considered for the different models. This was done to understand the presence of any unusual values in the dataset, that could be verified using the data in QGIS before proceeding with the actual analysis. The table below shows the values for all the attributes that are considered in the analysis.

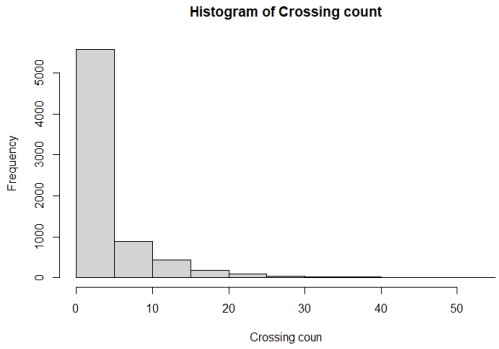
Sr. No.	Attribute	Parameter	Value	Sr. No.	Attribute	Parameter	Value
1	crashpoint	Min.:	0	10	bicycle	Min.:	0
	crashpoint	Median:	0		bicycle	Median:	397
	crashpoint	Mean:	1.157		bicycle	Mean:	651.8
	crashpoint	Max.:	29		bicycle	Max.:	4136
2	signal	Min.:	0	11	primary	Min.:	0
	signal	Median:	0		primary	Median:	0
	signal	Mean:	1.821		primary	Mean:	75.58
	signal	Max.:	29		primary	Max.:	2551.75
3	tram	Min.:	0	12	secondary	Min.:	0
	tram	Median:	0		secondary	Median:	0
	tram	Mean:	0.1081		secondary	Mean:	261
	tram	Max.:	6		secondary	Max.:	3241.1
4	busstop	Min.:	0	13	tertiary	Min.:	0
	busstop	Median:	0		tertiary	Median:	0
	busstop	Mean:	1.522		tertiary	Mean:	117.35
	busstop	Max.:	19		tertiary	Max.:	1610.91
5	railstop	Min.:	0	14	resident	Min.:	0
	railstop	Median:	0		resident	Median:	842
	railstop	Mean:	0.08549		resident	Mean:	
	railstop	Max.:	3		resident	Max.:	4439
6	crossing	Min.:	0	15	footway	Min.:	0
	crossing	Median:	0		footway	Median:	959.78
	crossing	Mean:	3.47		footway	Mean:	1556.42
	crossing	Max.:	51		footway	Max.:	10905.77
7	junction	Min.:	0	16	path	Min.:	0
	junction	Median:	0		path	Median:	252.09
	junction	Mean:	0.05405		path	Mean:	513.89
	junction	Max.:	4		path	Max.:	5901.13
8	education	Min.:	0	17	Weighted_Average.Speed	Min.:	0
	education	Median:	0		Weighted_Average.Speed	Median:	30
	education	Mean:	0.1197		Weighted_Average.Speed	Mean:	30.19
	education	Max.:	7		Weighted_Average.Speed	Max.:	59.83
9	biergarten	Min.:	0	18	supermarkt	Min.:	0
	biergarten	Median:	0		supermarkt	Median:	0
	biergarten	Mean:	0.03751		supermarkt	Mean:	0.3067
	biergarten	Max.:	2		supermarkt	Max.:	7

Table 4.1 Summary of all attributes used in the analysis

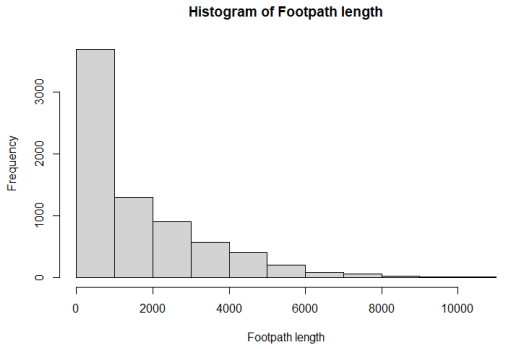
The attributes which show some possible outliers are, “signal”, “crossing”, “busstop”, “footway”, “bicycle”. One of the easiest ways to identify outliers is by using some form of plot. Here, this was done using the histogram for all the attributes with possible outliers.



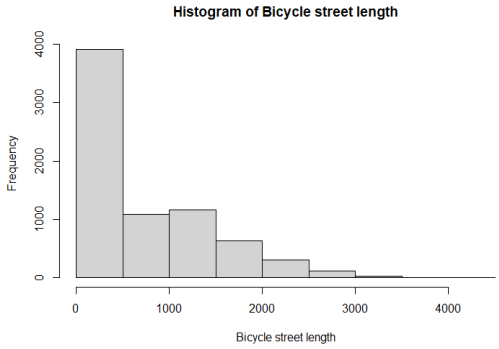
Histogram of number of traffic signals per grid cell for the period 2016-2019



Histogram of number of Crossings per grid cell for the period 2016-2019



Histogram of the length of footpath per grid cell for the period 2016-2019



Histogram of the length of bicycle path per grid cell for the period 2016-2019

Figure 4.5 Histograms for attributes with possible outliers

The histogram for the attribute's "signal", "crossings", "footpath" and "bicycle" is as shown above. The maximum values for these attributes were previously considered to be unusual, however, the histogram plots show that the data is smoothly distributed across the X-axis with almost no missing values in between, except for the attribute "signal". This suggests that the values, even though unusually high are not outliers, confirmed through the histogram plots. QGIS was also used to verify these high values after visualizing individual attributes and checking the corresponding grid cell against the original data.

After exploring the data, the next step was modelling the data and this is discussed in the next section.

4.2 Multiple Linear Regression

The first statistical model was a simple regression model. Linear regression is one of the most basic models which can be used to understand the relationship between two or more variables. It can be plotted simply by using the equation:

$Y = ax_n+b$, where y = dependent variable, $x_n = n^{th}$ independent or explanatory variable, a = slope of the regression line, b = intercept.

In r, this was done using the "lm" package in r, which uses a formula to run the regression of the independent variable against the dependent variable (R Core Team, 2021). For the first iteration, all the attributes were used in the linear model, but the attributes which were not significant were discarded and the rest were used for the final iteration.

The output for the final iteration is shown in the figure below:

Residuals:					
	Min	1Q	Median	3Q	Max
	-9.3514	-0.5508	-0.003	0.2116	19.4074
Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.98E-03	4.77E-02	0.063	0.95	
signal	2.20E-01	1.18E-02	18.616	<2E-16	***
tram	6.58E-01	5.25E-02	12.536	<2E-16	***
busstop	-1.26E-01	1.26E-02	-10.027	<2E-16	***
railstop	1.10E+00	7.79E-02	14.16	<2E-16	***
crossing	2.80E-02	5.82E-03	4.801	1.61E-06	***
education	8.27E-01	4.47E-02	18.484	<2E-16	***
biergarten	1.06E+00	1.11E-01	9.607	<2E-16	***
supermarkt	6.18E-01	3.26E-02	18.96	<2E-16	***
primary	5.65E-04	9.33E-05	6.063	1.40E-09	***
secondary	7.39E-04	7.05E-05	10.479	<2E-16	***
resident	3.02E-04	2.27E-05	13.274	<2E-16	***
WeightedAverage Speed	-7.15E-03	1.58E-03	-4.522	6.23E-06	***

Signif. Codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1
Residual standard error:	1.808	on	7239	degrees of freedom	
Multiple	R-squared:	0.5533,	Adjusted R-squared:		0.5526
F-statistic: 747.2 on 12 and 7239 DF,				p-value:	<2E-16

Table 4.2 Output summary for Linear Regression model

The final iteration of linear regression against the attribute "crashpoint" gives 12 highly significant attributes out of a total of 17 that were considered initially. Here the Adjusted r-squared value was 0.5526. But as compared to the first iteration with all the attributes, the adjusted R-squared value was reduced by a very small margin, from 0.533 to 0.5526. This is a result of having fewer attributes as part of the regression, even if the dropped

variables are statistically insignificant, they affect the model fit and the R-squared value. To understand the influence of the attributes on the dependent variable, the standardized variables should be looked at. They are shown in the table below.

Attributes	Standardized Coefficients
signal	0.2686
tram	0.1152
busstop	-0.0984
railstop	0.1255
crossing	0.0620
education	0.1559
biergarten	0.0770
supermarkt	0.1841
primary	0.0532
secondary	0.1214
resident	0.1248
Weighted Average Speed	-0.0404

Table 4.3 Standardized coefficients for Linear model

Here only two attributes show a negative correlation, they are “busstop”, and “Weighted Average Speed”. This negative correlation suggests that every time the value for these two attributes is increased, the number of crashes should decrease. The attribute with the strongest effect on the number of crashes is the attribute “signal”, which suggests that the greater the number of signals in a grid cell can result a greater number of crashes. At the same time, the attribute representing speed also shows a negative sign suggesting, an inverse relationship for the number of crashes although with a marginal effect. This suggests that when motot vehicles are driving faster, the likelihood of a bicycle crash is reduced. Another attribute with inverse relation is “busstop”, but with a significantly small effect.

Using the final linear regression model, we plot the observed and predicted values with the regression line.

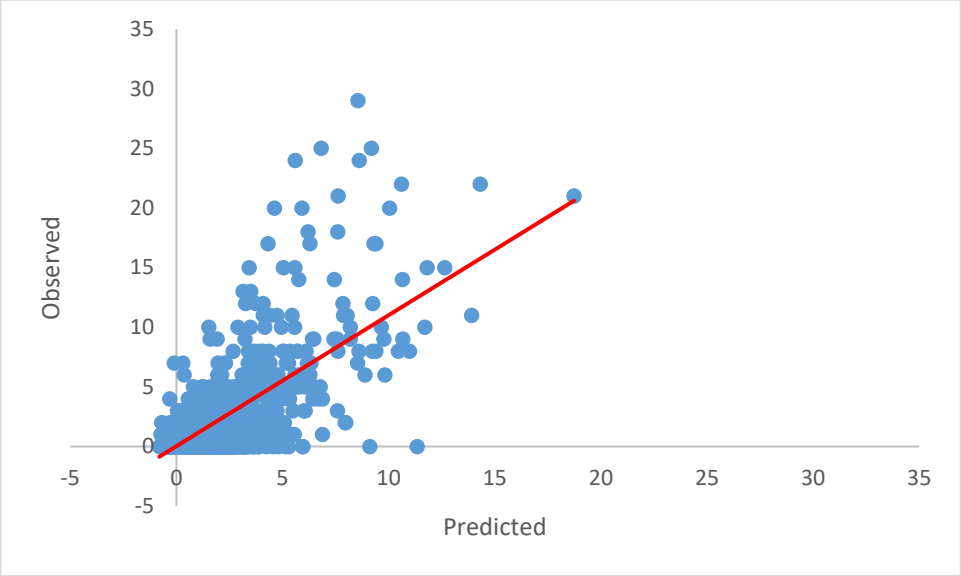


Figure 4.6 Predicted and Observed values for the linear regression model

As seen in the image above, some negative values are predicted by the model. This is because the data not being normally distributed and the linear regression model is not bound at the value 0. This results in the model predicting some negative values. Because of the presence of negative values and poor fit, this model cannot be used to make any inferences or predictions for the year 2020. the maximum predicted value for the linear regression is 18.745, which is significantly lower than the observed value of 29. The statistical fit, the presence of negative values and the under-representation of crashes make the linear regression model a poor choice to analyze crashes in this case.

4.3 Poisson Model

One of the most popular regression models for count data is a Poisson regression model. And the Poisson model for regression assumes that the Variance (Yi) is equal to the mean E(Yi). But more often than not, the crash data shows over-dispersion or under-dispersion due to the presence of zeroes. In (Miaou, 1994), the equation for Poisson regression is stated as:

$$\log(\mu_i) = \beta_0 + \beta_1 X_i$$

where, μ_i = conditional expectation of y_i, β_0 = is the intercept and β_1 = coefficient marked x. In the case of Poisson regression, there is no error term like in linear regression, as the μ determines both the mean and variance of the Poisson random variable. A variable is said to have Poisson distribution if y has positive integer values with the probability:

$$\Pr\{Y = y\} = \frac{e^{-\mu}\mu^{y}}{y!}$$

where, μ = parameter, and the mean E(Y) and the variance var(Y) are the same as μ when it is greater than 0. For our data, the unconditional mean of the outcome variable was found to be 1.16 and the variance was 7.31. Since the variance was significantly bigger than the mean, this suggested that there might be some overdispersion in the

model (Zhu, 2012). The Poisson regression was run using the “glm” package. This package provides an option to perform regression to be performed with eight family choices, here we use the family “Poisson”. The first iteration again uses all the variables. The summary of the first iteration is shown in the table below.

Deviance	Residuals:					
	Min	1Q	Median	3Q	Max	
	-5.4699	-0.8978	-0.5456	-0.1952	9.8114	
Coefficients:						
	Estimate	Std. Error	z value	Pr(>|z|)		
(Intercept)	-2.45E+00	6.56E-02	-37.26	<2.00E-16	***	
signal	4.13E-02	3.29E-03	12.538	<2.00E-16	***	
tram	1.02E-01	1.25E-02	8.148	3.70E-16	***	
railstop	1.21E-01	2.07E-02	5.846	5.04E-09	***	
education	2.41E-01	1.08E-02	22.288	<2.00E-16	***	
biergarten	2.80E-01	3.07E-02	9.107	<2.00E-16	***	
supermarkt	1.27E-01	8.86E-03	14.383	<2.00E-16	***	
bicycle	3.50E-04	1.85E-05	18.943	<2.00E-16	***	
primary	4.70E-04	3.62E-05	12.975	<2.00E-16	***	
secondary	4.27E-04	2.73E-05	15.656	<2.00E-16	***	
tertiary	1.92E-04	4.29E-05	4.471	7.77E-06	***	
resident	4.71E-04	1.16E-05	40.52	<2.00E-16	***	
footway	1.30E-04	6.74E-06	19.294	<2.00E-16	***	
path	-7.96E-05	2.23E-05	-3.577	0.000348	***	
Weighted Average Speed	1.47E-02	1.67E-03	8.821	<2.00E-16	***	

Signif. Codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ' '	1
(Dispersion parameter for Poisson family taken to be 1)						
Null deviance:	25318	on	7251	degrees of freedom		
Residual deviance:	10228	on	7237	degrees of freedom		
AIC:	16992					
Number of Fisher Scoring iterations:	5					

Table 4.4 Output Summary for Poisson model

The first thing to notice in the summary for the Poisson model is the Deviance residual, which shows the minimum, 1stQ, mean, 3rd Q and the maximum value. While looking at this data, there seems to be a lack of symmetry when all the values are compared together. Here the minimum value is -5.47, the median is -0.5456 and the maximum is 9.81 which indicates some sort of skew in the model. Next, there is the estimate, standard error, the z value and the p-value that is associated with it. While the standard error represents the average distance of the observed value from the regression line, the z

value is obtained by dividing the estimate by the standard error. The z value shows the difference between the mean for the specific attribute in the dataset.

In the case of Poisson regression, since the dependent variables take the log, their coefficients need to be exponentiated to obtain the true coefficients of the regression. Here Poisson model coefficient for Signal showed 4.13E-02, this means that the expected log count for one unit increase in signal is 0.04267 when all other attributes remain constant. A similar interpretation can be made for all the other variables. While only the attribute "path" shows a negative coefficient, inferring that for every log count increase in the length of an unspecified path in a grid cell, results in a reduction by -7.96E-05.

A plot of the predicted and actual values is shown below.

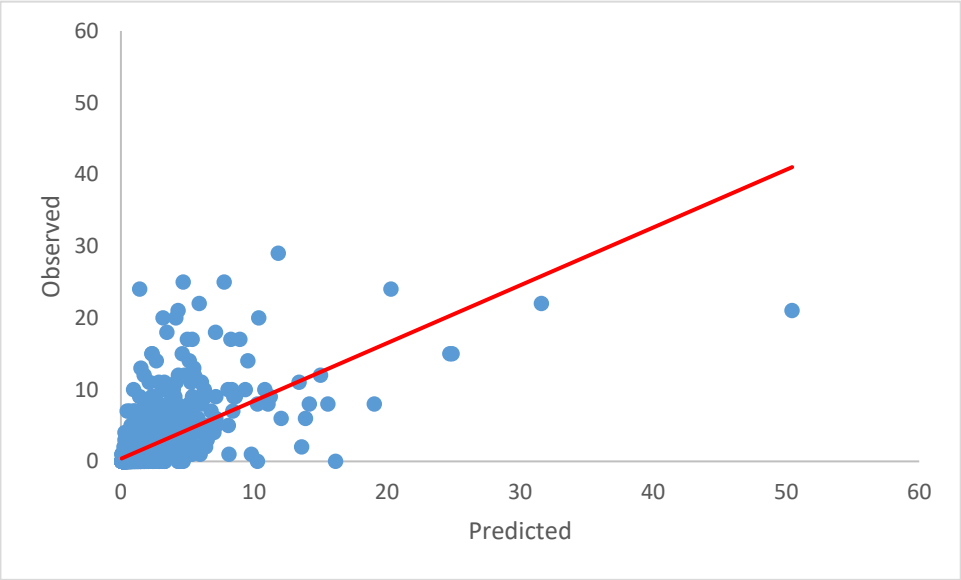


Figure 4.7 Predicted and Observed values for the Poisson regression model

The values predicted using the Poisson model need to be exponentiated to obtain the true values. These predicted values are plotted against the observed values. As seen, there is no presence of any negative values. While the highest observed value is 29, the highest predicted value is 50.

The model was checked for dispersion which can be calculated using the Pearson residuals of the model fit and the number of independent parameters. In R, this was done using the "dispersiontest" from the package "AER" (Kleiber & Zeileis, 2008), and the value is found to be 2.151. Since this value is bigger than 1, this suggests that there is some overdispersion in the Poisson model, which means that any predictions made with this model are bound to have an inclination, which makes it unsuitable to make any reliable interpretations. One of the most common causes of overdispersion in a model is the existence of excess zeroes, which can be addressed by a zero-inflated model (Lambert, 1992).

4.4 Negative Binomial and Zero Inflated Negative Binomial model

While the Poisson regression model is suitable for count models, oftentimes such data contains an excess number of zeroes. The Negative Binomial model, also known as the Poisson-Gamma model, can overcome the limitation of the Poisson model regarding the condition of mean and variance. The selection criteria of the model are dependent on the existence of overdispersion (Lord & Mannering, 2010). Since our data does exhibit overdispersion, variance is significantly greater than the mean, and the use of a Negative Binomial model is ideal in this situation (Park & Lord, 2009). A Negative binomial (NB) regression can be executed using the "glm.nb" package from R (Venables & Ripley, 2002). Although the literature suggests that a Zero-inflated Negative Binomial (ZINB) regression model will be better suited to the data that shows excess zeroes, skipping over the NB model is not ideal. While ZINB may be better suited theoretically, it must be verified against the regular NB model. So, a negative binomial model helped establish a comparison for a zero-inflated negative binomial model, which can be done using different statistical tests. The NB model was run with the same parameters that were used for the zero-inflated model. Below is the summary of the first iteration of the output.

Deviance	Residuals:					
	Min	1Q	Median	3Q	Max	
	-2.5036	-0.7351	-0.4189	-0.2265	4.7901	
Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)		
(Intercept)	-3.35E+00	9.81E-02	-34.154	<2.00E-16	***	
signal	5.58E-02	7.00E-03	7.973	1.54E-15	***	
tram	2.17E-01	2.92E-02	7.432	1.07E-13	***	
railstop	3.62E-01	4.71E-02	7.686	1.52E-14	***	
education	2.43E-01	2.67E-02	9.112	<2.00E-16	***	
biergarten	3.33E-01	7.14E-02	4.668	3.04E-06	***	
supermarkt	1.80E-01	1.98E-02	9.08	<2.00E-16	***	
bicycle	4.14E-04	3.21E-05	12.913	<2.00E-16	***	
primary	5.39E-04	6.67E-05	8.086	6.19E-16	***	
secondary	4.94E-04	5.21E-05	9.48	<2.00E-16	***	
tertiary	3.44E-04	7.28E-05	4.727	2.28E-06	***	
resident	5.27E-04	1.88E-05	27.983	<2.00E-16	***	
footway	1.39E-04	1.14E-05	12.215	<2.00E-16	***	
Weighted Average Speed	2.63E-02	2.50E-03	10.508	<2.00E-16	***	

Signif. Codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1
(Dispersion parameter for Negative Binomial (1.276) family taken to be 1)						
Null Deviance: 12751.6		on	7251	degrees of freedom		
Residual Deviance: 4923.4		on	7238	degrees of freedom		
AIC:	14578					
Number of Fisher Scoring iterations:			1			

	Theta:	1.276
	Std. Err:	0.0586
2x log likelihood:		-14548.2

Table 4.5 Output Summary for Negative Binomial Model

Here, all attributes were observed to have a positive coefficient except for the intercept itself. Looking at the coefficients, the three attributes with the strongest effect on the dependent variable were the attributes “education”, “railstop” and “biergarten”. The coefficient of railstop is 3.62E-01 or 0.362, which indicates that for every one-unit increase in signal, the expected log count of the attribute “crashpoint” increases by 0.362. The coefficient for the attribute “biergarten” is 3.33E-01 or 0.333, which suggests that with one unit increase in biergartens, the expected log count of “crashpoint’ increases by 0.333. And the coefficient for the attribute “railstops” is 2.43E-01 or 0.2431, which suggests that one unit increase in rail stops, the expected log count of “crashpoint” increases by 0.2431. This model will be used to compared the Zero-inflated negative binomial model to compare which gives a better statistical fit.

At the same time, an excess number of zeroes were observed in the crash data count, as seen in Figure 4.3. For this purpose, a zero-inflated model can be used. A zero-inflated model makes use of two distinct modelling procedures which can model crashes in two states, the zero-crash state also known as the count model and the non-zero crash state which is the binary logit model (Shankar et al., 1997). Since the data consists of both over dispersion and excess zeroes, the literature suggests that a zero-inflated negative binomial (ZINB) model should yield the best results. For a ZINB regression mode for which a response variable Y_i, where i is a positive integer greater than 0, has a probability mass function as:

$$\Pr(Y_i = y_i) = \begin{cases} p_i + (1-p_i)\left(\frac{\emptyset}{\mu_i+\emptyset}\right)^{\emptyset}, & y_i = 0, \\ (1-p_i)\frac{\Gamma(\emptyset + y_i)}{\Gamma(y_i+1)\Gamma(\emptyset)}\left(\frac{\mu_i}{\mu_i+\emptyset}\right)^{y_i}\left(\frac{\emptyset}{\mu_i+\emptyset}\right)^{\emptyset}, & y_i = 1,2,3,\ldots \end{cases}$$

where $0 \le p_i \le 1$, $\mu_i \ge 0$, $\emptyset$ is the dispersion parameter with $\emptyset > 0$ and Γ (.) is the gamma function (Garay et al., 2011).

The ZINB model is best suited for the data that has two kinds of zeroes, structural zeroes and sampling zeros (Washington et al., 2020). Structural zeros are the zeroes that are obtained for attributes that can only have a value of 0, for example, a grid cell that overlooks a lake in the city, cannot have a bicycle crash inside, so it will always remain 0. A sampling zero occurs when a grid cell with a possibility of a crash, has no crash occurring in it. This is modelled easily by the zero-inflated model. The count model and the zero models, both model the 0 values but the zero-inflation component adds additional zeros to data, hence the name zero inflation.

The Zero-inflated Negative Binomial model was executed using the “zeroinfl” of “pscl” package (Zeileis et al., 2008). This model can be analyzed as a Poisson model or as a negative binomial model. To compare the models, the metric of dispersion statistic was

used, which was calculated using the Pearson residuals and the independent attributes. Dispersion statistics can help find out the existence of under-dispersion or over-dispersion in a model.

The summary of the model is shown below:

Call:	zinflm <- zeroinfl(crashpoint ~ signal+tram+railstop+education+biergarten+supermarkt+bicycle+primary+secondary+resident+Weighted Average Speed | signal+tram+railstop+education+biergarten+supermarkt+primary+secondary, data = data2, dist = "negbin")				
Pearson residuals:					
	Min	1Q	Median	3Q	Max
	-1.1816	-0.4199	-0.2727	-0.1954	16.6837
Count model coefficients (negbin with log link):					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.25E+00	1.56E-01	-14.471	<2.00E-16	***
signal	5.25E-02	7.04E-03	7.452	9.17E-14	***
tram	2.40E-01	2.91E-02	8.251	<2.00E-16	***
railstop	3.83E-01	4.36E-02	8.768	<2.00E-16	***
education	2.69E-01	2.51E-02	10.73	<2.00E-16	***
biergarten	3.11E-01	7.21E-02	4.31	1.64E-05	***
supermarkt	1.75E-01	2.01E-02	8.712	<2.00E-16	***
bicycle	3.72E-04	4.58E-05	8.114	4.91E-16	***
primary	3.40E-04	6.58E-05	5.169	2.35E-07	***
secondary	3.22E-04	5.07E-05	6.341	2.28E-10	***
resident	4.44E-04	5.06E-05	8.777	<2.00E-16	***
Weighted Average Speed	2.04E-02	2.79E-03	7.33	2.30E-13	***
Log(theta)	4.57E-01	8.23E-02	5.558	2.73E-08	***
Zero-inflation model coefficients (binomial with logit link):					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.69E-01	1.35E-01	2.736	0.00622	**
signal	-1.86E+00	6.55E-01	-2.832	0.00462	**
tram	-1.18E-01	1.05E+00	-0.112	0.91044	
railstop	-5.82E-01	5.67E-01	-1.025	0.30526	
education	-7.95E-01	7.55E-01	-1.052	0.29284	
biergarten	-6.90E-01	6.01E-01	-1.149	0.25074	
supermarkt	-4.29E+00	4.48E+00	-0.959	0.33748	
primary	-6.82E-05	5.38E-04	-0.127	0.89917	
secondary	-2.03E-03	4.94E-04	-4.117	3.84E-05	***
Signif. Codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ' ' 1
Theta	=	1.5796			
Number of iterations in BFGS optimization:			58		
Log-likelihood:	-7242	on	22	Df	

Table 4.6 Output summary for the Zero Inflated Negative Binomial model

As seen in the table above, the call for the zero-inflated model in R consists of two parts. The first part is used for the count model and the second part is used for the zero-inflated logit component of the model. They are separated into two parts, which are separated by a "|" symbol in the call of the model. It can be observed that the model with count data has no attributes with a negative coefficient. All the attributes have a positive coefficient. While, if all the values become absolute zeros, indicating no movement, logically there are bound to be no crashes. From the correlations, the attribute "crashpoint" corresponds strongly with the attribute "signal". In the count model part, the coefficient for the attribute "signal" is 5.25E-02, which means that for each one-unit increase for signal, the expected log count for the number of crashes increases by 5.25E-02 when all the other variables remain constant. The attributes, "signal" and "secondary" are statistically significant for the zero-inflation component. At the same time, the log odds of being an excessive zero would decrease by 1.86 for every additional signal in a grid cell. In other words, the more signal in the grid cell the less likely that a zero would be due to a crash not happening. Or, the more the number of signals, the more likely that a crash happened in a grid cell. For the attribute "secondary", the log odds of being an excessive zero would decrease by 0.00203 for every additional meter of secondary street in a grid cell. Or in other words, the more secondary streets in a grid cell, the more likely that a crash will happen in that grid cell.

The exponentiated coefficients for the ZINB model are:

Model type	Attribute	Value
Count Model	Intercept	0.1054
	signal	1.0538
	tram	1.2716
	railstop	1.4661
	education	1.3087
	biergarten	1.3644
	supermarkt	1.1912
	bicycle	1.0004
	primary	1.0003
	secondary	1.0003
	resident	1.0004
	Weighted_Average.Speed	1.0206
Zero-inflation model	(Intercept)	1.4465
	signal	0.1563
	tram	0.8891
	railstop	0.5590
	education	0.4518
	biergarten	0.5014
	supermarkt	0.0136
	primary	0.9999
	secondary	0.9980

Table 4.7 Exponentiated coefficients for the Zero Inflated Negative Binomial model

The baseline odds for crashes happening in a grid cell are 1.4465. These odds are increased by one unit increase in the attribute signal by 0.15. The odds of a crash happening are decreased the most when the values for primary and secondary attributes are increased. One of the strongest influencers for crash probability is the attribute “railstop”. This suggests that grid cells with a high number of rail stops have a strong influence on crashes occurring. The next three attributes with a high effect on count data are “tram” “biergarten” and “education”, which represent tram stops, beer gardens and education institutes. The important thing to remember while interpreting these values is that they are not influenced by the zero-inflation model, which results in these changes in values as compared to previous models.

While the count model coefficients seem to show that the model is affected the most by the attribute “railstops”, the Zero-inflated negative binomial model is a two-part model. This means that the coefficients from both the count model and zero inflation model are responsible for the overall predictions being made by the model. So, the coefficients by themselves only explain the effect on one part of the model, because the second part of the model strongly affects the significance of other attributes.

Since the z-values measure the distance between the data point and the mean using the standard deviation, z scores can have a positive or negative sign. This depends if the z-value is higher or lower than the mean. This can be used to compare the value to the average. Below is the comparison of the z-values of the Poisson model and the Negative Binomial models.

Z-value comparison			
Attribute	**Negative Binomial model**	**Zero Inflated Negative Binomial model**	**Difference**
signal	7.973	7.452	0.521
tram	7.432	8.251	-0.819
railstop	7.686	8.768	-1.082
educatio	9.112	10.73	-1.618
biergarten	4.668	4.31	0.358
supermarkt	9.08	8.712	0.368
bicycle	12.913	8.114	4.799
primary	8.086	5.169	2.917
secondary	9.48	6.341	3.139
resident	27.983	8.777	19.206
Weighted Average Speed	10.508	7.33	3.178

Table 4.8 Z-values for Negative Binomial and Zero Inflated Negative Binomial model

As seen in the table above, the z-value for attributes is significantly higher for the Negative binomial model, as compared to the Zero-inflated negative binomial model. The third column shows the difference between the z-values for both models. As seen, the z-value

for the zero-inflated model is higher for only three variables “tram”, “railstop” and “education”. This also confirms that the Zero-inflated negative binomial model has lower standard errors as compared to the negative binomial model.

To test whether the ZINB model works better than the NB model, a chi-squared test statistic was calculated. This test was performed using the function “pchisq” in R (R Core Team, 2021), which is also known as the non-central Chi-Squared Distribution. By default, it calculates the left tailed probabilities, but by inserting the parameter “lower.tail” to be false, it can be adjusted. The difference of log-likelihoods and “df” or degrees of freedom is the difference between the independent parameters for the two models. Another way to obtain this p-value instead of adding the lower tail command is to subtract the value obtained from 1 as, “1- pchisq(“difference of log”, df)”. Between the ZINB and NB models, the p-value comes out to be 2.08E-10, which concludes that the ZINB model is a significant improvement over the NB model.

While the chi-squared test can be used to compare two models, the model selection should not be based solely on this test. For this reason, another test statistic that can be used to find a better model between the Negative binomial model and the zero-inflated negative binomial model was employed, which is the AIC test. Akaike's information criterion also known more commonly as AIC, is used to classify and test models amongst each other. While it does not suggest if the model is best overall, it does however rank the input models and can help identify the better model. AIC is suggested to be used for statistical regression models as it performs best with large datasets (Vrieze, 2012). The AIC can be derived from a model's likelihood function, the number of independent variables and the maximum likelihood estimate. (Akaike, 1974) gives the general equation for AIC as:

AIC = (-2)log(maximum likelihood) + 2 (number if independent model parameters)

The complexity of the model is indicated by the number of parameters used to estimate the model. While the true model, also known as the model used to generate the data is not part of the analysis, the AIC is efficient and will always choose the model that minimizes the mean squared error of the prediction (Vrieze, 2012).

The AIC can be calculated in R by simply running the command “AIC(model1, model2)“. The AIC value for the Negative binomial model was 14578.19, and for the Zero-inflated negative binomial model was 14528.55. As suggested by the literature, the Zero-inflated model is the ideal model when compared to the simple Negative binomial model since it has a lower AIC value.

4.5 Spatial Regression model

The first step in spatial regression is making the spatial weights, which influence the outcomes of spatial regression. Spatial weights are typically a positive matrix that specifies the relationship between two neighbours for each observation. In a spatial matrix a non-zero element “m_{ij}”, defines “j” as being a neighbour of “i”. And since an observation

cannot be a neighbour to itself, the weights for this diagonal element remain 0. Spatial weights are generally based on different types of contiguity, which may result in different weights for the same layout. can be formed with two methodologies, rook, and queen. For this analysis, the queen contiguity was used, which takes into account the neighbours connected via the vertices and the edges (Anselin, 2002). Queen contiguity was used for the analysis since it considers all the neighbours for assigning spatial weights. Since the grid used in the analysis shares no specific correlation or restrictions with each of their neighbours, using Queen contiguity is the logical step. A sample for the two types of contiguity is shown below:

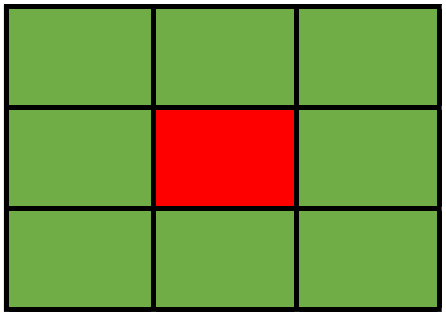


Figure 4.8 Example of Queen contiguity

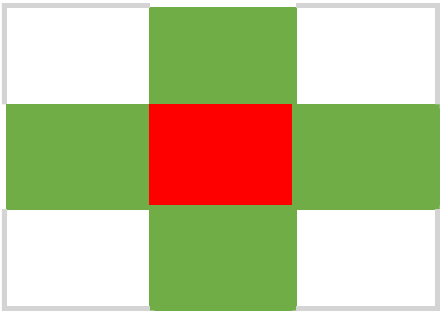


Figure 4.9 Example of Rook Contiguity

Figure 4.8 and Figure 4.9 show the pattern of Queen contiguity. Here the grid cell coloured in red is considered a neighbour with all the cells coloured in green, while for the rook contiguity, the cell central cell coloured in red is considered a neighbour with only the cells connected to it by the edge. This is the main difference between the different contiguity concepts used in the analysis of spatial regression models.

In r, this was done using two different commands, “poly2nb” and “nb2listw”, both are part of the “spdep” package in R. The “poly2nb” command is used to build a list of neighbours with contiguous boundaries from polygons and the “nb2list2” command is used to apply the weights to the neighbours based on its attributes and the chosen scheme (R. S. Bivand & Wong, 2018).

For this analysis two spatial models were considered, namely Spatial Autoregressive Model (SAR) also known as the Spatial lag model and the spatial error model. The Spa-

tial lag model is typically used when the dependent variable “y” is influenced by the values in its neighbouring units, while the spatial error model relies on the presence of spatial dependence in the error term of the neighbouring units (Saputro et al., 2019).

The equation for the spatial lag model and the Spatial error model can be referred in the book “Spatial econometrics: methods and models” and the notations are as mentioned below (Anselin, 1988).

Spatial lag model:

$$Y=\rho WY+X\beta+\epsilon$$

Where Y is the response variable, ρ(Rho) is the autoregression parameter which estimates the influence of the neighbouring units, W represents the spatial weights matrix, β is the vector representing the slopes for the predictors, and X is the predictor and ϵ is the error term.

Spatial Error model:

$$Y=X\beta+\lambda W\mu+\epsilon$$

Where Y is the response variable, λ is the autoregression coefficient, W represents the spatial weights matrix, and μ is the spatial error term.

The key terms for both the models to interpret are the ρ(Rho) and the λ(Lambda), these terms should be statistically significant for the models to determine the better model. Both the models were run with all attributes for the first iteration, but after discarding the insignificant attributes, the final model is as discussed below.

The Spatial lag model was run using the “lagsarlm” of the “spatialreg” package in R (R. Bivand & Piras, 2015). The summary for the spatial lag model is as shown below:

Call: lagsarlm(formula = reg3eq, data = spdata, listw = listw1)					
Residuals:	Min	1Q	Median	3Q	Max
	-4.51507	-0.286058	-0.031792	0.254658	11.29403
Type:	lag				
Coefficients:	(Asymptotic standard error)				
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.72E-02	5.90E-02	0.63	0.5287	
Weighted Average Speed	-8.21E-03	1.92E-03	-4.2688	1.97E-05	
Signal	1.13E-01	1.42E-02	7.9581	1.78E-15	
Railstop	5.90E-01	9.42E-02	6.2614	3.82E-10	
Crossings	3.15E-02	6.95E-03	4.541	5.60E-06	
Education	2.80E-01	5.54E-02	5.051	4.40E-07	
Biergarten	6.43E-01	1.34E-01	4.807	1.53E-06	
Supermarkt	3.64E-01	3.98E-02	9.1569	<2.22E-16	
Primary	7.67E-04	1.13E-04	6.8192	9.16E-12	
Secondary	4.89E-04	8.48E-05	5.7709	7.88E-09	
Footway	-8.88E-05	1.76E-05	-5.035	4.78E-07	
Rho: 0.68035,	LR test value	1205.5,	p-value:	<2.22E-16	
Asymptotic standard error:		0.016139			
	z-value:	42.155,	p-value:	<2.22E-16	
Wald statistic	1777,	p-value:	<2.22E-16		
Log likelihood:	-2814.031	for lag model			
ML residual variance (sigma squared):			1.2028,	(sigma: 1.0967)	
Number of observations:		1813			
Number of parameters estimated:			13		
AIC:	5654.1,	(AIC for lm:	6857.6)		
LM test for residual autocorrelation					
test value:	11.89,	p-value:	0.0005643		

Table 4.9 Output Summary for the Spatial Lag Model

The value of Rho which is the spatial lag parameter which shows how the neighbouring values of Y affect the primary value of Y, it has a positive effect and it is also statistically significant. Since in a spatial lag model, the value of Y depends also on the neighbors' value of Y and vice versa, the slope estimates of the model and their significance cannot be taken into consideration. Instead, there's a need to look at the impact of coefficients, both direct and indirect for each attribute which can be ran using the "impacts" command in R from the "spatialreg" package (R. Bivand & Piras, 2015). This is shown in table below:

Attributes	Direct	Indirect	Total
Weighted Average Speed	-9.05E-03	-0.016618507	-0.02566942
Signal	1.25E-01	0.229317937	0.354211021
Railstop	6.51E-01	1.194981636	1.845802694
Crossings	3.48E-02	0.063878636	0.098668762
Education	3.09E-01	0.566594665	0.875178268
Biergarten	7.09E-01	1.302610538	2.012049365
Supermarkt	4.02E-01	0.737814767	1.139649718
Primary	8.46E-04	0.001554217	0.002400688
Secondary	5.40E-04	0.000990774	0.001530378
Footway	-9.80E-05	-0.000179903	-0.000277884

Table 4.10 Direct, Indirect and Total impact of coefficients

Here, the “Direct” values show the effect on the variable Y for a grid cell “x”, if there is an increase of attributes by 1 for each individual attribute. The “Indirect” values, show the effect on Y for a grid cell number “x” if the neighbours increase their value by 1 for each individual attribute, which can also be considered as the effect on neighbours if there is a change in the value of Y for the grid cell “x”. The total values are the combined effect, which is used for analysis. Printing the summary of these effects, we obtain.

Simulated p-values:			
	Direct	**Indirect**	**Total**
Weighted Average Speed	3.09E-05	7.39E-05	4.83E-05
Signal	6.66E-16	6.75E-14	2.44E-15
Railstop	5.88E-10	5.96E-09	1.46E-09
Crossings	9.58E-06	1.76E-05	1.17E-05
Education	9.81E-08	8.00E-08	5.52E-08
Biergarten	1.37E-06	4.42E-06	2.30E-06
Supermarkt	<2.22E-16	<2.22E-16	<2.22E-16
Primary	2.35E-11	2.10E-09	2.43E-10
Secondary	1.03E-08	1.24E-07	3.50E-08
Footway	7.22E-08	1.76E-06	4.78E-07

Table 4.11 Simulated P values for the Spatial lag model

As seen above, the simulated p values as significant for all the selected attributes. The model summary also compares the spatial regression model to a linear regression model, which is run for the same parameters as the spatial lag model. It can be observed that the AIC value for the spatial lag model was 5654.1, and for the linear regression model was 6857.6, which suggests that statistically, the spatial lag model performs better than the linear regression model.

Next was the Spatial error model. This model was executed using the “errorsarlm” of the “spatialreg” package in R (R. Bivand & Piras, 2015). Below is the summary for the same:

Call: errorsarlm(formula = reg4eq, data= spdata, listw = listw1)					
Residuals:					
	Min	1Q	Median	3Q	Max
	-4.2605	-0.32556	-0.02176	0.21937	11.61444
Type:	error				
Coefficients:	(asymptotic standard errors)				
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.92E-01	1.71E-01	1.1212	0.2621875	
Signal	1.08E-01	1.44E-02	7.5284	5.13E-14	
Railstop	4.45E-01	8.74E-02	5.0913	3.56E-07	
Crossings	3.90E-02	7.25E-03	5.385	7.25E-08	
Education	2.19E-01	5.32E-02	4.1231	3.74E-05	
Biergarten	4.40E-01	1.26E-01	3.5687	0.0003588	
Supermarkt	3.20E-01	3.81E-02	8.398	<2.22E-16	
Primary	9.35E-04	1.20E-04	7.7734	7.55E-15	
Secondary	6.11E-04	8.87E-05	6.8833	5.85E-12	
Resident	1.97E-04	3.64E-05	5.4141	6.16E-08	
Lambda:	0.84438,	LR test value:	1083.3,	p-value:	<2.22E-16
Asymptotic	standard	error:	0.016241		
	z-value:	51.99,	p-value:	<2.22E-16	
Wald statistic	2702.9,		p-value:	<2.22E-16	
Log likelihood	-2864.297	for error model			
ML residual variance (sigma squared):		1.1905,	(sigma:	1.0911)	
Number of observations:	1813				
Number of parameters estimated:		12			
AIC:	5752.6,	(AIC for lm:	6833.9)		

Table 4.12 Output Summary for the Spatial Error model

Unlike the spatial lag model, the estimates for the Spatial error model (SEM) can be interpreted directly as marginal effects. First, the Lambda value is a positive 0.84438, which is statistically significant. Here the coefficient “railstop” shows a very strong effect on the number of crashes, which suggests that the more rail stops in a grid cell, increases the number of crashes by 0.445. This could be explained by the movement of people which results due to use of public transport and the movement towards the stops. Since there are no demand parameters in the model, there is just an assumption and cannot be verified. The spatial error model also generates the AIC values, as compared to the linear regression model. Here the SEM model shows AIC value of 5752.6 as compared to the AIC value of 6833.9 for the linear regression model. This suggests that the SEM model performs better than the regression model.

While the literature suggests that the SEM performs better than the SAR model in most cases (Rhee et al., 2016) (Jia et al., 2018), the comparison of both models suggests that the spatial lag model performed better as compared to the spatial error model. This was concluded based on the AIC values, which were 5748.962 for the spatial error model and 5654.1 for the spatial lag model, and the model with lower AIC values is considered to be a better fit. The next step was to check how well the models predict the values. This was done using the predict function, both the models were used to predict values and there were several inconsistencies detected. For the spatial lag model, many negative values were observed. And for both spatial error and spatial lag models, the maximum values were 16.2 and 9.1, which are considerably low as compared to the highest observed value. This suggests that both the spatial models do not function properly for the given dataset, although the spatial lag model gives a better output as compared to the Spatial error model which is due to a better statistical fit.

When looking at the grid cell numbers for the highest predicted crashes, they show similar results that the highest predicted crash value was the same as the observed crashes. This along with the presence of negative values suggests that the model requires more spatial information before making any interpretations.

4.6 Model Validation for the year 2020

For predictions, a different dataset was used which belonged to the year 2020. Since the zero-inflated negative binomial model gives the best statistical fit for the tested data, the same model was used to predict the number of crashes for the year 2020. The dataset used for predicting the values consists of all the same attributes that were used to train the data.

The prediction results, plotted against the actual number of crashes and a histogram for the predicted values are shown below:

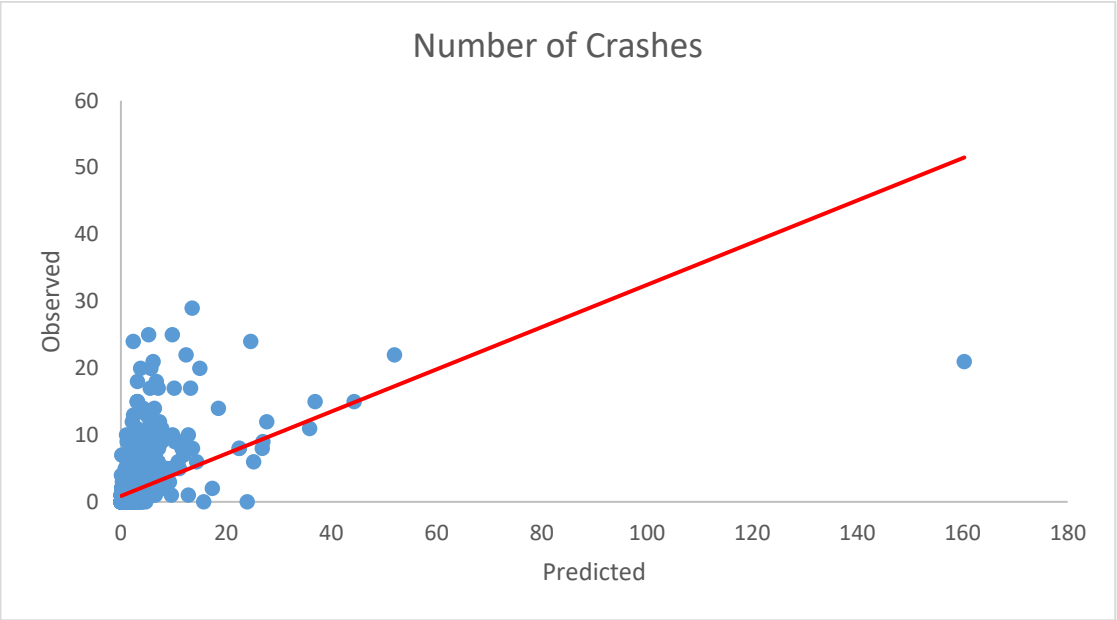


Figure 4.10 Predicted and Observed values using the Zero Inflated Negative Binomial model for the year 2020

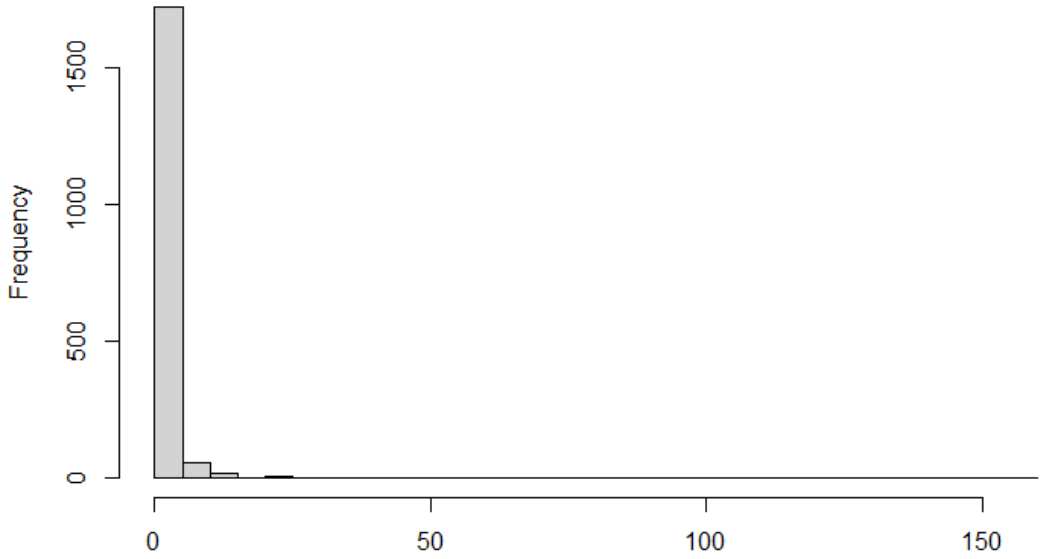


Figure 4.11 Histogram for the predicted values

As seen in both the plot and the histogram, no negative values are being predicted by the model, which should hold since the count model is a negative binomial model. The first observation made for the set of predicted values was that there is an outlier that was predicted by the model which suggests that 156.74 crashes may happen at a single location in a year. But this seems extremely high when compared to the observed data for which the highest value is 29. The histogram shows that most of the predicted crash values are saturated between 0 and 30, with almost no values from 50 to 160. So, we know that statistically, the model predicted a significantly high outlier in one instance.

So, the next step was to visualize the predicted data in QGIS, to conclude whether the location of this predicted crash point was also wrong or if this was a prediction error in the model. First, the actual crash data for 2020 was visualized to check the location of crash points. The grid visualized can be seen below:

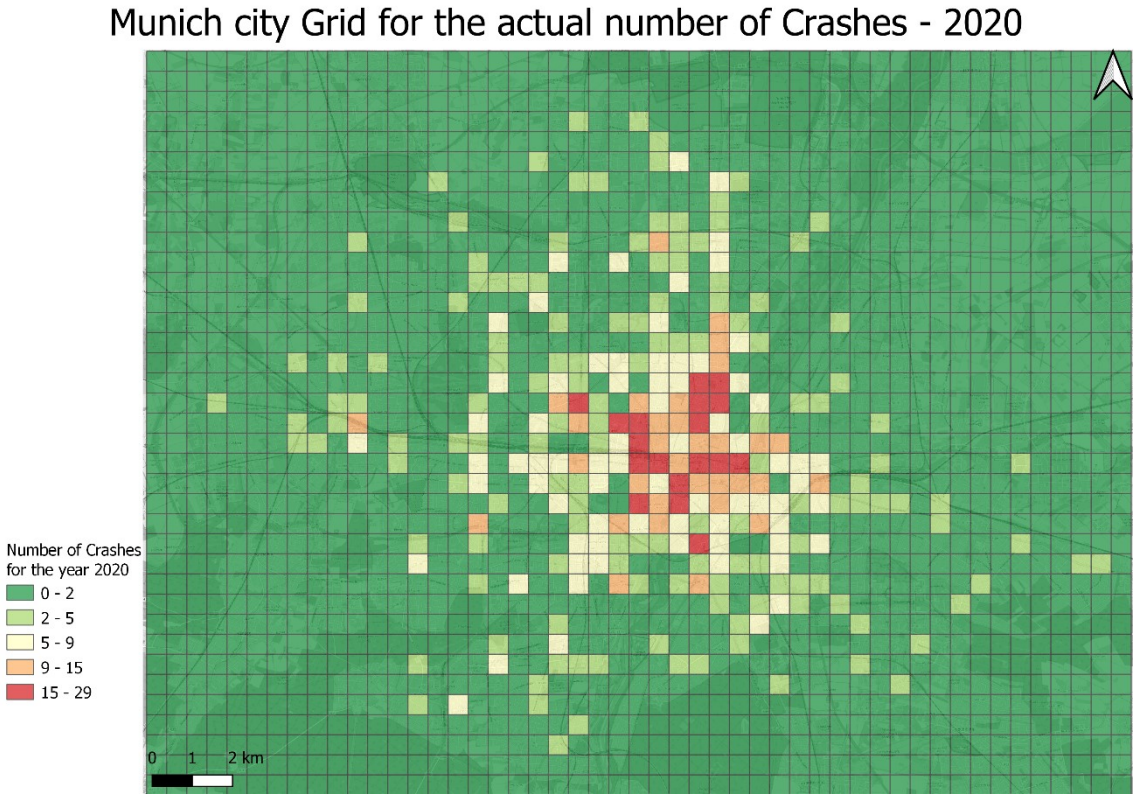


Figure 4.12 Munich city grid with observed crashes for the year 2020

As seen above, most of the high number of crashes are concentrated towards the centre of the grid, which also corresponds to the city centre of Munich. Then the grid cell numbers were identified to check their respective corresponding locations in the city of Munich, and the data was checked in QGIS. The cells with more than 15 number of crashes are shown below:

Cell ID	Number of Crashes	Cell ID	Number of Crashes
907	17	**946**	21
1016	17	**985**	21
1018	17	**909**	22
1024	17	**984**	22
870	18	**795**	24
1017	18	**1020**	24
908	20	**911**	25
1053	20	**1094**	25
1054	20	**1057**	29

Table 4.13 Observed crashes(more than 15 crashes per grid cell)

Next, the predicted number of crashes was visualized in QGIS. The data was imported to QGIS and joined to the existing grid dataset, using the common ID variable. The output is as shown below:

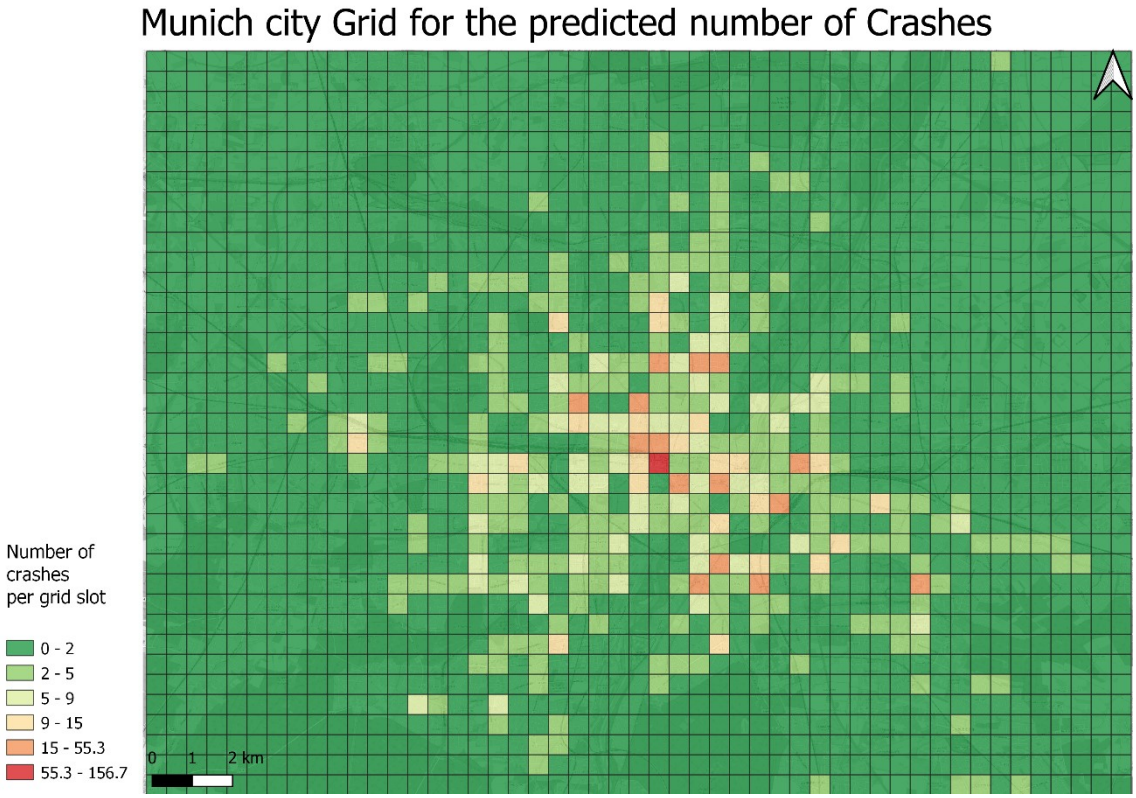


Figure 4.13 Munich city grid with predicted crashes for the year 2020

The predicted crashes are well distributed around the city. The highest number of crashes is observed in the grid cell number 946 and which is located between the central station of Munich and Karlsplatz which is to the east of the central station. The grid cells corresponding to the crash with the number of crashes more than 15 are as shown in the table below:

Cell ID	Number of Crashes	Cell ID	Number of Crashes
908	15.042	**1015**	25.632
1433	17.343	**1026**	27.261
906	18.484	**1062**	27.428
941	20.658	**945**	36.454
1170	23.721	**1058**	37.981
795	24.603	**1052**	43.263
1205	24.839	**984**	55.25
1137	24.904	**946**	156.743

Figure 4.14 Predicted crashes(more than 15 crashes per grid cell)

Taking a look at predicted values, it was observed that one point shows a significantly high number of crashes. All the attributes for ZINB had a positive sign, indicating that they had a positive correlation with the dependent variable. This is further corroborated by looking at the observed and predicted crash data and zones with a high number of crashes. For example, when looking at grid cell number 946, it can be noticed that this cell has mixed traffic which includes tram line, car traffic and bicycle traffic all one the

same lane, although this traffic is uni-directional, the tram movement is in both directions. This was observed in other grid cells with a high number of crashes as well, although this was confirmed by visual inspection and satellite images, as there is no data available that classifies the streets according to their right of way. Grid cell number 946 is shown below:

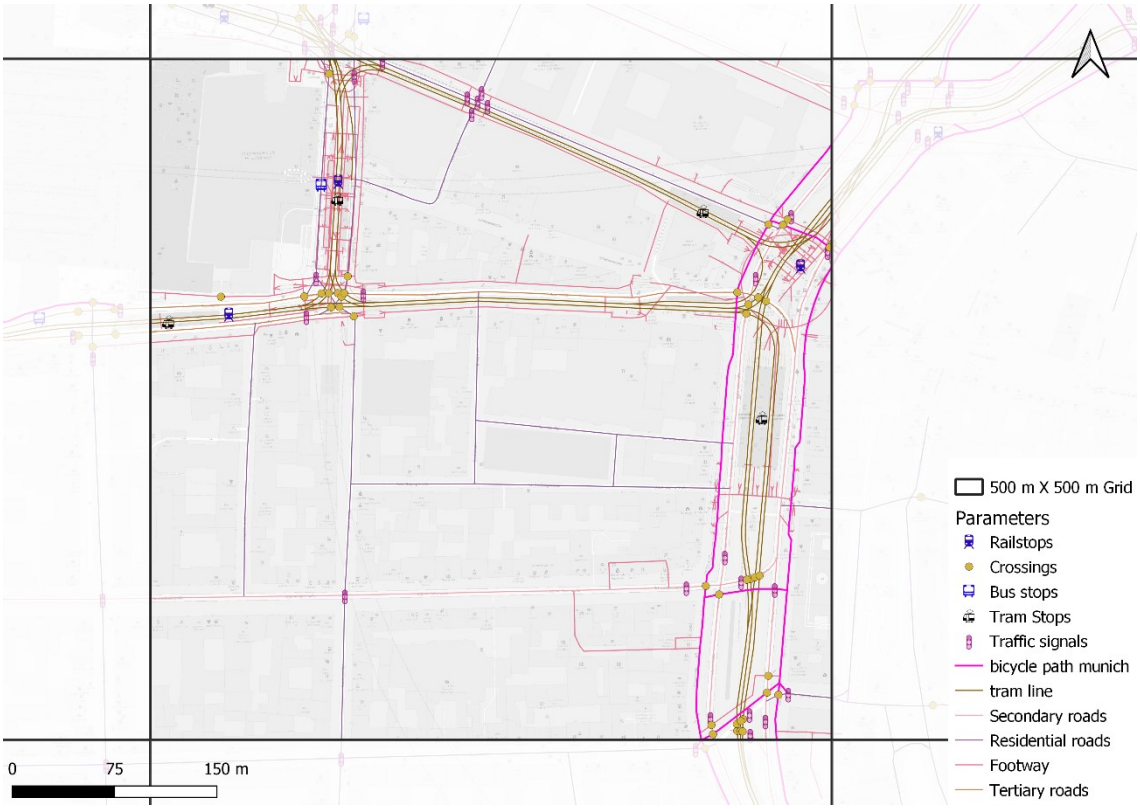


Figure 4.15 Grid cell number 946 with all the infrastructural parameters

While there is an obvious difference between the observed and predicted values, it is crucial to determine where the model is over or under predicting the crashes. This can be done by visualizing the difference between the two. To visualize the difference between predicted crashes and actual crashes over each grid cell, both the data were merged using the “ID” variable and the sum was calculated in a new column.

The difference is plotted on the grid is shown below:

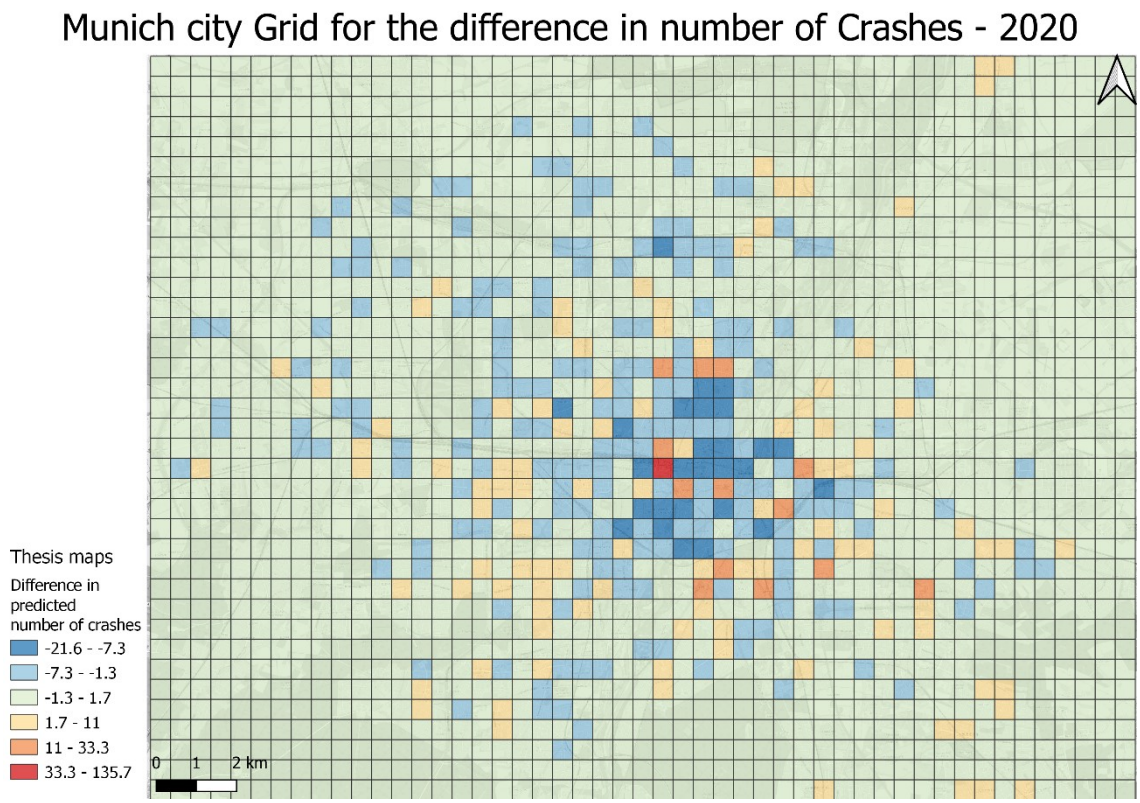


Figure 4.16 Munich city grid with the difference between the observed and predicted number of crashes

As seen in the image above, the colour scheme for the map shows the grid cells for which the crashes were underpredicted in shades of blue, and the grid cells for which the crashes were overpredicted in the shades of red. Although the predicted values show that many values were significantly underpredicted, the Zero-inflated Negative Binomial model predicted a total of 2467.954 crashes, as compared to 2350 observed crashes for the year 2020. This suggests that the model predicted a total of 117.9 additional crashes. This could be attributed to the outlier in the predicted values and the fact that the ZINB model does not predict any 0 values, but values slightly greater than 0.

When the outlier of the highest number of crashes is considered an exception, it can be observed that the predicted values are considerably lower in most grid cells as compared to the observed values. One of the reasons for this is the lack of additional variables that were removed to improve the statistical fit of the model. Another reason could be attributed to the year 2020 itself. It must be mentioned that the year 2020 had special circumstances due to the global Coronavirus pandemic. The implementation of lockdowns and restricted movement of people, saw a significant change in the use of vehicles, while there was a fall in the use of cars, there was an increase in the use of bicycles (Möllers et al., 2021) (Schweizer et al., 2021). This could be recognized as one of the reasons for the difference in the predicted and observed values.

Most of these grid cells correspond to the city centre of the city of Munich, but for a better classification of these cells, the grid cells were matched with the district map for the city of Munich. Using the “join locations by attribute” command in QGIS, the grid map and

the district maps were joined together and the crash values were summed up together. The grid cells which belong to the number of crash groups 9-15 and 15-29 are located close to the city centre. The resultant map with the number of crashes for individual districts is shown below.

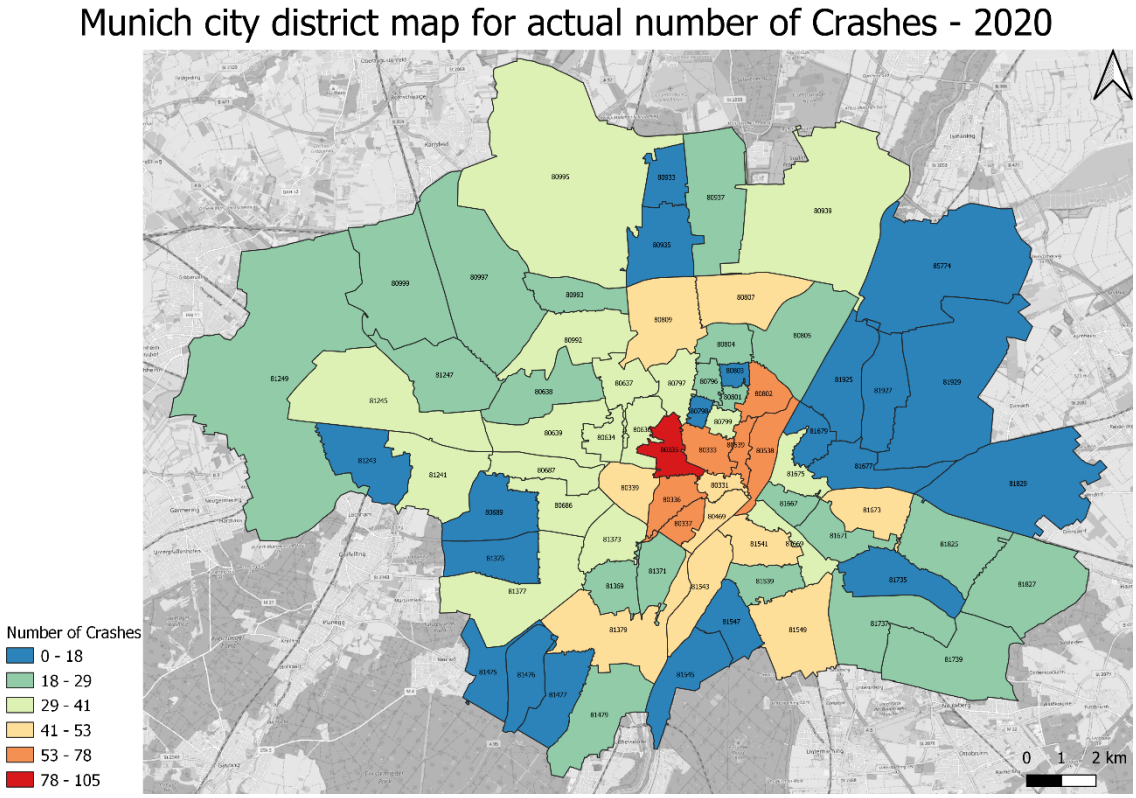


Figure 4.17 Munich city district map for observed crashes in the year 2020

Since the districts are much larger than the individual grid cells, the data group is increased further and results are displayed on a bigger scale. The districts identified with high number of crashes are Altstadt-Lehel, Ludwigsvorstadt-Isarvorstadt, Maxvorstadt, Sendling, Schwabing-West and Schwabing-Freimann. All of these districts are closely associated with the city centre of the city of Munich which is located in Altstadt Lehel.

These grid cell IDs were used to identify the locations in the city of Munich, and to consider the possible applicable solutions. Again, the predicted crash grid was combined with the district map of Munich to obtain the predicted number of crashes per district in Munich. The number of predicted crashes was visualized as shown below:

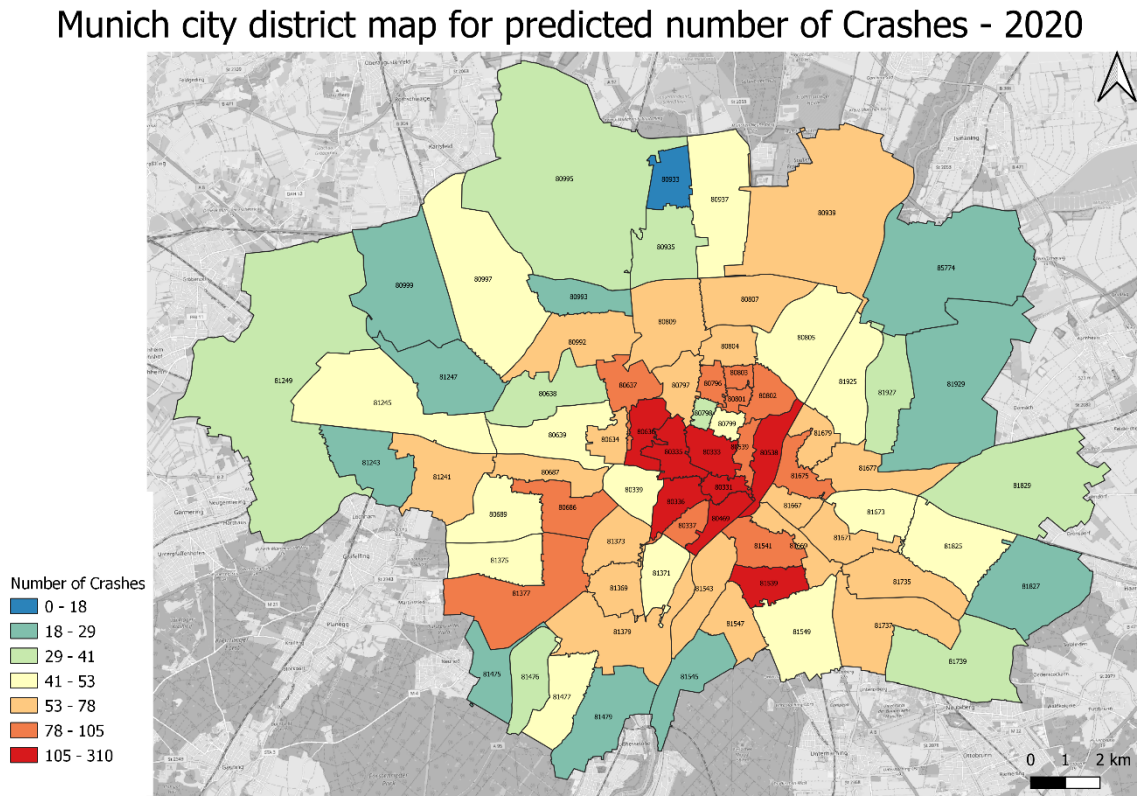


Figure 4.18 Munich city district map for predicted crashes in the year 2020

Upon looking closely at the predicted crash data for 2020, it was observed that most of the districts with high number of crashes are located close to the city centre. The districts were Altstadt-Lehel, Maxvorstadt, Ludwigsvorstadt-Isarvorstadt, Schwabing-Freimann, Schwabing West, and Au Haidhausen. While this corresponds to the actual crashes in the year 2020, there are many additional crash locations predicted. Some of these could be attributed to the fact that the model does not predict any 0 values but values more than 0, which could add up while looking at aggregated data. Another reason is that during the data analysis, the data for speed was worked upon and averaged to use in the model. As discussed before, one of the drawbacks of looking at district level aggregated maps is the smoothening of data, as this could misrepresent the severity of individual junctions. But a map like this helps understand the total number of crashes occurring over different districts which could help classify them if needed.

The highest crash value for the predicted values is 156.743, which occurs in the grid cell number 946 which overlaps with the districts of Altstadt-Lehel and Ludwigsvorstadt-Isarvorstadt. As seen in Table 4.7, the standardized coefficients for the ZINB model, “railstop” had the strongest influence on the dependent variable and the highest number of rail stops also belong to the grid cell number 946 and which corresponds to the section of the city, between Karsplatz and Hauptbahnhof. While the highest number of crashes for the observed crash data and the predicted number of crashes overlap with the district of Altstadt-Lehel and Maxvorstadt. The districts of Altstadt-Lehel and Ludwigsvorstadt-Isarvorstadt are neighbours to each other. While for the predicted crashes the value was significantly high, the value corresponds to the location for the observed high number of

crashes. This suggests that the district of Altstadt-Lehel and its surroundings have a set of characteristics which result in a high number of crashes. ZINB model also suggests a strong effect of educational institutes on crashes, while for the observed crashes it can be seen that the crash numbers were lower in the districts with universities, which could be due to the shift to online learning during the Coronavirus pandemic. The next coefficient with a strong effect is tram stops. Most tram stops are located on the street, and while there are some sections of the city with a separated space for trams, in the zones with a high number of crashes, this rarely seems to be the case.

5 Applications

While to reduce the number of crashes in a city, multiple solutions can be implemented, which range from junction specific to area-wide solutions. (Deliali et al., 2021) conducted a study to analyze the effect of separated and merged bicycle lanes and how they affect the number of accidents between car drivers and bicycle users. They concluded that the bike lanes which are located between the footpath and parking lane reduce the driver's ability to detect a cyclist. While, this could be a possible solution, to model this there is a need for data for parking lanes and the identifier for whether there is segregation or not. With this metric, the proposed model could analyze a solution like this practically, by calculating the length of each lane in individual grid cells, along with the total length of the road and adding it to the data for the regression model. At the same time, the ratio of bicycle streets as compared to all streets in a grid cell can help us look at the lack of bicycle lanes in grid cells with a high number of crashes. The data suggest that the grid cell with the highest number of crashes had 9% of the total length of streets in the grid cell as a bicycle lane. It was also observed that several grid cells with a high number of bicycle crashes share the road with motor vehicles and trams. A possible solution to address this is providing bicycle lanes in places with mixed rights of way. While this is difficult to execute in reality, as most streets require parking, a time-based parking system on such streets can help address this issue. Since there is no data available on open-source platforms that can classify the streets based on their right of way, this cannot be modelled.

To address the issue of a high number of bicycle crashes in the inner districts of the city of Munich, one possible solution was modelled using the Zero-inflated model. This was done by changing the specific attributes for the grid cells in a prediction dataset that was used for model validation. An attempt was made to explore the idea of a car-free zone in the city centre, which corresponds to the district "Altstadt-Lehel". To do this, all the attributes that correspond to the cars and their movement were removed from the model, this included the signals and length of streets which coincided with the secondary and residential streets. The attributes for the length of these streets and the data for speed were also set to 0. Based on the assumption of car users for the model as mentioned in section 3.4, the model should assume that there are no car users in these districts but only pedestrians and bicycle users. When the new number of crashes was predicted using the modified dataset, it resulted in the following number of crashes.

ID	New Predicted Crashes	Observed Crashes
982	1.005106	8
983	0.180527	15
984	2.247176	22
1020	0.572297	24
1021	0.642743	12
1057	0.741086	29
1058	2.305879	15

Table 5.1 Observed and Predicted number of crashes for car-free zones

As seen above, the crashes reduce drastically in the corresponding grid cells. While the crashes do not go down to 0, there are still some crashes even after removing the infrastructure attributes primarily associated with motor vehicles. This shows that if the inner district of the city of Munich is made a car-free zone, the bicycle crashes should reduce considerably. However, such a drastic proposal needs to be justified more and the connectivity around this region can help support this proposal. There are multiple S-Bahn stops, U-Bahn stations, bus stops and a tram line passing through this district. So, connectivity still exists even with the loss of the car as a mode of transport.

While there may be some underlying cause for the crashes in these zones attributed to the driver behaviour, environmental factors or other reasons, the predicted values seem to suggest that removing the cars addresses this problem. But such an extreme measure cannot be implemented without a proper study of the existing demand and users of bicycles and cars in the district of Altstadt-Lehel. So, this solution is proposed to be considered once the demand for this region has been included in the model along with more attributes that can incorporate area and drive behaviour characteristics.

6 Conclusion

The work done through the thesis first involved gathering the data through open-source platforms. There was a significant lack of data that could be used to analyze driver behaviour. But using the available resources, all the data was compiled in QGIS, which is a very useful software when it comes to visualizing spatial data and shapefiles.

The decision to use the grid for analysis proved to be useful and satisfactory in the sense that it helped quantify the infrastructure data at a macroscopic level. While the other option was to use the district boundaries for the city of Munich, an analysis of such a big area with the limited available data would have yielded poor results when analyzing it in any of the chosen models. A self-defined grid helped the models to make more focused predictions. While there is a possibility that a grid with larger defined zones can lead to more accurate results, this is due to data smoothening and not due to improved prediction of the model. The district maps were used but only for visual analysis and for concluding the districts with a high number of crashes in both observed and predicted datasets.

The literature review suggested that for analyzing large crash datasets, with a large number of zeros and count data, a Zero-inflated Negative Binomial model could be a good choice. The data was investigated for the excess zeroes, using the plots and statistical tests which confirmed the zero inflation and overdispersion and thus the direction of the modelling process. The comparison was done between the Linear regression model, Poisson model, Negative binomial (NB) model and a Zero-inflated Negative binomial (ZINB) model.

Through the analysis of different models, it was concluded that the ZINB model is the best model to analyze crashes among the selected set of models. The ZINB model took into account the overdispersion of data and also accounted for the excess number of zeros observed in the data which made the Poisson model and the negative binomial models unsuitable for analysis. While there were outliers in the prediction of crashes for the dataset for the year 2020, it fared much better than the other models. The location of this outlier also corresponded to the highest number of observed crash values as well.

Since the data consists of a large number of zeroes, to assume that the Zero-inflated Negative Binomial Model inherently shows a better fit to the model is incorrect, but based on the statistical tests like the likelihood ratio test and the AIC values, the ZINB model proves to be the best choice here. The AIC test applied to all the four models gives the following values and their results are as shown below:

Attributes	Value
Linear Regression model	29185.79
Poisson Model	16991.67
Negative Binomial model	14578.19
Zero-inflated Negative Binomial model	14528.55

Table 6.1 AIC values for all models

The AIC test for all four models also indicates that the Zero-inflated negative binomial model should be the best fit, compared to the other three models. This suggests that the zeroes in the data are generated from two processes. There are grid cells with unobserved crashes as well as grid cells with underreported crashes. This generally results from the data collection process. Literature suggests that the underreporting of crashes could result from many reasons like improper report filing for the incident, and not reporting the crash due to no or small injuries etc., which accounts for unobserved crashes in the dataset. The model shows that based on the data, a major influencing factor for bicycle crashes are motor vehicles, locations with a huge number of visitors and the shared right of way between motor vehicles and bicycles. This occurs in quite a few locations with high number of crashes, especially in the grid cells which correspond to the central districts. Other influencing factors for crashes were supermarkets, education institutes and beer gardens. This is likely because all three types of facilities attract people daily. Quite many students visit the universities regularly and supermarkets are frequented by everyone on a need basis, so the increasing number of crashes due to these facilities is rational.

While the difference in the predicted and observed values originates from two prime reasons, the first is the lack of additional data which comprises the driver behaviour attributes and environmental attributes and the second is the occurrence of the global Coronavirus pandemic in 2020 which increased the use of bicycles. This was also confirmed when the dataset was modified to model the central district to be car-free, even after removing the attributes accounting for motor vehicles, the model still predicted crashes in this section. Although this process was done only for 7 grid cells, it suggests the existence of underreporting of crashes and hence the idea of there being structural and sampling zeroes. This further supports the use of ZINB model.

The spatial models performed significantly poor as compared to the statistical models. A prime reason for this was the lack of variables that explained spatial relationships between the grid cells. While a comparative AIC value for both models prove them to be better than the regular linear regression model, their prediction values were significantly worse when compared to the ZINB model. The Spatial Lag model performed better than the spatial error model, as it displayed better prediction results as compared to the error model. While this was conducted as a provisional model, it showed potential to be explored more with the help of much more spatially rich data for better results.

After combining the predicted and observed crash values, it was concluded that the district of Altstadt-Lehel, Ludwigsvorstadt-Isarvorstadt and its surrounding neighbourhoods

are areas with the highest number of crashes. While this could be attributed to any number of reasons, based on the analysis of the data available, motor vehicles seem to be a big influential factor in these crashes.

6.1 Limitations

One of the major limitations of the thesis was the lack of demand data in the model. This means the lack of the key attributes like the movement of people from one region to the other, the number of cyclists and the number of motor vehicles, which would have helped quantify the crashes better. These attributes would also help get a better understanding of the number of crashes happening against the total number of users, which would help explain more details regarding the crashes. Like understanding whether the crashes are resulting due to driver behaviour, movement of people within different zones or due to poor infrastructure or lack thereof. Due to this, the model only takes into account the crashes happening due to infrastructural reasons. Although the model works under the assumption, that the existence of travel speed and bicycle infrastructure suggests the existence of car and bicycle users, it only addresses one aspect of the issue. The speed data was calculated in parts for assigning to the corresponding grid cell, which was not uniform. While the work done in QGIS included assigning the individual street segments to the corresponding grid cell, the process included some form of repetitions for the data. Since the length of road segments was not consistent, this resulted in some segments being repeated for adjacent grid cells. This was the cause of the speed average being higher in different cells on average. Although this was addressed to some extent by including the weights for length, the speed data not being accurate results in some inconsistency in the results like over prediction of crashes.

Another problem with the data gathering process was that the data was gathered from open-source platforms, so this data was only available for the year 2020 and not individually from 2016 through 2019. The data could not be back-dated with any reliable assumptions, as most of the infrastructure remained consistent in the city and there were no records of any changes for the same. This created a dataset with recurring infrastructure data for the period from 2016 - 2019, with the infrastructure data from the year 2020. While this dataset could not be treated as a combination of data for four different years, it was treated as just one dataset of grid cells where all the rows are treated as individual grid cells. There is also an issue of underreporting of crashes, which results in missing crashes thus resulting in excess zeroes. For the year 2020, motor vehicle movement saw a considerable reduction, due to reasons like lockdowns, or most people working from home. Another key variable that could have been considered is the length of the tram line per grid cell. During model Validation, it was observed that almost all grid cells with a high number of crashes observed, had a mixed right of way, typically classified as ROW-C. While, in the data search, no data was found on open-source platforms that could classify the streets based on their right of way, which could have proved to be significant in understanding the role of shared right of way and the number of crashes.

The spatial regression model, although theoretically superior has a big reliance on the input parameters. The spatial models require a substantial amount of data that could help explain the crashes and their relationship with neighbouring grid cells. While the regression seems simple in the sense that it allows for the model to consider the neighbour values and their attributes as weights, the relationship between neighbours must be explored more. The relationship may become further complicated by the use of even smaller grid structures for analysis since most spatial regression models studied dealt with larger areas.

6.2 Further Research

There is scope to improve the model further, this could be done by collecting more data from government agencies. Several other data that could explain more crash contributing factors like driver behaviour, environmental effects, vehicle condition etc., could help enhance this analysis further. One of the important things in scope would be to add the travel demand parameters in the model, which could help understand how people move in the city. This could result in a better understanding of how relatively safe or unsafe some parts of the city are and whether this is due to the movement of people or due to the presence of insufficient or unsafe infrastructure. A classification parameter can also be used to classify the whole city into different zones, like commercial or residential etc. Although such classification will require independent research as more often than not, modern cities are a mixture of different kinds of developments. Attributes like these could enhance the analysis of spatial models, which truly depend on spatially explorative data. Anything that can explain the landscape or geographical features better and help understand the reasoning behind the movement patterns, will be an ideal fit for spatial regression models.

The analysis should not be restricted to just statistical or mathematical models. There is a possibility to explore some newer analysis techniques which make use of self-learning algorithms based on machine learning. Although relatively new and the literature on such models are scarce, it could shine new light on the relationship between various spatial attributes of crash analysis.

List of References

Abdel-Salam, A., Guo, F., Flintsch, A., Arafeh, M., & Rakha, H. (2008). Linear regression crash prediction models. In *Efficient Transportation and Pavement Systems*. CRC Press. https://doi.org/10.1201/9780203881200.ch25

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

AL-Hasani, G., Asaduzzaman, M., & Soliman, A.-H. (2019, August). *Comparison of spatial regression models with Road Traffic Accidents Data r*. https://doi.org/10.11159/icsta19.31

Algeri, S., Aalbers, J., Morå, K. D., & Conrad, J. (2020). Searching for new phenomena with profile likelihood ratio tests. *Nature Reviews Physics*, *2*(5), 245–252. https://doi.org/10.1038/s42254-020-0169-5

Anselin, L. (1988). *Spatial Econometrics: Methods and Models* (1st ed., Vol. 4). Springer Netherlands. https://doi.org/10.1007/978-94-015-7799-1

Anselin, L. (2002). Under the hood Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*(3), 247–267. https://doi.org/10.1111/j.1574-0862.2002.tb00120.x

Bivand, R., & Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, *63*(18). https://doi.org/10.18637/jss.v063.i18

Bivand, R. S., & Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, *27*(3), 716–748. https://doi.org/10.1007/s11749-018-0599-x

Bundesministerium für Digitales und Verkehr (BMDV). (2022). *Nationaler Radverkehrsplan 3.0*. https://www.bmvi.de/SharedDocs/DE/Anlage/StV/nationaler-radverkehrsplan-3-0.pdf?__blob=publicationFile

Cai, Q., Abdel-Aty, M., Lee, J., & Eluru, N. (2017). Comparative analysis of zonal systems for macro-level crash modeling. *Journal of Safety Research*, *61*, 157–166. https://doi.org/10.1016/j.jsr.2017.02.018

Chiou, Y.-C., & Fu, C. (2013). Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. *Accident Analysis & Prevention*, *50*, 73–82. https://doi.org/10.1016/j.aap.2012.03.030

Deliali, K., Christofa, E., & Knodler Jr, M. (2021). The role of protected intersections in improving bicycle safety and driver right-turning behavior. *Accident Analysis & Prevention*, *159*, 106295. https://doi.org/10.1016/j.aap.2021.106295

Follmer, R., & Gruschwitz, D. (2019). *Mobility in Germany – short report*. *September*. www.mobilitaet-in-deutschland.de

Gao, X., Asami, Y., & Chung, C.-J. F. (2006). An empirical evaluation of spatial regression models. *Computers & Geosciences*, *32*(8), 1040–1051. https://doi.org/10.1016/j.cageo.2006.02.010

Garay, A. M., Hashimoto, E. M., Ortega, E. M. M., & Lachos, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models.

Computational Statistics & Data Analysis, *55*(3), 1304–1318. https://doi.org/10.1016/j.csda.2010.09.019

GEOFABRIK. (2022). *OpenStreetMap Data Extracts*. https://www.geofabrik.de/en/index.html

Hadi, M. A., Aruldhas, J., Chow, L.-F., & Wattleworth, J. A. (1995). ESTIMATING SAFETY EFFECTS OF CROSS-SECTION DESIGN FOR VARIOUS HIGHWAY TYPES USING NEGATIVE BINOMIAL REGRESSION. *Transportation Research Record: Journal of the Transportation Research Board*, 169–177. https://onlinepubs.trb.org/Onlinepubs/trr/1995/1500/1500-021.pdf

Hauer, E. (2001). Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. *Accident Analysis & Prevention*, *33*(6), 799–808. https://doi.org/10.1016/S0001-4575(00)00094-4

Imprialou, M., & Quddus, M. (2019). Crash data quality for road safety research: Current state and future directions. *Accident Analysis & Prevention*, *130*, 84–90. https://doi.org/10.1016/j.aap.2017.02.022

Jia, R., Khadka, A., & Kim, I. (2018). Traffic crash analysis with point-of-interest spatial clustering. *Accident Analysis & Prevention*, *121*, 223–230. https://doi.org/10.1016/j.aap.2018.09.018

Kim, K., Brunner, I. M., & Yamashita, E. Y. (2006). Influence of Land Use, Population, Employment, and Economic Activity on Accidents. *Transportation Research Record: Journal of the Transportation Research Board*, *1953*(1), 56–64. https://doi.org/10.1177/0361198106195300107

Kleiber, C., & Zeileis, A. (2008). *Applied Econometrics with R*. Springer-Verlag. https://cran.r-project.org/package=AER

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, *34*(1), 1. https://doi.org/10.2307/1269547

Landeshauptstadt München. (2022a). *Munich economy – key data*. https://stadt.muenchen.de/en/info/economic-data.html

Landeshauptstadt München. (2022b). *Munich Public Transport*. https://www.muenchen.de/int/en/traffic/public-transport.html

LeSage, J. P. (2005). Spatial Econometrics. In *Encyclopedia of Social Measurement* (pp. 613–619). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00343-1

LeSage, J. P. (2008). An Introduction to Spatial Econometrics. *Revue d'économie Industrielle*, *123*, 19–44. https://doi.org/10.4000/rei.3887

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001

Lord, D., Washington, S., & Ivan, J. N. (2007). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, *39*(1), 53–57. https://doi.org/10.1016/j.aap.2006.06.004

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, *37*(1), 35–46.

https://doi.org/10.1016/j.aap.2004.02.004

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/BM.2013.018

Medury, A., Grembek, O., Loukaitou-Sideris, A., & Shafizadeh, K. (2019). Investigating the underreporting of pedestrian and bicycle crashes in and around university campuses − a crowdsourcing approach. *Accident Analysis & Prevention*, *130*, 99–107. https://doi.org/10.1016/j.aap.2017.08.014

Miaou, S.-P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, *26*(4), 471–482. https://doi.org/10.1016/0001-4575(94)90038-8

Miler, M., Todić, F., & Ševrović, M. (2016). Extracting accurate location information from a highly inaccurate traffic accident dataset: A methodology based on a string matching technique. *Transportation Research Part C: Emerging Technologies*, *68*, 185–193. https://doi.org/10.1016/j.trc.2016.04.003

Möllers, A., Specht, S., & Wessel, J. (2021). *The impact of the Covid-19 pandemic and government intervention on active mobility* (No. 34). https://www.wiwi.uni-muenster.de/ivm/sites/ivm/files/documents/forschung/diskussionspapiere/workingpaper34.pdf

Nettleton, D. (2014). Selection of Variables and Factor Derivation. In *Commercial Data Mining* (pp. 79–104). Elsevier. https://doi.org/10.1016/B978-0-12-416602-8.00006-6

OpenStreetMap Wiki. (2022). *Main Page*. OpenStreetMap Wiki. https://wiki.openstreetmap.org/w/index.php?title=Main_Page&oldid=2301410

Park, B.-J., & Lord, D. (2009). Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention*, *41*(4), 683–691. https://doi.org/10.1016/j.aap.2009.03.007

Pew, T., Warr, R. L., Schultz, G. G., & Heaton, M. (2020). Justification for considering zero-inflated models in crash frequency analysis. *Transportation Research Interdisciplinary Perspectives*, *8*, 100249. https://doi.org/10.1016/j.trip.2020.100249

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Rhee, K.-A., Kim, J.-K., Lee, Y., & Ulfarsson, G. F. (2016). Spatial regression analysis of traffic crashes in Seoul. *Accident Analysis & Prevention*, *91*, 190–199. https://doi.org/10.1016/j.aap.2016.02.023

Saputro, D. R. S., Muhsinin, R. Y., Widyaningsih, P., & Sulistyaningsih. (2019). Spatial autoregressive with a spatial autoregressive error term model and its parameter estimation with two-stage generalized spatial least square procedure. *Journal of Physics: Conference Series*, *1217*(1), 012104. https://doi.org/10.1088/1742-6596/1217/1/012104

Schweizer, A.-M., Leiderer, A., Mitterwallner, V., Walentowitz, A., Mathes, G. H., & Steinbauer, M. J. (2021). Outdoor cycling activity affected by COVID-19 related epidemic-control-decisions. *PLOS ONE*, *16*(5), e0249268. https://doi.org/10.1371/journal.pone.0249268

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention*, *29*(6), 829–837. https://doi.org/10.1016/S0001-4575(97)00052-3

Shinar, D., Valero-Mora, P., van Strijp-Houtenbos, M., Haworth, N., Schramm, A., De Bruyne, G., Cavallo, V., Chliaoutakis, J., Dias, J., Ferraro, O. E., Fyhri, A., Sajatovic, A. H., Kuklane, K., Ledesma, R., Mascarell, O., Morandi, A., Muser, M., Otte, D., Papadakaki, M., … Tzamalouka, G. (2018). Under-reporting bicycle accidents to police in the COST TU1101 international survey: Cross-country comparisons and associated factors. *Accident Analysis & Prevention*, *110*, 177–186. https://doi.org/10.1016/j.aap.2017.09.018

Statistische Ämter des Bundes und der Länder. (2022). *Untfallatlas*. https://unfallatlas.statistikportal.de/_opendata2021.html

TOMTOM. (2022). *Road analytics*. https://www.tomtom.com/products/road-traffic-data-analytics/

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, *57*(2), 307. https://doi.org/10.2307/1912557

Washington, S., Karlaftis, M., Mannering, F., & Anastasopoulos, P. (2020). *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429244018

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*, *127*, 51–53. https://doi.org/10.1016/j.econlet.2014.12.029

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*, *27*(8). http://www.jstatsoft.org/v27/i08/

Zhu, F. (2012). Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *Journal of Mathematical Analysis and Applications*, *389*(1), 58–71. https://doi.org/10.1016/j.jmaa.2011.11.042

Zorn, E. (2010). *Radverkehr in München*. https://www.muenchen.de/rathaus/dam/jcr:9cd1c7ba-8032-4621-8db7-287bb16ffbae/Radverkehr-Muenchen-2007.pdf

List of Abbreviations

BMVI	Bundesministerium für Digitales und Verkehr informiert
TAD	Traffic analysis districts
NB	Negative binomial
ZIP	Zero inflated Poisson
ZINB	Zero inflated Negative binomial models
SAR	Spatial autoregressive model
SEM	Spatial error model
TAZ	Traffic Analysis zone
OLS	Ordinary Least Square
AIC	Akaike's Information Criterion

List of Figures

List of Tables

Declaration concerning the Master's Thesis

I hereby confirm that the presented thesis work has been done independently and using only the sources and resources as are listed. This thesis has not previously been submitted elsewhere for purposes of assessment.

Munich, May 15th, 2022

Pawan Shambhu Singh