Technical University of Munich – Professorship for Modeling Spatial Mobility
Prof. Dr.-Ing. Rolf Moeckel
Arcisstraße 21, 80333 München, www.msm.bgu.tum.de

# MASTER'S THESIS

# Analyzing and predicting road safety based on land use and road network properties. Case study of Hong Kong

Author:

**Róbert Kiss**

Supervision:

**Dr. Carlos Llorca Garcia**

**Prof. Dr. -Ing. Rolf Moeckel**

Date of Submission: 2017-10-10

## Abstract

As the world's population grows, so does transportation: our roads are more and more filled with traffic and this comes with a growing number of crashes. Because of this, safety measures become increasingly important in order to reduce the number of casualties in crashes.

Transportation Safety Planning is the method of planning cities with safety in mind. During the process, the future crash numbers are estimated based on the properties of the network and some other data, like estimated traffic volumes. However, in some lesser developed regions a problem arises: there is not enough data to precisely estimate future crash numbers with the conventional methods.

This thesis is looking for an answer for the following questions: is it really necessary to use all that data for crash estimation? Would it be possible to estimate crashes with an acceptable accuracy only based on land use and road properties? If so, how do these properties affect the crash occurrences?

In the process of finding out the answers for these questions, Hong Kong is used as an example. Hong Kong as one of the world's busiest cities is an ideal subject for an analysis like this. One year of crash data has been obtained from there and was plotted on a map together with land use and road properties in order to find out the relationship between these two variables and the crashes. Poisson and negative binomial models were used for the crash estimation, from which finally only the results of the Poisson-model were considered due to the unfitness of the negative binomial models.

According to the results, there is indeed a relationship between land uses, road properties and crashes, especially strong between the last two. Even though with some limitations, but this method can be used in any part of the world for crash estimation, and might be a good alternative for more complicated models requiring more data.

## Acknowledgements

First and foremost, I would like to thank Dr. Carlos Llorca Garcia, my supervisor for all the support he has given me throughout the process of writing my thesis. Starting from finding a topic through developing the models up until evaluating the results, he was never tired of answering my questions and leading me into the right direction. This work could not have been done without him.

Furthermore, I would also like to thank Dr. Tony Sze from the Department of Civil and Environmental Engineering of The Hong Kong Polytechnic University, who provided me the data of all injury crashes in Hong Kong from the year 2015. The data was obtained from the Transport Information System of the Hong Kong Transport Department. Dr. Sze has also volunteered to be the distant supervisor of my thesis, thus providing me local knowledge and guidance which I am highly grateful for.

# Table of Contents

# 1. Introduction

Traffic crashes are part of our lives. Every year nearly 1.3 million people die in road crashes, and an additional 20-50 million get injured or disabled (Association for Safe International Road Travel, 2017). When transportation has become necessary in almost everyone's life in order to fulfill daily tasks and function in life, it is of utmost importance to make it as safe as possible. Children walking to school, businessmen taking taxis to the airport, grandparents taking the bus home from the market: we have to make sure, that all of them get to their destinations safely.

There are several ways of making transportation safer. Introducing new regulations in order encourage people to drive slower through the cities; introducing congestion charge in order to decrease traffic in the inner areas of the city; or developing public transportation in order to provide people a good alternative for using their car. It is also possible to increase safety already on the design level: designing intersections with a safer layout, or using different kind of life-saving equipment along high-speed roadways and high-risk areas.

However, what if it would be possible to design safer cities already on the town-planning level? With today's constantly growing population, new residential areas are planned and built all around the world on a daily basis. Space is scarce in cities, therefore they have to grow, not just vertically and in density, but also area-wise. This might not be the best solution because of the longer commutes that result from more and more widespread cities, but it is a necessary part of dealing with the growing population of cities.

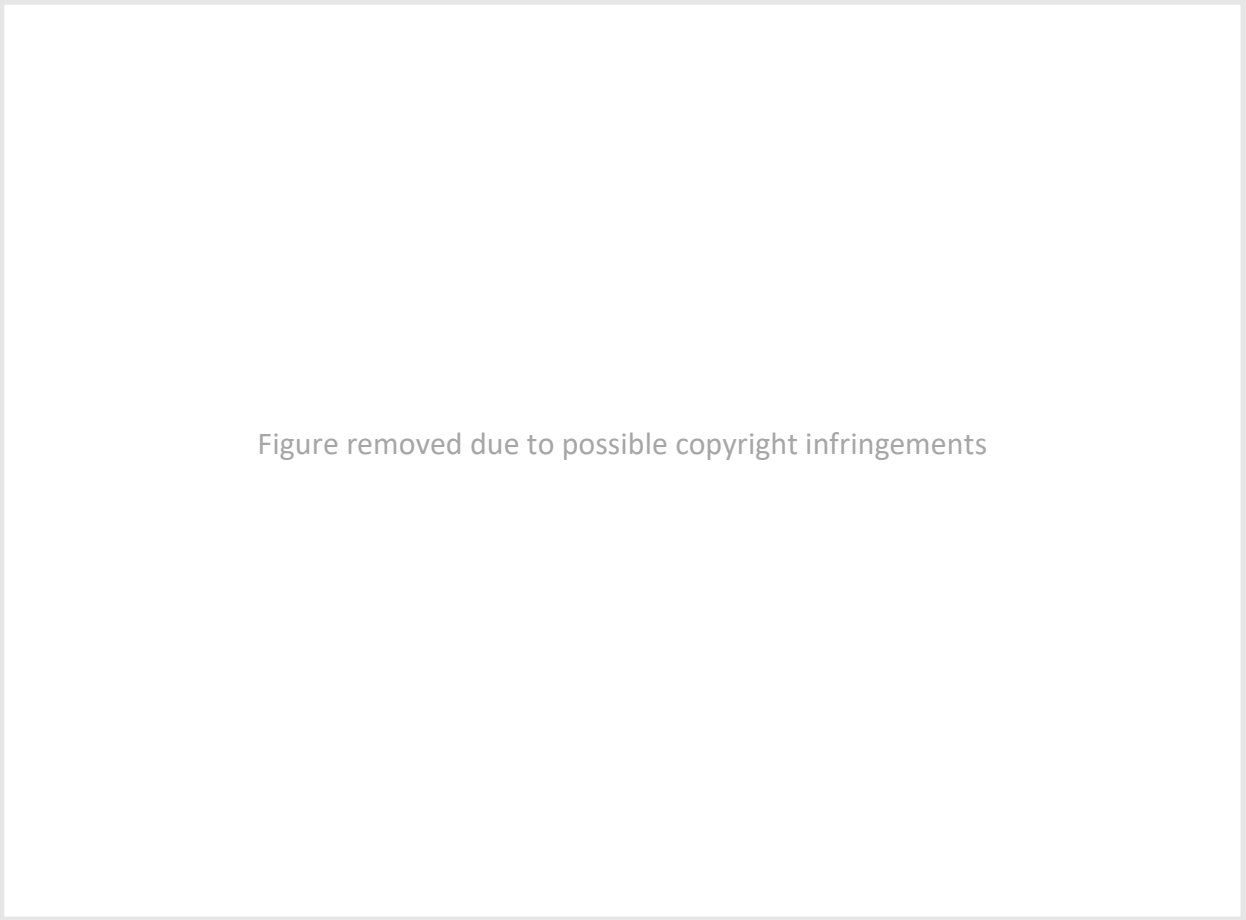
If there would be a model that can tell us the relation between the number and severity of crashes and some geographical properties related to town-planning, like land-use and road network properties, we would be able to design our cities in a safer way already from the beginning. In fact, this is already a known phenomenon: it is called Transportation Safety Planning (TSP).

TSP is the method of integrating safety into the transportation planning stage with a comprehensive, multimodal and data-driven approach, in order to reduce transportation fatalities and injuries within the new development (US Department of Transportation, 2017). This is very much in line with current thesis' topic: designing new residential areas consciously with safety in mind, using road types and land uses that make an area safer in comparison to others. In order to do this, a model is needed to estimate the future crashes. This is what present study is trying to come up with.

As the subject of the crash analysis Hong Kong has been chosen, because of the good availability of crash, land use and road network data, and because of the diversity of these factors. Hong Kong is a Special Administrative Region of the People's Republic of China, it is located on the south-eastern tip of China, being bordered from the north by the Guangdong province of China

and being surrounded with the South China Sea from the other three sides. In addition to this, Hong Kong has a very mountainous landscape, making only a small percentage of the area of Hong Kong suitable to build residential areas. For this reason, Hong Kong's population of 7.31 million people (according to 2015 data) lives on less than 25% of the total 1105.7 square km area of Hong Kong. This accounts for a population density of approximately 26,485 people per square km if we only consider the developed land areas, which is amongst the highest in the world (GovHK, 2017).

**Figure 1: The map of Hong Kong (Source: maps.google.com)**




In addition to the high traffic volumes generated by the sheer number of people living in the area and the population density, Hong Kong's mountainous area also results in narrow and curvy roads, which can also account for a high number of traffic crashes in the area.

All in all, the previously mentioned properties make Hong Kong ideal an ideal subject of this analysis.

According to the crash data acquired from the Transport Information System of the Hong Kong Transport Department, 16170 crashes have happened in Hong Kong through the year of 2015.

The crash database also contains coordinates, which makes it possible to plot the crashes on a map in order to analyze them. This way it is possible to assign crashes to specific land use and road network properties, in order to come up with a model that predicts the number and severity of crashes based on these properties.

The main objective of developing this model is twofold: firstly, the goal is to describe reality in order to see which of the variables have the most significant effect on crashes. In this analysis no traffic volume data will be used, only land use and road network properties, therefore it will be interesting to see how close are the results to other models that do use traffic volume data. If the results are not very far off, that means that traffic volumes might not be that important factors in crash prediction as we previously thought.

Secondly, if the results are not very far from one another, that will mean that it is possible to predict traffic safety without using traffic volumes. This can come very useful in the cases of some less developed locations, where traffic volume data might not be available. In these places when planning new areas or even after just building some roads, it will be possible to predict the safety of the new network with this model, even if there is not much data available.

In comparison to other models that use much more variables, using this model is much faster and less bothersome, therefore in cases when there is either not enough data or not enough time (and a less precise result is also acceptable), this model can be used instead of the more complicated ones. Developing this model is the objective of this thesis.

# 2. Literature review

Throughout the last few decades extensive research has been done in the field of crash analysis, however, in most cases the spatial factors have been more or less ignored. This is mostly due to the limited possibilities of computer analysis in the earlier days. Since then, computer analysis has advanced, and the recent developments in spatial modeling have enabled researchers to explore the spatial correlations of crashes. (Yazdani-Charati et al., 2014)

This thesis is focusing on creating a crash estimation model using some spatial variables in order to be able to use the results when planning new areas. The act of transportation planning with safety in mind is a known phenomenon, it is called Transportation Safety Planning or TSP.

C. Siddiqui (2012) has already explored the possibilities of TSP in his doctoral dissertation. In this study, the suitability of the current traffic-related zoning planning process is examined on the example of West Central Florida. To do this, several research objectives are being investigated, including exploring the existing key determinants in traditional transportation planning (trip generation/distribution data, land use types, demographics, etc.) in order to develop an effective and efficient TSP framework.

The analysis of crash data was performed using nonparametric approaches, classical statistical methods and Bayesian statistical techniques. The most important variables were determined by using nonparametric statistical techniques with different trip related variables and traffic related factors. The significance of spatial autocorrelation in crashes was also investigated, which was something new in comparison to earlier studies.

Motor vehicle crashes were classified as on-system (higher speed limits, traffic from different TAZs) and off-system (local roads with low speed limits) crashes. It was found that crashes occurring on on-system are more influenced by roadway and traffic related factors than off-system crashes. Therefore, for on-system crashes all other variables were disregarded, while for off-system crashes all zonal variables were considered.

After the analysis it was found that the land-use types “industrial” and “retail/office” have a positive association with the amount of on-system crashes, while the land-use types “kindergarten or school” and “urban” have the same association with the amount of off-system crashes (Siddiqui, 2012).

Because the purpose of this thesis is similar to the above work (namely that we would like to develop an estimation model than can be used to help TSP), we can use this paper to identify and address the issues that we might encounter during the research, one by one.

## 2.1. Modelling techniques

First and foremost, it is very important to find out which modelling technique is the best to use for the estimation. Many previous research has been done in this topic, all trying to determine which model fits best for crash analysis.

Two types of generalized linear models (GLM): Poisson regression models and negative binomial regression models are widely used in traffic crash analysis. Poisson distribution is ideal to describe traffic accidents because of their rare occurrence. However, when crashes are analyzed on a grid-network, many of the grid cells can have zero amount of crashes, which leads to overdispersion. In these cases Poisson-distribution might not be the best fit for the data, since it is not very good at handling data with excessive zero counts. To solve this problem, a possible solution is to use negative binomial regression, or the zero-inflated versions of the previously mentioned two models. (Songpatanasilp et al., 2015)

Because of the same reasons, many studies use negative binomial regression models in their analysis. These kind of planning-level crash estimation models are feasible and should be used in Transportation Safety Planning (De Guevara et al., 2004).

That being said, negative binomial regression models can in many cases represent a better fit to crash analysis than Poisson distribution models. They are normally used to establish the relationship between crashes and contributing factors. It is convenient to use them because they are capable of taking uncorrelated heterogeneity into account. However, they might not be able to do the same with spatial correlation. This is the reason why most advanced analyses use Bayesian models. They are capable of doing the same as negative binomial regression models, but in the meantime they are also competent when it comes to modeling spatial correlation, therefore they are considered a better device for crash estimation than the previous models. (Quddus, 2008)

Bayesian analysis is a statistical method which estimates the parameters of a hidden distribution based on the observed one. It starts with a prior distribution that might be based on anything and it is commonly assumed to be a uniform distribution. Based on this the likelihood of the observed distribution needs to be calculated as a function of the parameter values. This function will be multiplied with the prior distribution and normalized in order to acquire unit probability. This is called the posterior distribution. Then the mode of the distribution becomes the parameter estimate, and the probability intervals can be calculated using standard procedures (Weisstein, 2017).

Even though many studies consider them superior to the standard GLM procedures, it is not that obvious that Bayesian models are always a much better fit than traditional models. J. Aguero-Valverde and P. P. Jovanis (2006) tried to estimate the annual county-level crash frequency in

Pennsylvania, and compared the results when using a full Bayesian hierarchical model in comparison to traditional negative binomial estimates. The variables they used were socio-demographics, weather conditions, transportation infrastructure and amount of travel. They found that the estimates were similar for both models, even though the variables that proved to be significant varied between them (Aguero-Valverde & Jovanis, 2006).

Apart from using different, more developed regression models like the Bayesian hierarchical model, sometimes modifying the original GLM can also prove to be useful. Using the zero-inflated versions of GLMs was one example for this, but some researchers have tried out a different way. In a first study they used a conventional GLM approach with the assumption of a negative binomial error structure (Hadayeghi et al., 2007), and then later in another study they used both negative binomial and Poisson regression models. In this second study, the accuracy of GLMs was compared to that of geographically weighted Poisson regression models. The findings showed that the GWPR models generally perform better than both of the GLMs (Hadayeghi et al., 2010).

Some studies also try to determine if macro or micro models are more fitting for such an analysis. Results show that micro models might be more accurate, but macro models require less data and work better for non-traffic engineering issues and also for long term transportation planning (Huang et al., 2016).

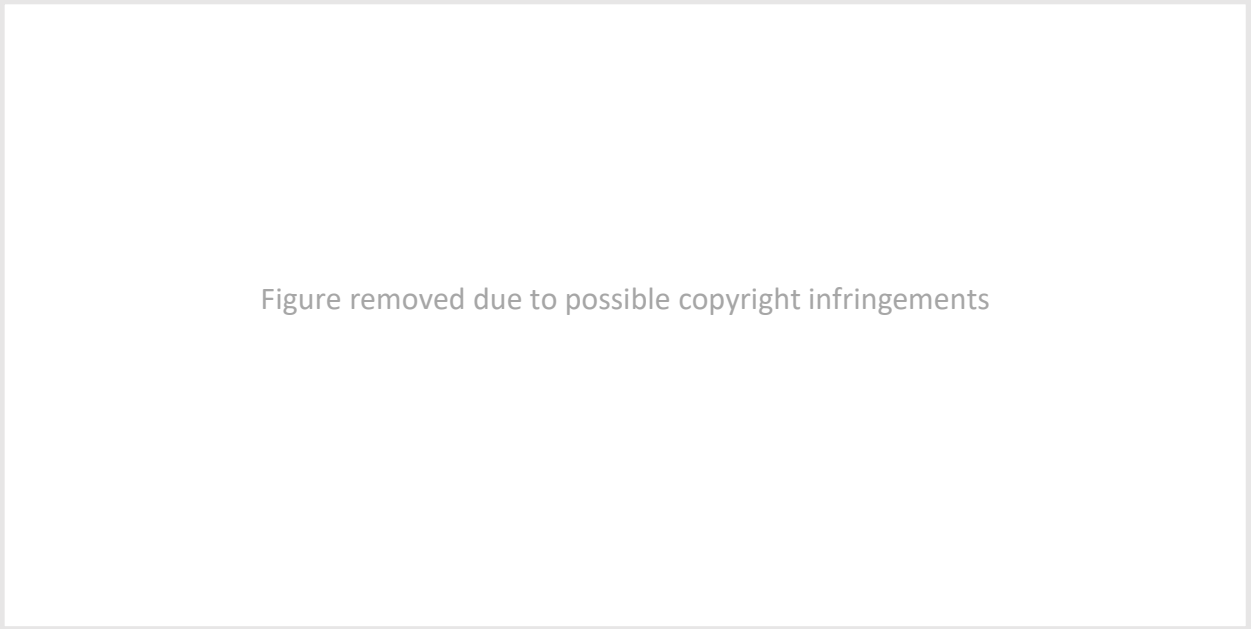
In conclusion, when choosing a model for the estimation, the two main possibilities are GLM or Bayesian models. Even though the latest is widely considered more advanced and better fitting for crash analysis than the GLM models, researchers have found that in many cases the Bayesian model does not provide a considerably better estimate than the traditional models. Also, in the case of GLM, various distribution types can be used: the most widely used ones are the negative binomial and Poisson distribution, and the zero-inflated versions of these which are developed in order to overcome the problem of overdispersion. All these models can be good choices for crash estimation for one reason or another. In this thesis however, the easiest methods (negative binomial and Poisson) will be used, in order to see how do they compare to other models using more complicated methods.

## 2.2. Variables and parameters

Apart from what kind of models to use, there is another issue which needs to be discussed: what kind of variables are feasible to include in the model, which ones are most likely to prove to be significant and therefore need to be concentrated on?

The causes of crashes are generally categorized into three classes: road environment, vehicle attributes and human behavior (Thomas et al., 2013). The following diagram shows the ratio between these factors:

**Figure 2: Contributing Crash Factors (US Department of Transportation, 2011)**




As seen on Figure 2, the factor that has the most influence on crashes is by far the group of human factors. Road environment factors are a much smaller group, and vehicle factors are almost negligible. For this reason, vehicle factors will not be further considered during this analysis. Even in the case of road environment factors, there is also a human influence in the majority of the cases. This makes sense, since it is very rare that the infrastructure is designed so poorly that it would cause a crash on its own. Instead, almost always a human error is also needed in order to create a crash. However, as seen on the diagram, road environment factors together with a human error can cause crashes in quite a substantial ratio (28%). Since research mostly concentrates on the human factors (like socio-demographical variables) and from the road environment factors they usually only consider traffic volumes, in this thesis we will try to focus on other, less researched factors like land use or road network properties.

According to K. Kim and E. Yamashita, different land uses generate and attract different types of trips, which also effects the volume of traffic. Therefore, it would be justifiable to assume that these factors would be the most relevant variables when estimating crash numbers. However, in reality, crashes are more of a factor drivers and travelers rather than land use properties. Despite this the analysis has been done to determine the relevance of land uses to crashes, and it has been found that there is indeed relevance between the two. In general, traffic volumes and other factors might be more relevant than land use factors, but in some cases land uses with less traffic have proven to be more prone to having crashes than other land uses with higher traffic. This

was mostly due to the timely distribution of crashes, but it shows that land uses are indeed relevant factors in crash estimation (Kim & Yamashita, 2002).

It seems that there are more types of crashes, and they differ in their properties and how they should be handled. Some crashes are more likely to cluster geographically than others. The reason for this is that some types of crashes are more related to roadway factors and these cluster more easily; other crashes are more related to human or other factors, and these do not have a spatial correlation (Strauss & Lentz, 2009). If this is true, that means that in some cases exploring the correlation of spatial factors could prove more useful, and in some others, temporal, socio-demographical and other factors may be more relevant for the research. In the followings, some examples are listed of earlier studies using different variables. The findings are also mentioned.

Boulieri et al. (2017) investigates the spatial and temporal correlations in the level of severity of road crashes. The results show important associations in spatial variables, and a downward temporal trend. According to the findings, in cities there is a higher risk of light accidents, while in suburban areas there is a bigger chance of severe crashes.

Dissanayake et al. (2009) models the effect of land use and temporal factors on child pedestrian casualties. The results show that secondary retail and high density residential land use types and some others are associated with child casualties, however, for some of these (eg. for educational sites) this is only true for different time periods.

According to S. B. Kusselson (2013), commercial and industrial land use, higher mean household income and a lower percentage of undevelopable land use have a significant influence in increasing crash risk. However, this research has been done only examining frontage roads sections near Houston, Texas.

In a study using negative binomial models, zonal VKT, major and minor road kilometers, total working and household population and intersection density were found to be correlated positively with the amount of crashes, while higher posted speed and higher congestion in the zone had a negative correlation. (Hadayeghi et al., 2003).

Huang et al. (2010) uses a Bayesian model to determine crash rates by using the variables of VMT and population. The results are essentially the same when looking at only serious crashes or all crashes. Higher traffic and population will result in higher crash risk.

Using Poisson regression models, Ivan et al. (2000) estimates crash rates as a function of traffic density, land use, light conditions and time of day. They have found that different variables prove to be significant for single-vehicle and multi-vehicle crashes. However, their analysis took place on rural highway sections, therefore the results of this study might not be significant for urban crash analyses.

W. Oris (2011) examined the spatial correlation of fatal car crashes in Kentucky in his Master-thesis. He used temporal and demographical variables to identify spatial patterns. Through rate calculation analysis of crash locations and daily traffic it was determined that roads with high speed limits and winding topography led to the highest number of crashes and highest rate of fatal crashes per 1,000 daily vehicles. Planar kernel density estimation showed temporal and socio-demographical patterns (eg. hot spots involving alcohol occurred in close proximity to bars or restaurants), while the results of network kernel density estimation showed that most hot spots were in high traffic areas or where major roads converged with secondary roads.

The correlation between crashes and weather conditions has also been investigated in a study. However, this analysis was done on county-level, and it has found that some areas are more prone to the clustering of weather-related crashes than others. The areas where more correlation was found were the ones experiencing a higher amount of precipitation. In these areas, weather has been found to be a contributor in a high number crashes (Khan et al., 2008).

The role of street network properties has also been specifically explored. Marshall and Garrick have performed an analysis on 24 cities in California, trying to determine how does the street network affect safety. According to their findings, street network characteristics do play a role in road safety outcomes; more specifically low street network density comes with a high risk of severe crashes, and high density comes with a lower risk. The worst results are to be found in street networks with high density and low connectivity, or the other way around. (Marshall & Garrick, 2010)

Pulugurtha et. al (2012) used only land use data for the crash estimation. It was found that many variables such as population, number of household units, employment, traffic production and attraction and some others were closely correlated to the land use characteristics, therefore they did not need to be taken into account separately. Even some land use categories like urban residential commercial, rural district and mixed use district were correlated to other land use categories, therefore these were also not considered. Finally, it was found that land use has a strong correlation to crash rates. Interestingly but understandably, the correlation in the case of single family residential areas was negative. Negative binomial models were used for the estimation.

It seems that almost every study has taken a different approach in terms of which variables to concentrate on or even which statistical method to use. Still, they all ended up with similar results, which suggests that even though the accuracy of models may vary based on the used variables and statistical methods, but in the end this might not affect the results that much as one would think.

More importantly, the research of Kim & Yamashita (discussed in the beginning of this section) has proven that land use factors are indeed relevant when it comes to crash estimation. Furthermore, the last two examples (Marshall & Garrick and Pulgurtha et. al) have demonstrated

that performing the estimation only using land use and road network properties as variables is a viable option. Therefore, in this analysis these two variables will be used to build a model, in order to see how competent this model is in comparison to more complicated models.

## 2.3. Spatial units

After determining the modelling technique to use and the variables to concentrate on, there is still a third, seemingly less complicated question that needs to be decided: what kind of spatial unit should be used during the analysis?

C. Siddiqui (2012) has an answer for this question too: in the doctoral thesis that was already mentioned in the beginning of the literature review, he also investigated the goodness-of-fit using different spatial units, like TAZs, block groups or census tracts. It was found that severe and pedestrian crash models had similar fits for TAZs and BGs, but better than for CTs. This indicates that these models are affected more by the size of the spatial unit, rather than the zoning configurations. Because of the wide usage of TAZs within different kind of analyses and even long range transportation plans (LRTP), finally TAZs were selected as the basis of the current analysis. However, it was acknowledged that because TAZs are in many cases bordered by major roadways, some crashes might occur near or on the boundaries of the zones which can cause inaccuracies in the model. For this reason, pedestrian crashes were modelled using a hierarchical Bayesian framework separately for boundary and interior crashes. It was found that the goodness-of-fit was better in these models than the ones which do not consider location within the TAZ.

The question of which geographical unit to choose is more elaborately investigated in another paper written by M. Abdel-Aty, C. Siddiqui and others (2013). Here they compare TAZs, BGs and CTs for total crashes, severe crashes and pedestrian crashes. The objective of the study was to investigate the effect of zonal variation on the previously mentioned models. The results show that the significance of explanatory variables is not consistent among these models. TAZs might be the most commonly used geographical units by transportation planners, but they are more suitable for LRTPs than for crash analysis. Therefore, at the end of the study the exploration of other zone systems is recommended (Abdel-Aty et al., 2013).

Following up with this topic, a new zonal system for traffic safety analysis has been developed. The new system is called Traffic Safety Analysis Zone (TSAZ), and solves several issues experienced when using the TAZ system. Models using TSAZs have a better fit than the ones using TAZs. However, even though in TSAZs the amount of boundary crashes is lower than with TAZs, it is still too high not to consider it during the analysis. Therefore, it is necessary to investigate

this topic further in order to come up with new zonal systems which fit even better for crash analysis. (Lee et al.,2014)

Because just as in the case of modelling techniques and variables, also in the case of spatial units it is not clearly decidable which approach is the best, we will use the simplest technique once again. In this case, that is the 1x1 km raster grid that was used by P. Songpatansilp et al. (2015) in a study that will be mentioned in detail in the next chapter. This method has several advantages in comparison to other, more complicated methods, and since current analysis is trying to use the simplest techniques, the raster grid proved to be the best solution.

## 2.4. Thesis implications

It seems that with modeling techniques, variables and also spatial units, the more complicated method we use, the better result we get, at least in most cases. But how much better are those results in comparison to when using simple methods? Is it really worth to put all the extra energy and time into the research, or maybe the results are good enough using simple methods too? This thesis is not trying to be one of the most complicated researches ever done in this field, because of the limitations of time and resources. Instead, we will try to determine if it is possible to come up with a good forecasting model, with only using the basic methods in every area of the research. This can prove very useful when a model needs to be developed for example in a developing country, where resources and especially traffic, socio-demographic and other kind of data are scarce.

Before proceeding to the research, the three most relevant papers to this study will be listed here. These three papers are very similar to present thesis in terms of their research topic. After describing these studies, the differences of this thesis in comparison to the papers will be highlighted, justifying why this topic has to be explored.

P. Songpatanasilp et al. (2015) have performed a traffic accident risk analysis based on road and land use factors, using the example of Tokyo. They used a 1 x 1 km grid in order to analyze how the land use and road related factors influence road safety. GLM models and their zero-inflated versions were used for the estimation, and it was found that the negative binomial model had a better fit than the Poisson-regression model, and also the zero-inflated versions' goodness-of-fit was superior in comparison to the basic GLM models. Regarding the effect of land use factors on safety, they found that crashes occur more frequently in commercial areas and less frequently in residential areas. However, they note that these results are limited to Tokyo, and performing similar analysis on other areas would be necessary.

The Belgian Science Policy Office has estimated the traffic impact of land use in relation to road infrastructure, using three case study areas in Belgium. They used multi-modal transport models for the analysis, but the output appeared to be unfit for this purpose. Therefore, they decided to put more focus on the evolution of land use and traffic measures and their impact on road safety. Two of the three case studies have proved that there is a significant relation between road accidents and the balance between land use and road / traffic characteristics. (Belgian Science Policy Office, n.d.)

Finally, Q. Guo et al. (2015) have explored the role of street patterns in zone-based traffic safety analysis. They used a zone-based Hong Kong database, and the topological characteristics of street patterns were estimated with Space Syntax. Then a joint probability model was adopted to analyze crash frequency and severity. In addition to the characteristics of street patterns, speed, road geometry, land-use patterns, and temporal factors were also considered.

They classified the study zones into three categories according to geographical layout: grid, deformed grid and irregular. Connectivity, depth and integration was determined for the elements of the road network. Speed data also was acquired from 480 GPS-equipped taxis traveling on the road network of Hong Kong. They found that street pattern characteristics play an important role in zone-based traffic safety analysis. Crash severity is significantly related to integration, road density, junction density, average speed, land-use patterns, and temporal factors. Furthermore, commercial areas are associated with lower crash severity and increased speed is significantly associated with more severe crashes.

These three papers and especially the last one have very much in common with present thesis' topic. However, our goal is to perform an analysis as simple as possible, and compare it to other researches to see if the analysis is still viable. Q. Guo et al. also use Hong Kong as an example, but they explore the relevance of many different variables including speed and temporal factors. Furthermore, they consider connectivity, depth and integration as parameters of the road network, and they work with geographical layouts like grid, deformed grid or irregular. To simplify things, in our research we will only consider two factors: land use and road network properties. However, when examining the road network, we will not concentrate on the geographical layout; instead, the percentage of different road categories within a zone will be used, which has never been done before.

And when talking about spatial units, it seems that the most widespread approach is to use TAZs. However, TAZs might not be available in some less developed regions of the world. Furthermore, as some previous studies mention, TAZs has the disadvantage of the zones being bordered by major thoroughfares, which of course means that a high percentage of crashes will be located on the border of the zones. To avoid this, we will use the same method as P. Songpatanasilp et al. used in their study; namely a 1x1 km raster grid. This not only avoids the biggest disadvantage of using a TAZ-system, but it is also a very simple method which can be used in any location regardless of the data available.

Regarding the modeling techniques the situation is similar as with the variables and spatial units: the simplest methods will be used in order to see how good of a model can be built only using the basic methods. Because of the low diversity of used variables (only two variables, land use and road network properties are used), it would also not be justifiable to use overly complicated models. Furthermore, the restrictions in time and resources make it logical to use one of the simplest models. And lastly, as previously mentioned, by using simple GLM models for the analysis, we will be able to find out how good are the results these basic models provide in comparison to their more complicated versions.

# 3. Data preparation

The following chapter contains the process of how the data was prepared for the analysis. First the crash data is introduced in detail, then preliminary analysis is performed on the data only based on the basic information already at hand, and finally the mapping process is presented, which is the last step before the development of the models can begin.

## 3.1. Introduction of dataset

The basis of this analysis is the crash data acquired from the Transport Information System of the Hong Kong Transport Department. This database contains information about 16170 crashes: these are all the injury crashes that happened throughout the year of 2015 within the limits of HKSAR (Hong Kong Special Administrative Region). The crashes are ranked by the seriousness of injuries (light, serious, fatal), and are assigned to geographic locations, which allows them to be plotted on a map.

The dataset is split into two different databases: the first one is concentrating on the casualties, and the second one on the participating vehicles. If there were more casualties (meaning injuries in this case) or more vehicles, the dataset will have more rows for the same crash. The basic attributes of the crashes are the same in both databases, these are the attributes that are related to the crash environment. There are 39 of these; all of them are listed in Table 1 together with the possible values for each field:

**Table 1: Common fields of crash data**

| | | | |
|---|---|---|---|
| 1 | REF | Reference No. | |
| 2 | SEVERITY | Severity | 1=Fatal, 2= Serious, 3= Slight |
| 3 | POLDIV | Police Division | |
| 4 | DBOARD | District Board Area | |
| 5 | HIT | Hit and run | 1= Yes, 2= No |
| 6 | ACC_DATE | Date of accident | DD/MM/YY |
| 7 | ACC_TIME | Time | HH/MM |
| 8 | WEEK_DAY | Day of week | 1= Mon, 2= Tue, 3= Wed, 4= Thu, 5= Fri, 6= Sat, 7= Sun |
| 9 | ST_NM | Street Name | |
| 10 | IN_70M_JCN | Within 70m of junction | 1= Yes, 2= No |
| 11 | SECND_ST | Second Street name | |
| 12 | IN_20M_JCN | Within 20m of junction | 1= Yes, 2= No |
| 13 | IDEN_FTR | Identifying feature | |
| 14 | GRID_E | Easting Grid | |

| | | | |
|---|---|---|---|
| 15 | GRID_N | Northing Grid | |
| 16 | PREC_LOCTN | Precise location | |
| 17 | HAPPEN | How accident happened | |
| 18 | NO_VEH | Number of vehicles | |
| 19 | NO_CSU | Number of causalties | |
| 20 | WEATHER | Weather | 1= Clear, 2= Dull, 3= Fog/mist, 4= Strong Wind, 9= Not known |
| 21 | RAIN | Rain | 1= Not raining, 2= Light rain, 3= Heavy rain, 9= Not known |
| 22 | NAT_LGT | Natural Light | 1= Daylight, 2= Dawn/ Dusk, 3= Dark, 9= Not known |
| 23 | ST_LGT | Street Lighting | 1= Good, 2= Poor, 3= Obscured, 4= Not lit, 5= None, 6= Daylight, 9= Not known |
| 24 | SPEED_LMT | Speed Limit | |
| 25 | TRAFF_AID | Condition of Traffic Aids | 1= Poor markings, 2= Other poor aids, 3= No signficant deficiencies, 9= Not known |
| 26 | TRAFF_CONG | Traffic Congestion | 1= Severe, 2= Moderate, 3= None, 9= Not known |
| 27 | RD_SURFACE | Road Surface | 1= Wet, 2= Dry, 9= Not known |
| 28 | AT_NEAR | At or Near | A= Roadworks (Govt), B= Roadworks (Utilities), C= Construction materials, D= Landslip/ fallen tree, E= Flooding, F=Timber walkway, G= Others, H= None, Z= Not known |
| 29 | XING_LMT | On a crossing controlled by | 1= Zebra, 2= Traffic signal, 3= Police, 4= Crossing patrol, 5= Cautionary Crossing, 8= None |
| 30 | XING_15M | Within 15m of crossing controlled by | 1= Zebra, 2= Traffic signal, 3= Police, 4= Crossing patrol, 5= Cautionary Crossing, 6= Footbridge/ subway, 8= None |
| 31 | JCN_CTRL | Junction control | 1= No, 2= Stop, 3= Give way, 4= Traffic signal, 5= Police, 6= Not junction |
| 32 | JCN_TYPE | Junction type | 1= Roundabout, 2= T-junction, 3= Staggered, 4= Y-junction, 5= Slip road, 6= Cross-roads, 7= Multiple, 8= Private access, 9= Other, 10= Not within 20M |
| 33 | RD_TYPE | Road type | 1= One way, 2= Two way, 3= Dual Carriageway, 4= More than 2 carriageway |
| 34 | CW_WIDTH | Carriage Width | |
| 35 | NO_LANE | Number of Lanes | 1= One lane, 2= Two lanes, 3= More than two lanes |
| 36 | RD_CLASS | Road Classification | 1= Primary Distributor, 2= Private Road, 3= Other |
| 37 | VEH_MOVE | Vehicle Movements | 1= One moving veh, 2= 2 in same direction, 3= 2 from opposite direction, 4= 2 from different roads, 5= >2 from same direction, 6= >2 from opposite direction, 7= >2 from different roads |
| 38 | OVERTAKE | Overtaking | 1= One vehicle overtaking, 2= 2+ vehicle overtaking, 3= No overtaking |
| 39 | ENV_CONTRI | Contributory Factor | |

As seen in Table 1, every crash has a reference number as the first attribute. Some crashes have multiple entries in the database, because in case of several injuries or vehicles there will be multiple rows with the same basic attributes but with the specific data related to the given person

or vehicle. In these cases it is important to be able to identify that multiple rows belong to the same crash, and this can be done with the help of the reference number. The second attribute is the severity, the values can be fatal, severe or slight. In case there were multiple injuries in a crash, the most serious injury will be considered here. The exact time of the crash is also part of this database, divided into date, time, and weekday to make investigations regarding the day of the week easier. The number of the roadway (in the official Hong Kong system) where the crash happened can also be found in the database, together with the number of the second roadway in case the crash occurred in an intersection. There are also two separate fields showing if the crash happened either within 70, or within 20 meters of an intersection. This information can come useful, but the vital part of our analysis are the next two fields: the coordinates of the crashes according to the Easting and Northing Grid. These two numbers make it possible to have the crashes plotted on a map, which is a basic need for this research. The numbers show the location of the crashes in a Hong Kong based grid system, with meter values showing the distance from an origin located in the sea near Hong Kong. After these two fields another two comes with very useful information: the description of the precise location of the crash, and also a description explaining how the crash exactly happened. Because these are free text fields, there is no way to process all of them in our research, but in individual cases they can come very handy when a specific crash's location or circumstances need to be checked.

The next two attributes are the number of casualties and vehicles, which also shows us how many rows will the crash have in the casualty and vehicle databases. After this comes the weather and lighting: according to the database, the weather can be clear, dull, foggy/misty or with strong wind; there is a separate field to show if there was any rain at the time of the crash and if there was, how heavy; and the third and fourth fields show the natural light and street lighting. Next the attributes related to the road environment follow: speed limit, condition of traffic aids, traffic congestion and road surface condition (dry or wet) are listed. Then the properties of the junction are listed, like junction control, junction type, and if there is a pedestrian crossing nearby. Finally, after listing the roadway properties (road type, number of lanes, road classification) the vehicle movements are mentioned, including if there was overtaking involved. Additionally, at the end also any contributing factors are mentioned. These could be several things from slippery road through potholes to obscured vision because of different reasons.

These are all the attributes related to the crash environment. The values of these fields for the same crash will always be the same, even if there are several rows because of multiple casualties or involved vehicles.

After this part, the data specific to injured persons or vehicles follows. We will start the introduction with the data related to vehicles.

**Table 2: Vehicle database**

| | | | |
|---|---|---|---|
| 1 | REF | Reference No. | |
| 2 | VEH_NO | Vehicle Number | |
| 3 | DRIVER_AGE | Age of driver | |
| 4 | DRIVER_SEX | Sex of driver | 1= M, 2= F, 9= Not known |
| 5 | LICEN_TYPE | Driving License Type | 1= Full, 2= Not Valid, 3= None, 4= Temporary, 5= International, 6= Other, 7= Learner, 8= N/A, 9= Not known |
| 6 | YR_MFT | Year of Manufacture | |
| 7 | VEH_AGE | Vehicle age | |
| 8 | VALID_LIC | Valid License | 1= Yes, 2= No, 8= N/A, 9= Not known |
| 9 | VALID_INS | Valid Insurance | 1= Yes, 2= No, 8= N/A, 9= Not known |
| 10 | VEH_CLASS | Vehicle Class | |
| 11 | VEH_OWNER | Vehicle Owner | 1= Private, 2= Military, 3= Police, 4= Fire service, 5= Ambulance, 6= Other Govt Dept, 7= Other, 9= Not known |
| 12 | MANOVRE | Main Vehicle Manouvre | 1= Straight ahead, 2= Changing lanes, 3= Overtaking o/s, 4= Overtaking n/s, 5= Crossing traffice, 6= Turning right, 7= Turning left, 8= U-turn, 9= Slowing/ stopping, 10= Stopped in traf, 11= Starting in traffic, 12= Leaving n/s park, 13= Leaving o/s park, 14= Parked, 15= Reversing, 16= Driverless moving, 17= Ran off road, 18 Other, 99= Not known |
| 13 | COLLIDE | Vehilce Collision With | 1= Vehicle, 2= Pedestrian, 3= Animal, 4= Object on c'way, 5= Traffic sign post, 6= Lamp/ teleph post, 7= Road sign, 8= Tree, 9= Wall/ bridge prpt, 10= Utility co equip, 11= Bollard, 12= Fire hydrant, 13= Pedestrn barrier, 14= Crash barrier, 15= Road works, 16= Hoarding/ walkway, 17= Hawker stall, 18= Other, 19= None, 99= Not known |
| 14 | VEH_LGT | Vehicle light | 1= None, 2= Parking lights, 3= Headlights dipped, 4= Headlight main beam, 8= N/A, 9= Not known |
| 15 | PT_IMPACT | First powint of impact | 1= Front, 2= Back, 3= Offside, 4= Nearside, 5= No impact, 9= Not known |
| 16 | DIR_FROM | Direction from | |
| 17 | DIR_TO | Driection to | |
| 18 | DAMAGE_POS | Part of vehicle damaged | |
| 19 | DAMAGE_SEV | Damage Severity | 1= No, 2= Slight, 3= Severe, 9= Not known |
| 20 | FTYRE_TYPE | Front Tyres Type | 1= Radial, 2= Cross-ply, 3= Both, 8= N/A, 9= Not known |
| 21 | FTYRE_COND | Front Tyres Condition | 1= Legal, 2= Illegal, 8= N/A, 9= Not known |
| 22 | RTYRE_TYPE | Rear Tyres Type | 1= Radial, 2= Cross-ply, 3= Both, 8= N/A, 9= Not known |
| 23 | RTYRE_COND | Rear Tyres Condtion | 1= Legal, 2= Illegal, 8= N/A, 9= Not known |
| 24 | GV_LOAD | Goods vehilce loading | 1= None, 2= Secure, 3= Insecure |
| 25 | OVERLOAD | Vehicle overloaded | 1= Yes, 2= No |

| | | | |
|---|---|---|---|
| 26 | MVE_REQ | MVE report requested | 1= Yes, 2= No |
| 27 | FIRE | Caught fire | 1= Yes, 2= No, 9= Not known |
| 28 | SKID_OVRTN | Skidding/ overturn | 1= Skidding, 2= Skidding and overturn, 3= Jacknife, 4= Jacknife and overturn, 5= Overturn, 6= None, 9= Not known |
| 29 | DEF_ALLEGE | Vehicle defects alleged | 1= Yes, 2= No |
| 30 | VEH_CONTRI | Vehilce Contributory factors | |
| 31 | DRI_CONTRI | Driver Contributory factors | |
| 32 | LICEN_CODE | Driving License Code | 1= China, 2= Other, 8= N/A, 9= Not known |
| 33 | STEERING | Steering | 1= Right, 2= Left, 8= N/A, 9= Not known |

Regarding the vehicles, the first attributes describe the driver, with mentioning his or her age, sex and the type of their driving license. Then the vehicle properties follow, like age, vehicle class, validity of vehicle license and insurance, and if the owner of the vehicle is a private person or an organization. After this the circumstances of the crash are mentioned, like main vehicle maneuver, first point of impact, direction, part of vehicle damaged and damage severity. It is also mentioned here what did the vehicle collide into, namely another vehicle, pedestrian, animal or fixed object. Then some properties are mentioned that can be an important help for deciding who was at fault. These are the type and conditions of tires, and in case of trucks if the loading was insecure or over the weight limit. There is a separate field that shows if there were any alleged vehicle defects. After listing the vehicle and driver contributory factors, the type of steering (left or right) is also mentioned, and with this, the attributes of vehicles are finished.

**Table 3: Casualty database**

| | | | |
|---|---|---|---|
| 1 | REF | Reference No. | |
| 2 | CAS_NO | Casualty Number | |
| 3 | CAS_AGE | Casualty Age | |
| 4 | CAS_SEX | Casualty Sex | 1= M, 2= F, 9= Not known |
| 5 | INJURY | Degree of injury | 1= Fatal, 2= Serious, 3= Slight |
| 6 | ROLE | Role of casualty | 1= Driver, 2= Passenger, 3= Pedestrian |
| 7 | SB_WORN | Seat belt or crash helmet worn | 1= Yes, 2= No, 9= Not known |
| 8 | LOCATN_INJ | Location of injury | A= Head, B= Upper trunk, C= Lower trunk, D= Arms, E= Legs |
| 9 | IN_VEH_NO | In vehicle number | |
| 10 | SEAT | Seat occupied | 1= Rear, 2= Front nearside, 3= Driver, 4= Standing in lower deck, 5= G/V Compartment (fixed), 6= G/V Compartment (w/o fixed), 8= Standing in upper deck, 9= Not known |

| | | | |
|---|---|---|---|
| 11 | PED_LOCATN | Pedestrian Location | 1= Footpath/ verge, 2= Refuge/ Central strip, 3= On controlled crossing, 4= Within 15M of controlled crossing, 5= Carriageway, 8= Other, 9= Not known |
| 12 | PED_ACTION | Pedestrian Action | 1= Walking (back), 2= Walking (face), 3= Standing, 4= Boarding, 5= Alighting, 6= Falling or jumping from, 7= Working at vehicle, 8= Other working, 9= Playing, 10= Crossing from nearside, 11= Crossing from offside, 99= Not known |
| 13 | PED_CIRCUM | Special Circumstances | 1= Footpath overcrowded, 2= Footpath obstructed, 3= One side no footpath, 4= Two side no footpath, 5= Ran onto road, 6= Climbed over barrier, 9= None |
| 14 | DIRECTN_FR | Direction from | |
| 15 | DIRECTN_TO | Direction to | |
| 16 | CAS_CONTRI | Contributory factor of casualty | |

As for the casualties, their age, sex, degree of injury and role (driver, passenger, pedestrian) is mentioned. From the data we can also find out if the injured person used a seat belt or wore a helmet, and the data also contains the location of the injury on their body and the seat they occupied if they were sitting in a vehicle. For pedestrians, the next two fields show their location at the time of the crash, and their action. Any special circumstances are also mentioned in the data; this could be an overcrowded or obstructed footpath or maybe that the pedestrian ran onto the road. At the end, as in the case of the vehicle attributes too, directions and contributory factors are mentioned.

## 3.2. Preliminary analysis

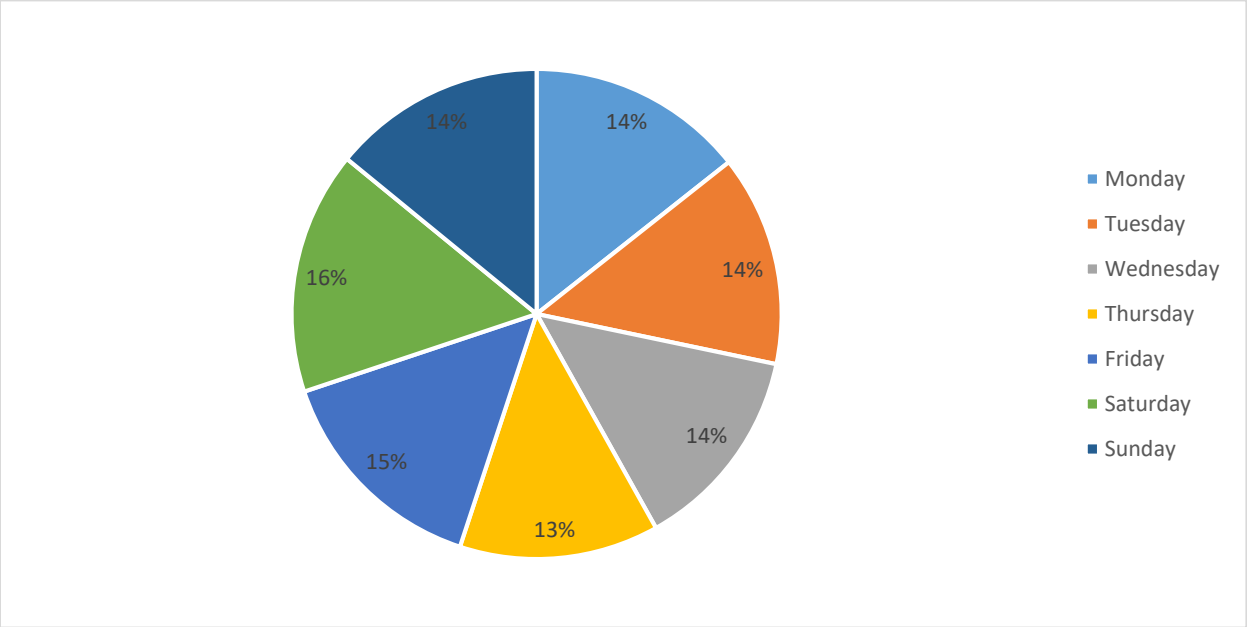
As seen from the description above, the data is very detailed and is easily usable for different kind of enquiries. Before proceeding to the actual analysis of the data some preliminary analysis can already be done at this stage. Questions that could be answered are for example the followings:

- Age group and sex most likely to be involved in a crash
- Percentage of heavy vehicles or taxis being involved in the crashes
- Types of junctions most likely to have crashes
- Ratio of pedestrian crashes in comparison to vehicle crashes
- Etc.

Some inquiries similar to these have been performed on the data. The results are presented in the followings.

## Crashes by weekdays:

**Figure 3: Number of crashes by weekdays**

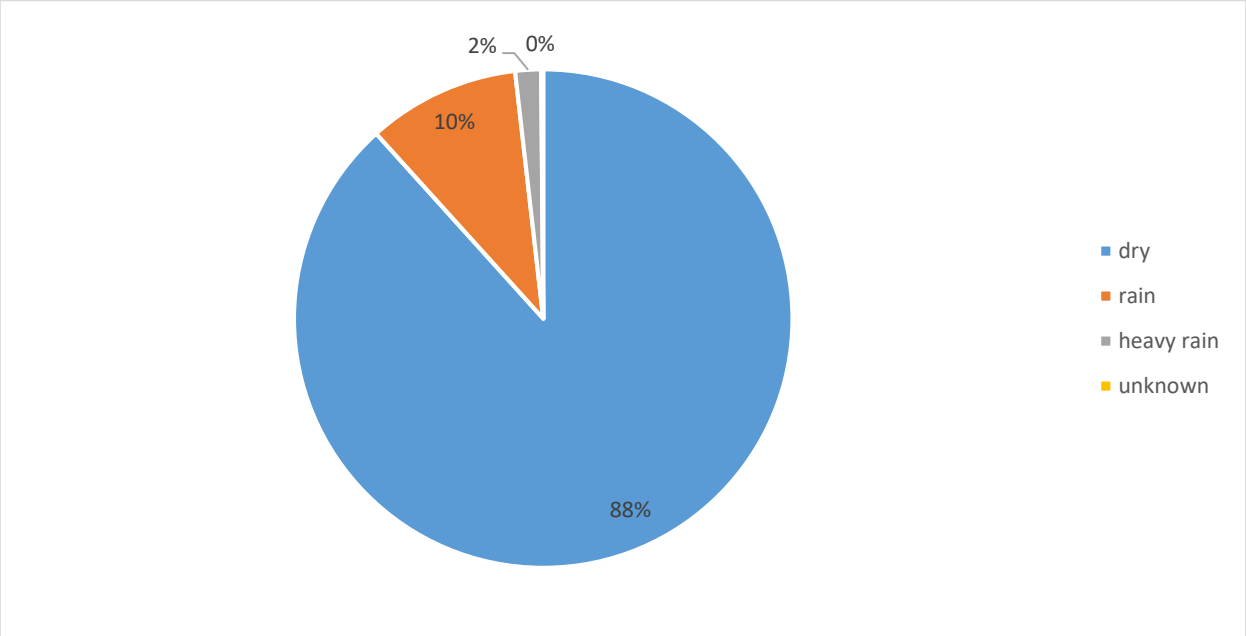



As seen above on Figure 3, the percentage of crashes happening on different weekdays is almost exactly the same for every weekday, therefore it is fair to say that the distribution of crashes throughout the week is consistent. However, there is a slight increase in the percentage on Fridays and also on Saturdays. An increased number of crashes on Fridays is nothing unique, in most big cities the Friday afternoon rush hour is the most severe from all rush hours. This is partly due to people already leaving for holidays, and partly to students for example, who only return home every weekend. Furthermore, in addition to the increased traffic, people are also tired after a long day of work and are also distracted by the idea of the coming weekend, therefore accidents happen more easily.

The reason why this scheme continues to Saturday too, is that many firms in Hong Kong obligate their workforce to work a half day (usually from 9 AM to 1 PM) on Saturday too (Boland, 2017). Therefore, the same pattern that was happening on Fridays, repeats itself on Saturdays too.

## Weather conditions:

**Figure 4: Weather conditions**



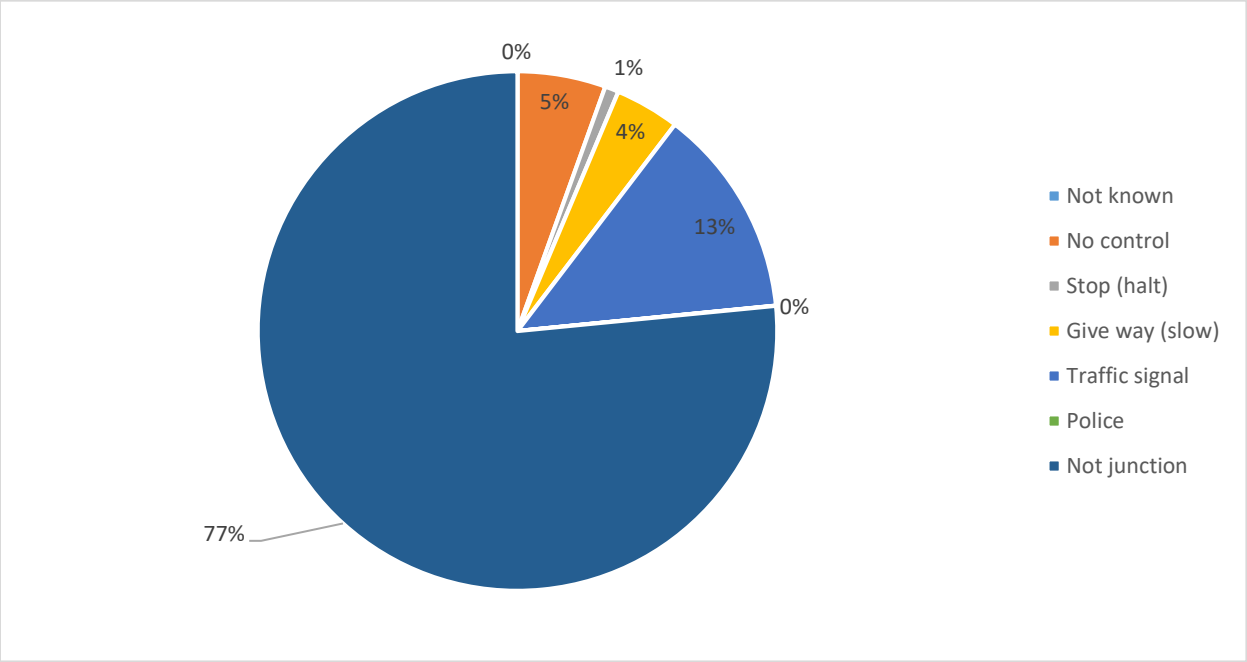

The results of this enquiry are somewhat unexpected. According to the data, 88% of all crashes happened during dry weather conditions. This would be an expected result in the case of a city with a comparatively dry climate, like cities in Southern California for example. However, Hong Kong's climate is sub-tropical with abundant rainfall, which in many cases comes in the form of a tropical cyclone or at least a thunderstorm (Hong Kong Observatory, 2015). The average annual amount of rainy days is 137 days (World Weather and Climate Information, 2016), which equals to 37.5% of a year. Therefore, it is surprising to see that in these conditions only 12% of the crashes occur in rainy weather, and a mere 2% of them in heavy rainfall. It would be justifiable to presume that a high percentage of crashes happen during rainy conditions, because of the bad driving conditions that come with a heavy storm: slippery road, reduced visibility, etc. But according to the data, this is not the case, in fact, the reality is quite the opposite. There has been no further research done on the reason of this, but possible reasons could be the followings: firstly, the people of Hong Kong are used to driving in rainy conditions, even in heavy rain. Therefore, they might become involved in less crashes in rainy weather than people living in drier climates. Secondly, from experience the people of Hong Kong know how dangerous these conditions can be, therefore they probably drive much more carefully and slowly during wet conditions. And lastly, there might not be that many people out on the streets when it is raining heavily, therefore the number of pedestrian crashes would also be reduced.

However, in dry conditions the driving culture in Hong Kong is very bold. It is not uncommon to see taxis or minibuses rushing through the city or driving on the edge of drifting out in the curves

of the roads in the mountainous areas. This fact can also account for a higher percentage of crashes within dry conditions in comparison to other regions of the world.

## Junction control:

**Figure 5: Junction control**




Another question that can be already answered based on the data is where do most of the crashes happen. Without knowing anything about the local characteristics, one could presume that most of the crashes happen at unsignalized intersections where a car hits a pedestrian or another car. This seems likely at least in the case of crashes with injuries. However, according the data, an overwhelming majority of crashes in Hong Kong actually happen between two intersections, and even the ones happening at intersections occur at traffic signals and not at "give way" signs or intersections without control. This finding shows that the main types of crashes in Hong Kong could be different from what we have presumed. The cause of these crashes could be determined by looking at the police reports of each single crash. However, based on local experience the following causes could be common in Hong Kong: firstly, in Kowloon (especially Mong Kok) and in and around Central on Hong Kong island not just the traffic, but also the amount of pedestrians is extremely high. The flow of traffic on the other hand is very slow in general. Still, if there is a gap in traffic, the average Hong Kong driver (especially taxi drivers) will speed up as much as possible to make use of this gap. Furthermore, pedestrians in Hong Kong do not have the right of way, which makes crossing the road in some cases very difficult and even dangerous (unless there are elevated pedestrian walkways present, which provide a good solution for this problem). Hong Kong drivers drive very fast even when turning,

which is again, very dangerous for pedestrians crossing the road at the corners. Therefore, many of the crashes in Hong Kong are most likely to be pedestrian crashes happening in the busy central districts. Apart from these crashes, it is also likely that drivers lose focus in a long-lasting traffic jam, so probably many of the crashes are also sideswipe or rear-end crashes occurring while being in one of those jams.

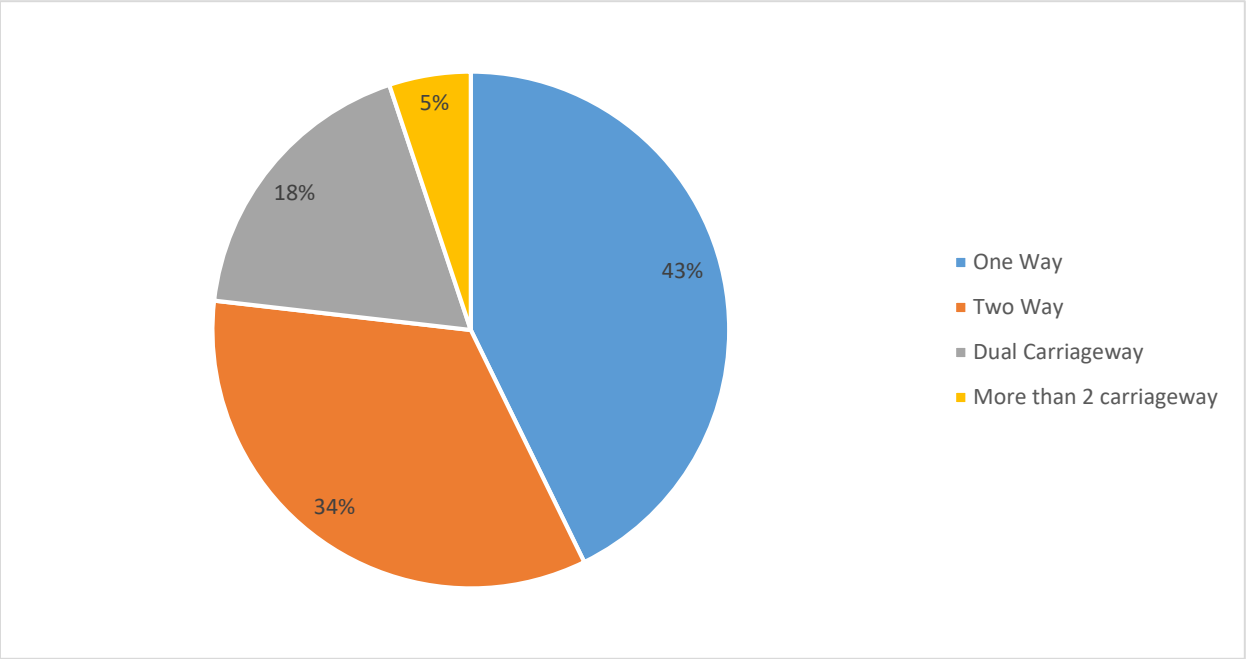
The other reason for many crashes occurring between intersections is the mountainous geography of Hong Kong. Immediately after leaving the busy central districts, drivers can find themselves on very curvy, mountainous roads. Since they are used to this environment, they will drive comparatively fast. However, these roads are not only curvy but also very narrow, which together with the high speeds is a good recipe for a crash. Crashes on these roads also count as crashes between intersection, therefore they would account for increasing the number of these kind of crashes.

After discussing junction controls, it is logical to proceed to road types being involved in crashes.

## Road types:

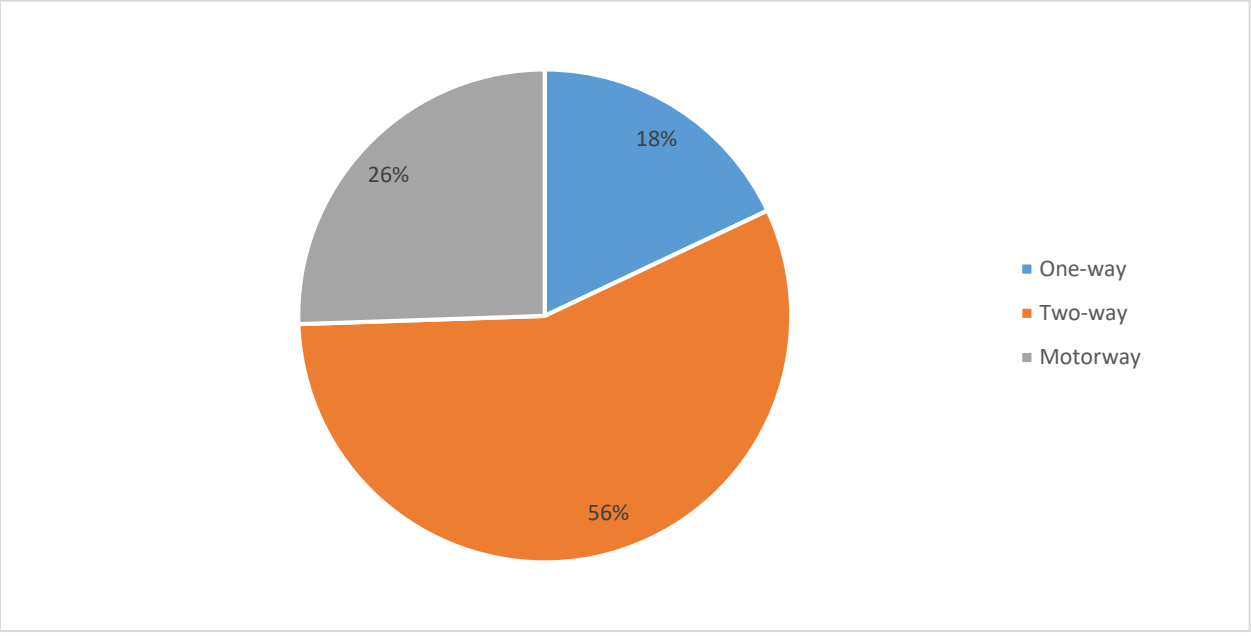
As seen from Figure 6, one-way and two-way roads are the overwhelming majority when it comes to the location of crashes. Dual carriageways seem to be relatively safe, which is not surprising because these kind of roads are considered safer everywhere in the world because of the division of different directions and the lack of at-grade junctions. “More than 2 carriageway” seems to be a category not very widely used outside Hong Kong: it represents sections of motorways where there are more than two carriageways, most probably before or after junctions. Because these sections do not represent a high percentage of the road network, it is natural that the crashes happening on these sections will represent a low percentage as well.

**Figure 6: Share of road types in crashes**




What is more interesting is the high percentage of one-way crashes, which is even higher than on normal, two-way roads. This result is somewhat unexpected, therefore it will be better to put in into context by looking at the percentages of different kinds of road in Hong Kong, irrespective of crashes.

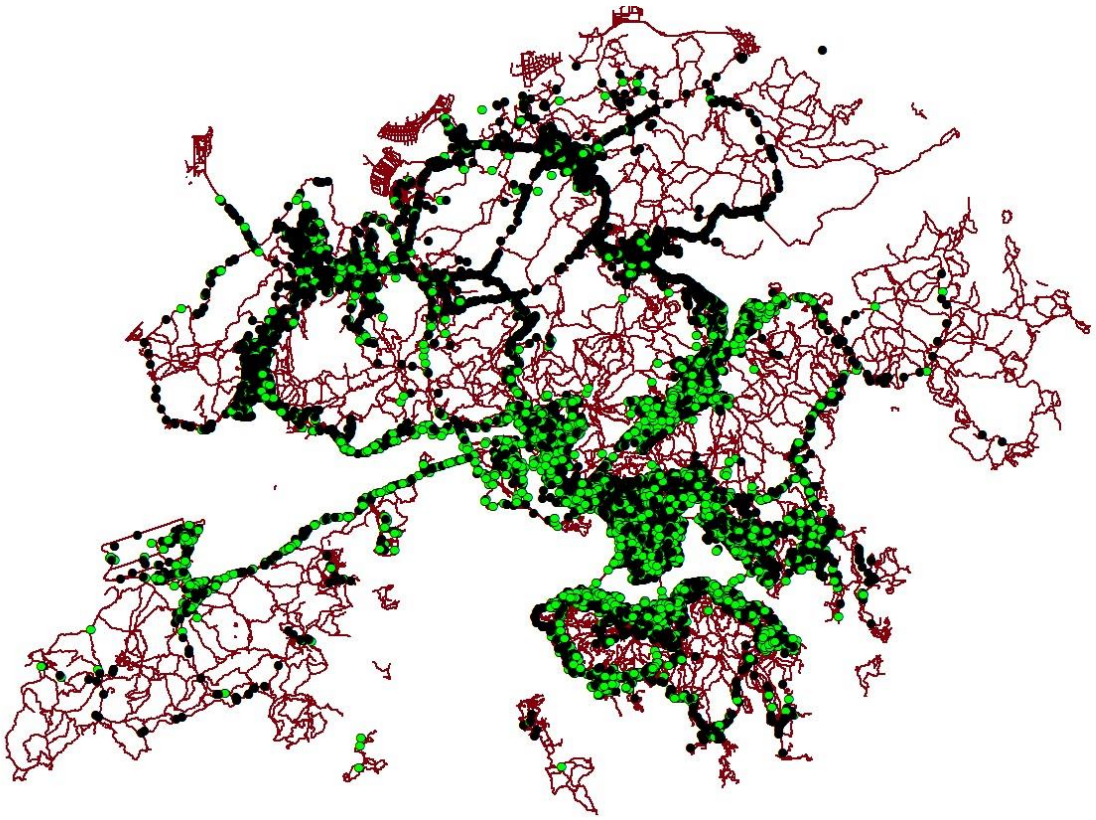
**Figure 7: Roads of Hong Kong**



According to the data acquired from the road network layer (the details about the layer will be mentioned in the following chapter) out of 10621 km of roads, 1909 km are one-way and 2712 km are motorways or trunk roads. This means that one-way roads represent 18%, and motorways 26% of the road network. However, when evaluating this data, it turned out that originally all motorway-sections and even trunk roads are considered one-way, therefore the length of all of these sections had to be subtracted from the length of all one-way roads, in order to get the length of actual one-way roads.

This raises the question though: what is the situation with the crash data? Maybe the percentage of one-way crashes is so high, because some of the crashes occurring on motorways have been categorized as one-way crashes? In order to find out the answer, a query has been performed after the crashes have been plotted on the map (the procedure of this will be presented in the next chapter as well). On the following map, one-way crashes are displayed with green color, while all other crashes have black color:

**Figure 8: One-way and two-way crashes in Hong Kong**

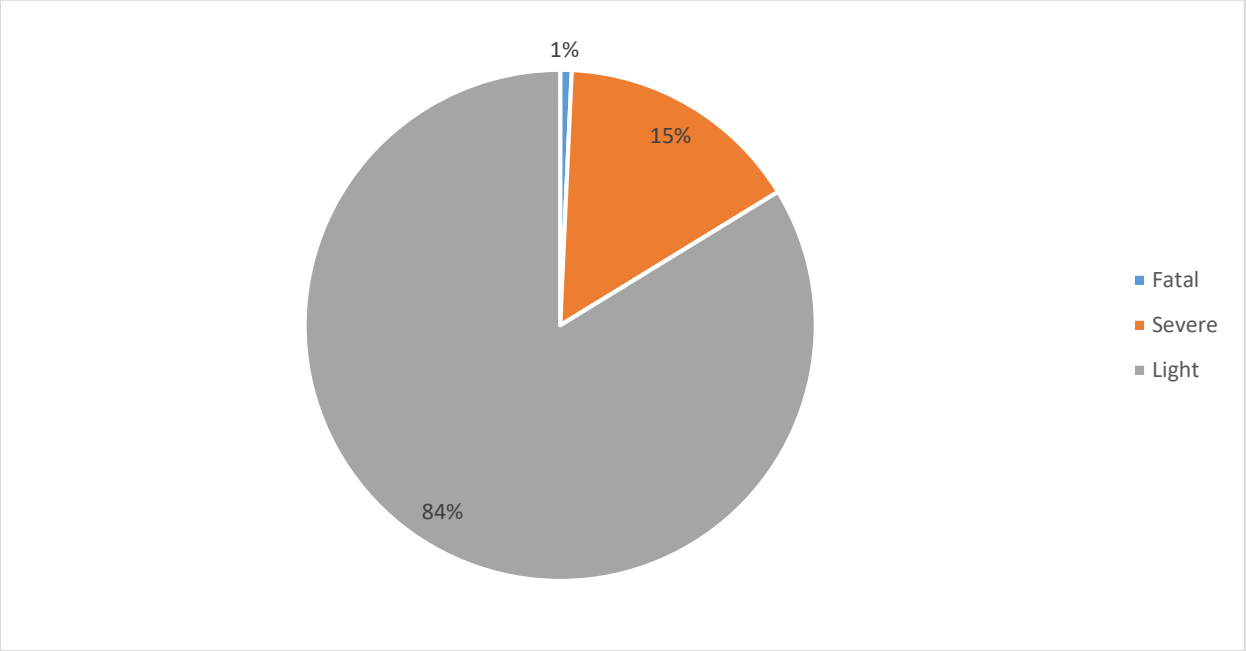


As seen on Figure 8, one-way crashes are present on most major roads, even on the causeway connecting Lantau Island (the island in the south-west) with Kowloon, which definitely considers as a motorway. Therefore, we can say that it has been proven that a high percentage of motorway crashes have incorrectly been categorized as one-way crashes. For this reason, unfortunately

there is no way to tell the real percentage of one-way or motorway crashes. The only result that comes out of this enquiry, is that the normal, two-way roads represent 56% of the road network, and they are involved in 34% of the crashes, meaning that they have proportionally less crashes than their share would justify.

## Severity:

**Figure 9: Severity**




One of the most relevant questions is the severity of crashes. As expected, over 80% of the injuries are light, 15% are severe and only 1% are fatal. These numbers are quite conventional and would probably not be very different in any other developed country (though in developing countries the percentage of fatalities could be higher).

## Correlation of speed limit and severity of crash:

***Table 4: Crashes by speed limit and severity***

| Number of crashes in 2015 | | | | |
|---|---|---|---|---|
| **Severity** | **Speed limit (km/h)** | | | |
| | **50** | **70** | **80** | **100** |
| **light** | 12006 | 762 | 321 | 140 |
| **serious** | 2227 | 122 | 62 | 33 |
| **fatal** | 105 | 5 | 3 | 1 |

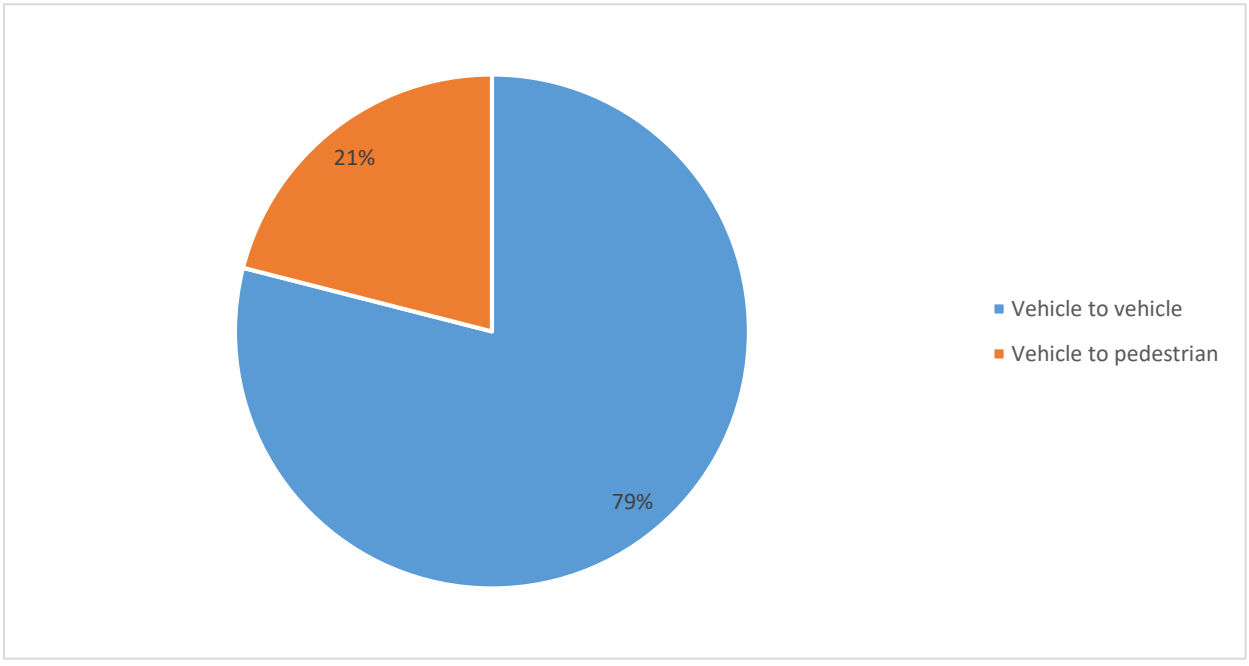
Another interesting question could be: do crashes taking place at higher speeds have more serious outcomes? At the first look it seems like the number of fatal crashes is negligible on roads with speed limits higher than 50 km/h. However, the reason for this is that the amount of all crashes taking place on these roads is also only a small fragment of all crashes. Percentage-wise the results are actually fairly stable: with any speed limits, the percentage of fatal crashes in comparison to all the crashes will always be between 0.5 and 0.7%, in the case of 50 km/h limits just as well as with higher limits.

A more useful fact that we can learn from this part of the data is rather that a vast majority of crashes happen on road with 50 km/h speed limits, which means that the significance of crashes taking place on curvy, mountainous roads might not be as high as previously assumed.

## Participants of crashes:

The question that already came up with several previous enquiries in this section, is the percentage of vehicle to vehicle and vehicle to pedestrian crashes. This is relevant when discussing severity (pedestrian crashes will probably be more severe than others) or location (most pedestrian crashes happen in the busy central areas).

**Figure 10: Participants of crashes**




As seen on Figure 10, the ratio is around 20-80 for crashes involving pedestrians or only cars. When discussing severity earlier, it was established that 1% of the crashes were fatal, and 15% of them serious. There is a good chance that a high percentage of these crashes overlap with the

21% of all crashes which involved pedestrians. In fact, there is a way to check this based on the data. The results are the following:

- From 117 fatal crashes 81 involved pedestrians (almost 70%)
- From 2510 severe crashes 784 involved pedestrians (only around 30%)

Based on these results it can be seen that the hypothesis that most fatal crashes involve pedestrians proved to be true. However, severe injuries in most cases seem to occur in case of crashes only involving cars, therefore the overlap between serious and pedestrian crashes have not proven to be true.

## Vehicle types:

If the majority of severe crashes do not involve pedestrians, the question could be, what kind of vehicle do they involve? And in general, what kind of vehicles are most likely to cause a crash or be part of one? This is the last question that needs to be answered within the preliminary examination of the data.

**Figure 11: Vehicle types participating in crashes**

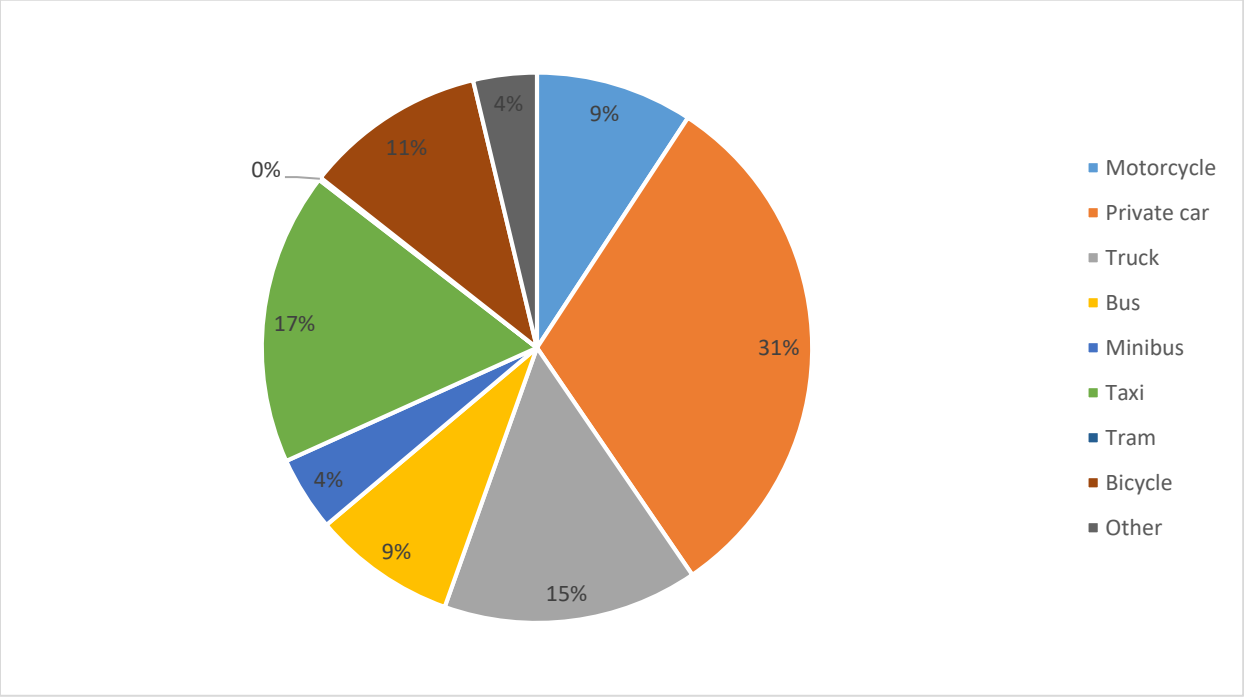
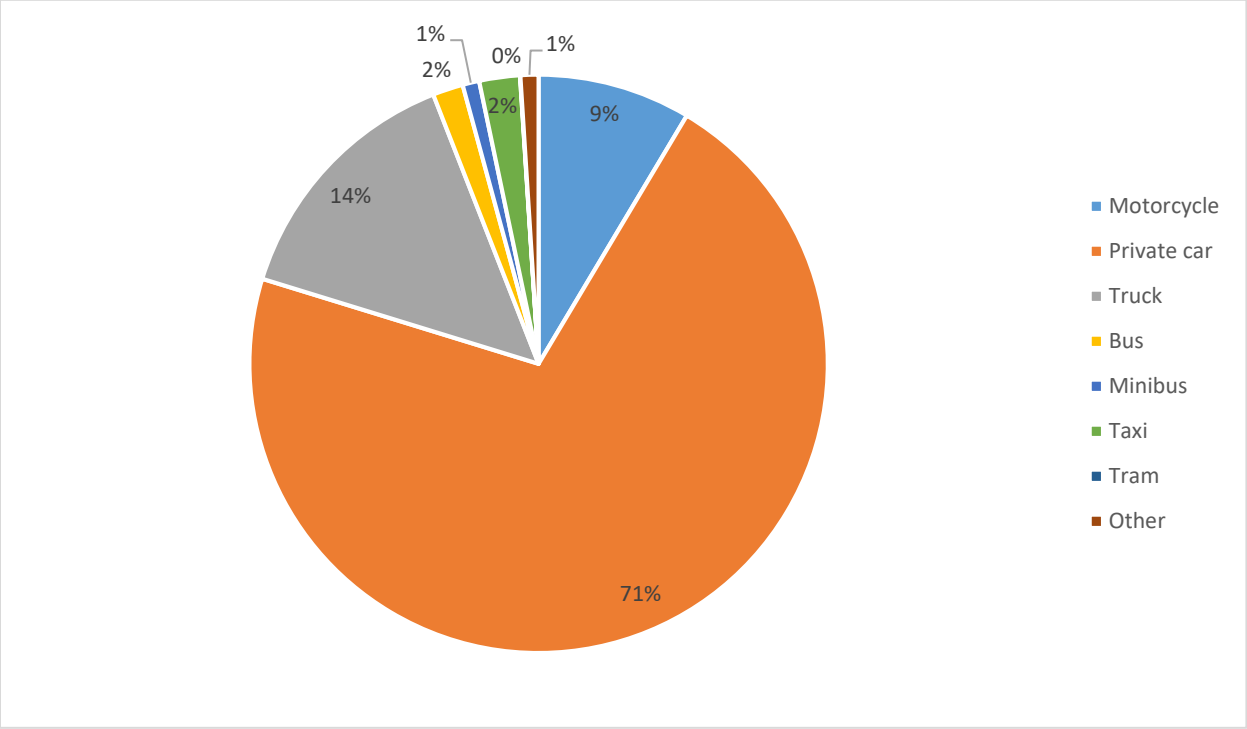



Figure 11 shows the percentage of different vehicle categories being involved in crashes. In order to interpret this data however, we would need to know how many of these different kinds of vehicles exist in Hong Kong in the first place. This data can be acquired from the homepage of the Hong Kong Transportation Department (Transport Department, 2017).

According to the data, the number of public minibuses in Hong Kong is limited at 4,350 vehicles, while the same is true for taxis, only instead of 4,350 the respective number here is 18,163. Apart from these numbers, the data contains the amount of private vehicles, buses, trucks, motorcycles and other kinds of vehicles in each year since 2012. For the purpose of the analysis the 2015 data has been used, given that the crash data is also from that year. Unfortunately, the number of bicycles in Hong Kong is not part of the data and could also not be found in other sources, therefore the following chart only includes vehicles with internal combustion or electric engines.

**Figure 12: Vehicle types in Hong Kong**




The most obvious finding is the overwhelming majority of private vehicles, which is of course no surprise. What comes as a surprise however, is the relative high percentage of trucks, motorbikes, buses and especially taxis within the vehicles involved in crashes. As seen on the charts, even though private vehicles account for around 70% of all vehicles, they are only involved in around 30% of the crashes. However, motorbikes and trucks represent the same percentage in crashes as in the general population of vehicles, which makes them relatively frequently involved in crashes in comparison to private vehicles. The biggest difference between the overall and in-crash representation can be experienced in the cases of buses, minibuses and taxis: they only account for 2, 1 and 2 percentages of the vehicle population respectively, but their involvement in crashes is 9, 4 and 17 percent in the same order. This shows that buses and minibuses are also very frequently involved in crashes, but taxis are by far the most likely to participate in a crash.

This is of course partly due to the fact that private vehicles are only being used a few hours maximum during the day, while taxis, buses and trucks are on their way the whole day. In case of trucks, this and the size of the vehicle (difficult to navigate on the busy and/or narrow roads) might be the only reasons. However, in the case of taxis and buses there might be other reasons behind the results. Taxi drivers in general drive very aggressively in Hong Kong, and it looks like this behavior shows up in the crash statistics. Many of the incidents caused by taxis are probably pedestrian-related, given that taxis in general do not give any respect to pedestrians. While by law they do not have to, given that pedestrians in most cases do not have the right of way in Hong Kong, many foreigners are not aware of that. Private car drivers might approach pedestrians with a bit more caution, even if they are the ones having the right of way. Therefore, this violent behavior of taxi drivers against pedestrians and the unawareness of some pedestrians can cause some very serious crashes.

The reason for the high involvement of buses and minibuses is also similar: in the case of the first one already the sheer size of buses (most of them are double-decker) is enough to explain many minor crashes, like side-swiping a car or hitting a pedestrian with the side-mirror. Minibuses however behave on the roads more like taxis: because of their size they are very agile, and their drivers drive in a similar manner to taxi drivers. In 2012 the government ordered all minibuses to be equipped with speedometers in the passenger area in order for the passengers to see if the minibus driver drives too fast. (Yau, 2016) These speedometers start beeping once the speed hits 80 km/h, but even this does not always stop minibus drivers to drive with excessive speeds for an extended amount of time.

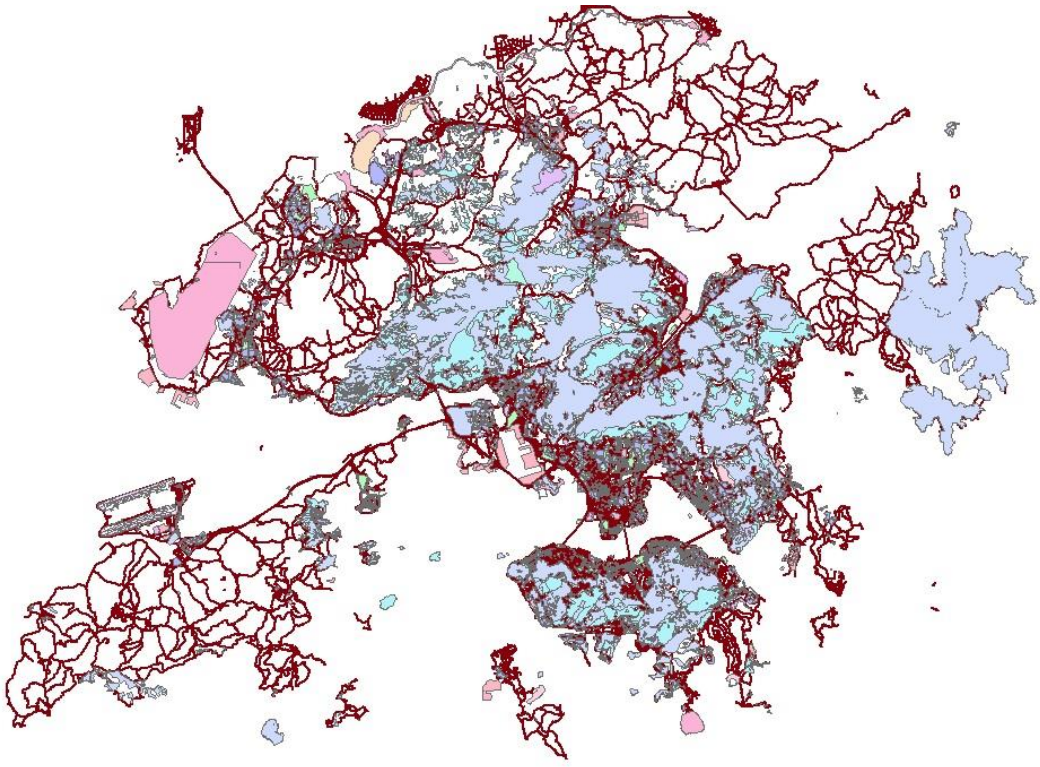
Lastly, the percentage of bicycles and motorcycles being involved in crashes is also relatively high, even though in the case of bicycles there is not data available showing their share of vehicle population. Still, motorbikes stand on 9 and bicycles on 11%. This is definitely a high percentage, given that none of these two transportation methods are very common in Hong Kong (based on own experience). Bicycles and motorbikes, especially scooters are a very common form of transportation in China, Vietnam and other underdeveloped countries, in these places they might even make up a bigger share of the vehicle population than cars. However, in Hong Kong the situation is different, their private cars, trucks, buses and taxis account for the majority of the vehicles, and unless using these vehicles, people either walk or take the metro. This is due to the fact that riding a motorbike and especially cycling in Hong Kong is not a very safe transportation method. The bicycle infrastructure is almost non-existent, the vehicle traffic is very heavy and there is no respect for cyclists just as well as for pedestrians. This is the reason why are there very few bicycles in Hong Kong, and why they are still involved in a relatively high amount of crashes.

## 3.3. Mapping process

After getting to know the data by answering the basic questions about the nature of crashes, the actual analysis can be performed. The goal of this thesis is to find the connection between the amount and severity of crashes and the network and land-use properties. To do this, the crashes need to be plotted on a map first.

As a first step before plotting the crashes, a map is needed to plot the data onto. The program used for the mapping purposes is ArcGIS, therefore the base map is provided by the program. However, there is still need for acquiring the layers of land uses and road network. In the case of Hong Kong it is sometimes difficult to find layers like this, so the search had to be extended to whole China. With this method, it was possible to download a shapefile containing the road network of whole China, which then needed to be cut to the extent of Hong Kong. In the case of land use there was a layer found which only covers Hong Kong, so in that case there was no cutting needed. All layers were acquired from OpenStreetMap.

**Figure 13: The Hong Kong road network together with the land-use layer**

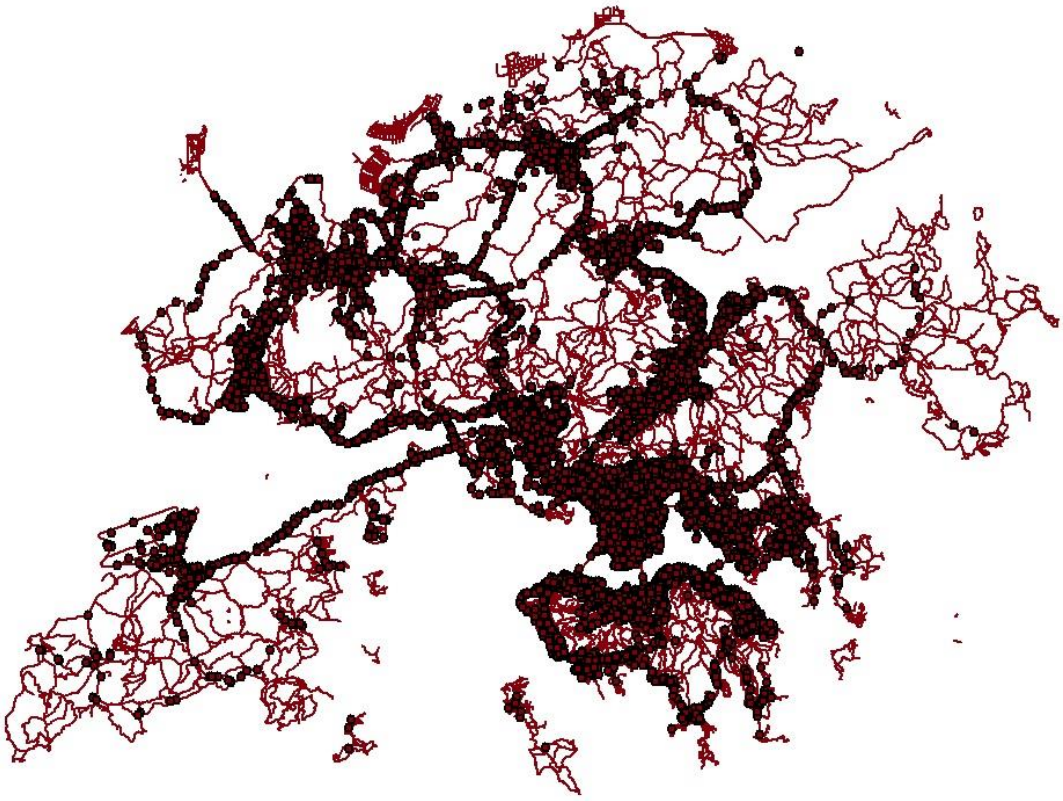


As seen on Figure 13, the land-use layer does not cover 100% of Hong Kong, but the missing parts are generally mountainous, natural areas, therefore this is not a problem from the perspective of the analysis.

After acquiring the required layers, the next step was to plot the crashes. The crash data has a "precise location" field, which describes the location of the crash, but this field does not contain any coordinates, therefore it would be very hard to do the plotting based on this field. Thankfully, the data also contains coordinates, but unfortunately these coordinates do not follow the format of the most commonly used coordinate systems; instead they are provided in meters, showing the distance of each crash from a common origin point. Hong Kong has two local coordinate systems, the Hong Kong 1963 Grid and the 1980 Grid System. Plotting the data in the latter system resulted in a data-cloud that has the shape of Hong Kong's road network, but was located in Laos instead of the right place. Using the 1963 Grid to plot the data seemed like a more successful approach at first, the data was located in Hong Kong, but after more thorough examination it turned out that the data only covered around 1/3 of Hong Kong. Apparently, the reason for this is that the data was in feet, while the coordinate system works with meters. However, even after converting the feet values into meters, the data somehow still did not fit the road network, the crashes were off and this issue did not seem to be fixable with a simple method.

After all the approach with the 1980 Grid System turned out to be working better. After adding 800.000 to all coordinate values in both North-South and East-West directions, the data moved from Laos to the area of Hong Kong, and fit the road network almost perfectly. However, the fit was still not perfect, the crashes visibly followed the road network, but they were still off with some meters in each direction. After examining some crashes which's exact location could be easily determined with the help of the location description, turned out that the whole dataset needed to be moved 180 meters to the south and 240 meters to the east in order to perfectly fit to the road network.

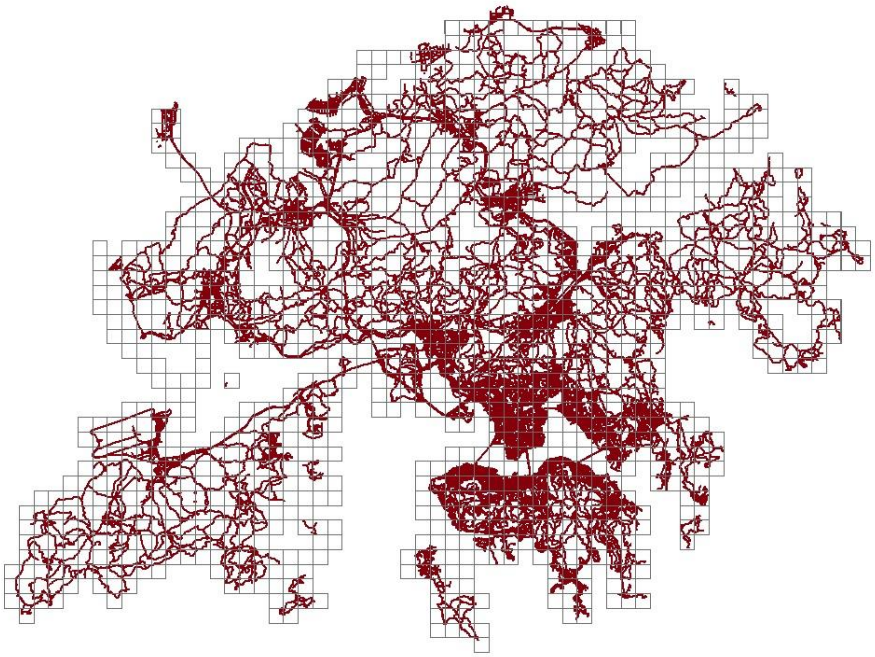
**Figure 14: Crashes on the Hong Kong road network**



As seen on Figure 14, the crashes fit the road network perfectly, with a few exceptions. These exceptions are crashes with false location, but there are so few of these that they will not interfere with the analysis. Unfortunately, 1128 out of the 16170 crashes are without location, therefore these cannot be plotted on the map. Therefore these crashes are only included in the preliminary analysis, where the exact location of the crashes does not have such a high significance.

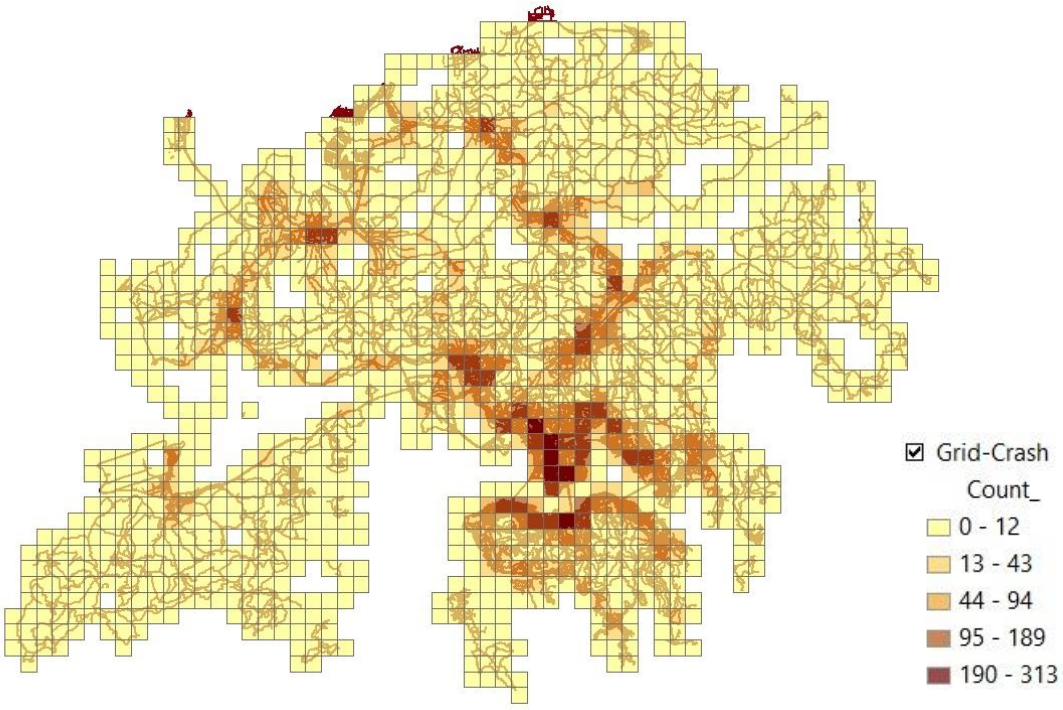
After plotting the crashes, their properties in relation to the local land-use and road network properties had to be analyzed. For this, Hong Kong needed to be split to smaller zones which can be handled more easily. As mentioned in the literature review, instead of using a conventional TAZ-system (Traffic Analysis Zone), for the purpose of this analysis a simple raster grid with 1 $km^2$ raster size will be used, in order to simplify the method leading to the analysis. The advantage of this solution is that the content of each raster will be random, which means that zones with mixed development will also be part of the analysis, as opposed to using TAZs. Furthermore, one of the disadvantages of using TAZs is that the zones are usually bordered by major thoroughfares which on their own are locations for many of the crashes. Therefore, a high number of crashes end up being located between two TAZs. (Siddiqui, 2012) With using a grid network that covers Hong Kong randomly (not based on any geographical layouts) this can be avoided.

**Figure 15: 1x1 km raster grid over Hong Kong**



After creating the grid, each crash was assigned to the cell it is located inside. This provides the opportunity to see which raster cells have the most crashes located within them. This information is shown on Figure 16:

**Figure 16: Number of crashes by zones**



As expected, the raster cells with the highest number of crashes are located in the city center and along the major roadways. This information on its own does not help very much with the analysis. Therefore, the next step is to join the road network and land use layers to the raster cells, in order to see what is exactly located in each cell apart from the crashes.

Before doing this, these two layers will be introduced more thoroughly. As previously mentioned, in this analysis the goal is to find the correlation between land use and road network properties, and crash occurrences. Of course to do this, 3 kinds of data are needed: crash, land use and road network data. The crash data has already been introduced in the previous chapter, but up until now there was no detailed description about the land use and road network data. This data is to be found in the shape of ArcGIS layers, which were acquired from OpenStreetMap.

The road network data is very simple indeed. The shapefile contains the type of road, in some cases the name of the road (this is missing in most cases) and if the road is a one-way road. The length of the road sections has also been added to the data during the analysis.

**Table 5: Original and merged road categories**

| **Road categories** | | |
|---|---|---|
| **Original** | **Merged** | **Ratio (length)** |
| motorway | motorway | 14.06% |
| motorway link | | |
| primary | primary | 12.91% |
| primary link | | |
| trunk | | |
| trunk link | | |
| residential | residential | 14.31% |
| service | service | 9.04% |
| secondary | secondary | 6.52% |
| secondary link | | |
| tertiary | tertiary | 7.72% |
| tertiary link | | |
| construction | disregarded because of insignificance | 34.68% |
| cycleway | | |
| footway | | |
| path | | |
| pedestrian | | |
| steps | | |
| track | | |
| living street | disregarded because of lack of data | 0.76% |
| road | | |
| unclassified | | |

Table 5 shows the categories the different road types were sorted into in the original layer file, and the new categories that have been created by merging some of the original ones. Many categories have been left out of the analysis, partly because they contained roads only usable by pedestrians, therefore were not significant from the point of this analysis, or because they only contained an insignificant amount of data (around 1% of other categories), and based on their names it was impossible to determine which other categories should they be merged into.

The layer containing the land use data has similar fields to the one containing the road network data and it is very detailed. Practically every building, park or any other area has a separate entry. For these entries there are two important fields of information in the layer: the category of the land use and the name of the area, which in this case is almost always filled out (and not missing, as in the case of road sections). Additionally, the size of each land use cluster in $m^2$ has been added during the analysis.

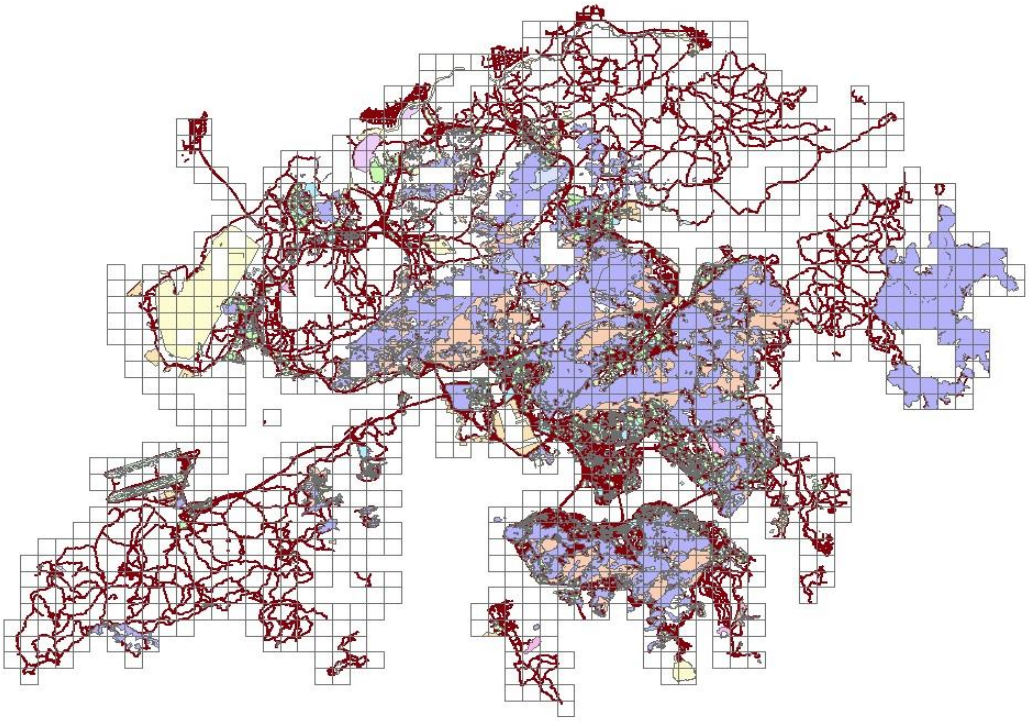
**Table 6: Original and merged land use categories**

| **Land use categories** | | |
|---|---|---|
| **Original** | **Merged** | **Ratio (area)** |
| industrial | industrial | 7.39% |
| residential | residential | 37.10% |
| retail | commercial | 2.51% |
| commercial | | |
| park | recreational | 9.96% |
| recreation ground | | |
| allotments | other (non-daily use) | 43.04% |
| cemetery | | |
| farm | | |
| forest | | |
| grass | | |
| meadow | | |
| military | | |
| nature reserve | | |
| orchard | | |
| quarry | | |
| scrub | | |

As in the case of road categories, the land use categories also had to be merged. In this case also a similar percentage of data had to be disregarded as in case of the road network data, because more than half of the land use categories were land uses not used daily by people, therefore they were not very useful for the analysis. Retail and recreation ground have been merged into other categories because the land uses were similar, but for these two land uses there were barely any occurrences in the data (0.11 and 0.68% respectively).

After discussing the features of road network and land use layers, there is only one step left before the modeling can begin: the crashes, land uses and road sections have to be assigned to the respective grid cells they are located in.

***Figure 17: Land use and road network together with the grid of Hong Kong***



However, there is a problem: as seen on Figure 17, the different areas of land-use and the different road sections are much longer than one raster cell. Assigning them to the cell which covers the most of their area could be a solution, but some land uses cover even as many as 10 cells. We could also assigne them to each and every raster cell they go through, but the problem with this solution is that we would like to know the area and length of the land-uses and road sections within each cell, in order to be able to determine the properties each raster cell has. Therefore, before joining the layers, both the land-use and the road network layers need to be cut to the extent of the raster cells. This is possible in ArcGIS with the command “Intersect”, which cuts every road and land-use area at the exact same spot where the raster cell intersects it, therefore the result will be several pieces of roads and areas which all fit in one or another raster cell.

After this step relating the land-uses and road sections to the grid cells becomes possible, just like we did with the crashes before. When this is done, we have the following information at hand: for each raster cell, it is known how many and what kind of crashes (how serious) happened; how many meters of roads are to be found within the cell from each category; and how many $m^2$ of different land uses are to be found within the cell.

# 4. Model specification

At this point, everything is prepared for the modeling to begin. For the purpose of the analysis, a statistical modeling program called R was used. Poisson and negative binomial models were built up in order to see which one fits the data better.

To build up the models using R, first the information needed to be summarized: in one csv document all grid cells needed to be listed, and then in separate columns the number of light, serious and fatal crashes happening in each cell, the length of each kinds of roads within the cells, and finally the area of different land uses within each cell. This document is readable by R and therefore could become the basis of the analysis. Note: the raster cells containing no roads have been excluded from the analysis.

The analysis was prepared using a Poisson regression model in the first place, and then comparing it to a negative binomial model. Before starting the modeling however, the three crash types needed to be aggregated, otherwise the model would become too complicated. For this, both to fatal and serious crashes a weighing factor was assigned, which shows how many light crashes are one serious or fatal crash worth. There are no fix values for this, but according to the US Department of Transportation (2015), the costs for different kinds of crashes are the following:

- Average fatal crash cost = $6,800,000;
- Average injury crash cost = $390,000;
- Average PDO crash cost = $12,000.

Note that in this comparison not light and serious injury crashes, but PDO and injury crashes are used. Because in our analysis PDO crashes are not taken into account, it has to be assumed that the cost of injury crashes is an average for light and serious crashes together. This would mean that one fatal crash is worth 17.44 average injury crash. In Hong Kong, there are approximately 85% light crashes and 15% serious ones.

$$\frac{\$6{,}800{,}000}{\$390{,}000} = 17.44$$

$$17.44 = 0.15\ x\ Serious + 0.85\ x\ Light$$

Unfortunately, there is not enough data to solve this equation, therefore an assumption had to be made in order to be able proceed further. The assumption was the following: one fatal crash is worth 3 serious crashes. There was no data found to support this assumption (or any other number), but using the number 3 seems like a logical value, especially because using this value the equation will provide realistic numbers:

$$17.44 = 0.15\ x\ 3 + 0.85\ x\ Light$$

$$Light = \frac{17.44 - 0.45}{0.85} = 20$$

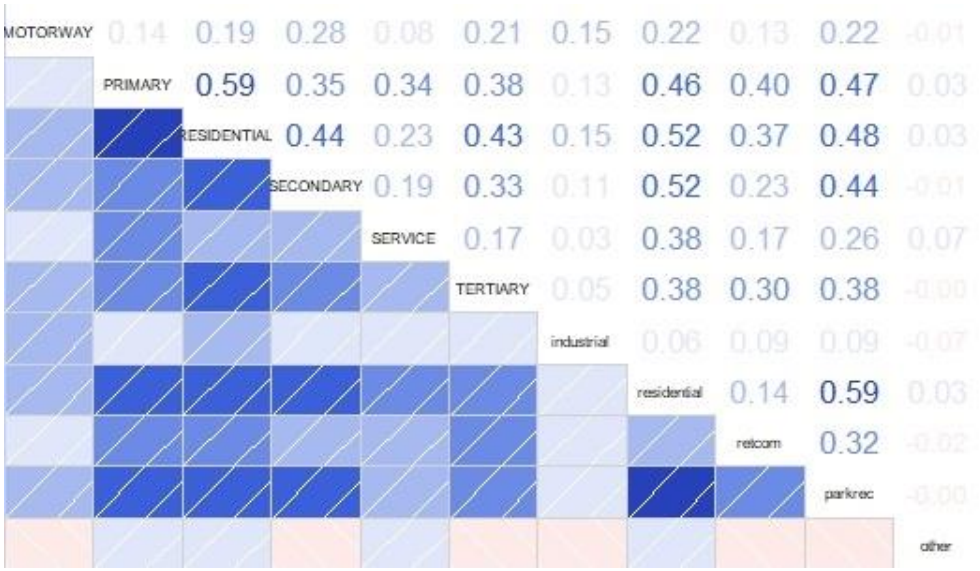
So according to the equation, 20 light crashes would be equivalent to one fatal, which would mean that fatal crashes would have the weight of 20 light crashes. Also, as assumed, 3 serious crashes would be equal to one fatal. This however, would mean that in comparison to light crashes, serious crashes would have the weight of 20/3=6.666. This value was rounded up to 7 in order to make the calculations easier. With doing so, calculating the aggregated number of crashes (KSI) happens the following way:

$$KSI = 20\ x\ Fatal + 7\ x\ Serious + Light$$

After determining the value of KSI, developing the models could begin. In the first case, Poission-regression was used in order to determine the effects of land-use and road network properties onto the crash occurrences. In order to test the goodness of the models, the tests started with the most basic model possible (without any variables), and then the variables were added one by one.

Before that however, a last test needed to be done in order to determine if any of the variables are closely correlated to each other. This is necessary because if two variables are very strongly correlated, it is not allowed to include both of them in the model, because the result will be similar to using the same variable twice. The correlation of the variables has been checked with R's correlogram function and the result can be seen on Figure 18.

**Figure 18: Correlogram**



As seen on Figure 18, primary and residential road categories have the highest correlation to each other, which is understandable since the two categories are also very similar. Furthermore, the primary, residential and secondary road categories are also somewhat correlated to residential and recreational land uses. However, none of these correlations are that high, that it would justify leaving out some of the variables from the models.

Interesting to see that industrial and other land uses are almost completely independent from other variables. This comes as no surprise in the case of the category "other", since that category is a mix of completely different land uses not used by people on daily basis, and because of this it also had to be excluded from the analysis. The category "industrial" is however still included, but according to the correlogram, the industrial spaces (in the case of Hong Kong) do not attract any specific land uses or road categories.

## 4.1. Poisson-model

Now that it has been decided that all variables can be included in the models (except for "other"), the modeling can start with the most basic model possible:

$$\log(Y) = a$$

In this equation, Y is the number of yearly aggregated crashes, and "a" is a constant, which is in this case equal to the average. The model was defined in the following way (using R):

*model1 = glm(formula = aggregatedCrashes ~ 1 , data = dataBase, family = poisson(link=log))*

To determine the models' goodness relative to each other, the Akaike Information Criterion (AIC) was used. The AIC number is an estimator that shows the relative quality of statistical models for the same dataset. The lower the AIC number, the better the fit of the model is for the given set of data. (Moffatt, 2017)

After running the model, the result has the AIC value of 78950. This on its own does not mean anything, the AIC-number only becomes useful if it will be compared to another model's value. In this case by the decrease of the value it can be seen how much has the model improved in comparison to its previous version.

For the second model the length of motorways has been added as variable:

$$\log(Y) = a + b\ x\ l_{motorway}$$

*model2 = glm(formula = aggregatedCrashes ~ 1+ MOTORWAY, data = dataBase, family = poisson(link=log))*

In this case, the AIC number has decreased from 78950 to 72508. This result is obviously better than using no variables in the first model, but the improvement is surprisingly low, which means that adding only the length of motorways as a variable to the model does not make the model much more precise. Let us continue with adding the length of primary roads, too:

$$\log(Y) = a + b\ x\ l_{motorway} + c\ x\ l_{primary}$$

AIC=44245. This however, is a substantial improvement in comparison to the previous models, indicating that the length of primary roads is far more relevant for the goodness of the model than the length of motorways. What happens if we leave the length of motorways out of the equation?

$$log(Y) = a + c\ x\ l_{primary}$$

AIC=48467, meaning that the result is somewhat less precise as together with the length of motorways, but the difference is rather low.

The models that have been prepared in the next steps with different variables and the respective AICs will be summarized in the following table:

**Table 7: Models describing road network properties and their respective AIC-value**

| Road network properties | | |
|---|---|---|
| **Model Nr.** | **Variables** | **AIC** |
| 1 | none | 78950 |
| 2 | 1 + motorways | 72508 |
| 3 | 2 + primary | 44245 |
| 4 | 3 + residential | 35759 |
| 5 | 4 + secondary | 32270 |
| 6 | 5 + service | 31069 |
| 7 | 6 + tertiary | 30147 |

Table 7 contains the models from model Nr. 1 to model Nr. 7. Each model includes one more variable than the previous one. Starting with model Nr. 1 which has no variables, then Nr. 2 with only the length of motorways as variable, Nr. 3 with the length of motorways and primary roads, etc. As seen in Table 7, the biggest effect on the goodness of fit of the model is created by adding the primary and residential roads. In the next step, the same procedure will be done for the land uses.

**Table 8: Models describing land use properties and their respective AIC-values**

| Land use properties | | |
|---|---|---|
| **Model Nr.** | **Variables** | **AIC** |
| 1 | none | 78950 |
| 2 | 1 + industrial | 77544 |
| 3 | 2 + residential | 52826 |
| 4 | 3 + commercial | 45992 |
| 5 | 4 + recreational | 43561 |
| 6 | 5 + other | 43559 |

In the case of land use properties, adding industrial land use has almost no effect on the fit of the model. Residential and commercial land uses have a comparatively high effect, especially the first one. However, adding the land use "other" has practically no effect on the result. This was expected, since "other" contains land uses not used by people on a daily basis. Because of the insignificance of it (together with other reasons mentioned earlier), this category will be excluded from the final model.

After determining that all previously considered variables can be included in the analysis, the final model will be the following:

$$\log(Y) = a + \sum b \; x \; l_{road} + \sum c \; x \; A_{land\ use}$$

## 4.2. Negative binomial model

The same procedure has been performed for the negative binomial model as for the Poisson-model described in the previous chapter, namely starting from the most basic equation, adding variables one by one until reaching the final model. The equations are exactly the same as the ones described before. However, already from the beginning it seemed that the negative binomial model does not fit the data as well as the Poisson. One reason for this was the low change in the AIC-number. The basic model's (without variables) AIC-value of 6956.5 has barely decreased with adding variables, as the full model with all the variables has the AIC of 6493.8. This means that the full model is not much more precise than the one without variables, which indicates that this model might not be a good fit for this batch of data.

To check if this regression is really not a good fit for the data, the final model's results have been inspected. The results were the following:

**Table 9: Results of the final negative binomial model**

| | Estimate | Std. Error | Pr(>\|z\|) | Signif. |
|---|---|---|---|---|
| (Intercept) | 1.53764 | 0.07403 | < 2e-16 | *** |
| MOTORWAY | 0.31247 | 0.0432 | 4.70E-13 | *** |
| PRIMARY | 0.36058 | 0.04313 | < 2e-16 | *** |
| RESIDENTIAL | 0.1765 | 0.03804 | 3.48E-06 | *** |
| SECONDARY | 0.2679 | 0.06694 | 6.29E-05 | *** |
| SERVICE | -0.04206 | 0.0519 | 0.417755 | |
| TERTIARY | 0.29312 | 0.07886 | 0.000202 | *** |
| industrial | 3.82392 | 2.14705 | 0.074911 | . |
| residential | 4.57736 | 1.1611 | 8.07E-05 | *** |
| commercial | -0.07865 | 5.24597 | 0.988039 | |
| recreational | -1.5158 | 3.00257 | 0.613676 | |

As seen in Table 9, many of the land use categorizes have proven not to be significant when using negative binomial regression. Commercial land use would not only be insignificant, but according to the data it would even be negatively associated with the crash occurrences, which seems highly unlikely.

To test if the model is really not a good fit, its estimated results have been compared to the experienced values, and the same has been done with the results of the Poisson-model. These results have been plotted on diagrams. The results for the Poisson and negative binomial distributions are presented in the followings in the same order:

**Figure 19: Poisson-model: Estimated numbers of crashes in relation to observed numbers of crashes**

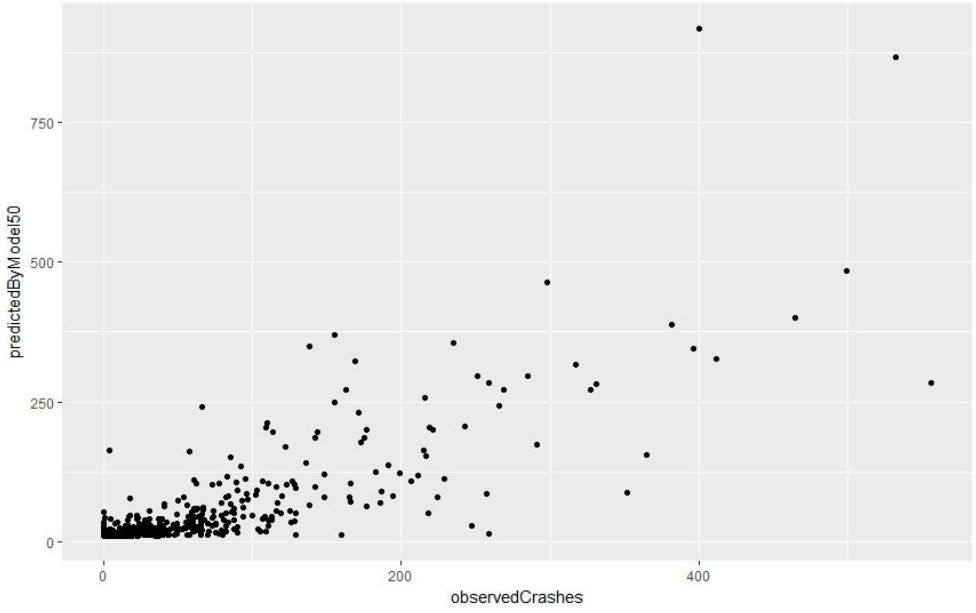
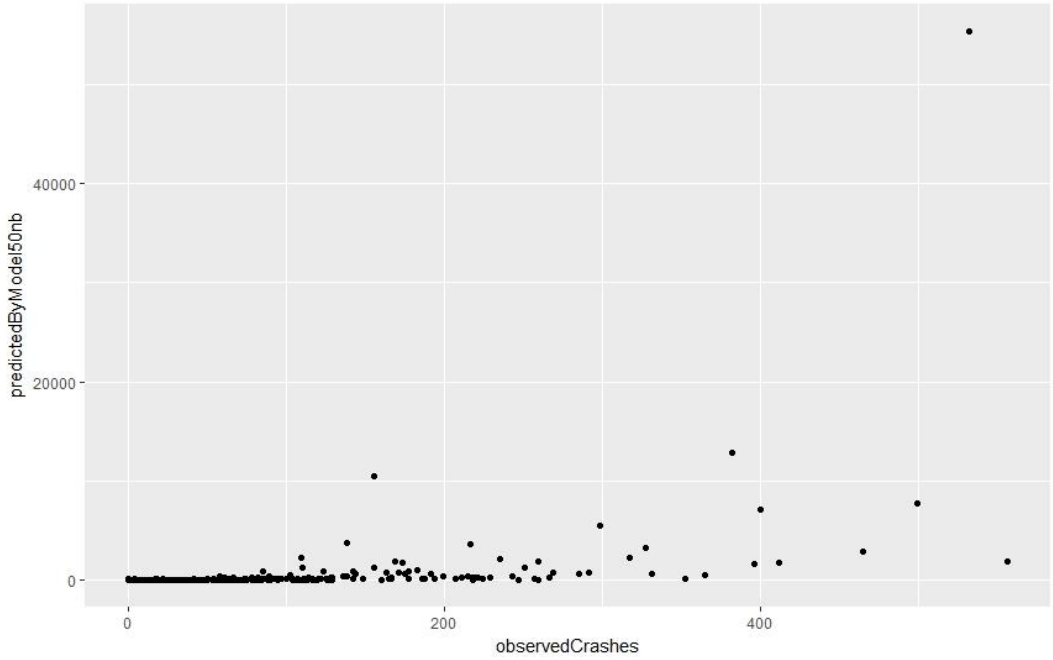
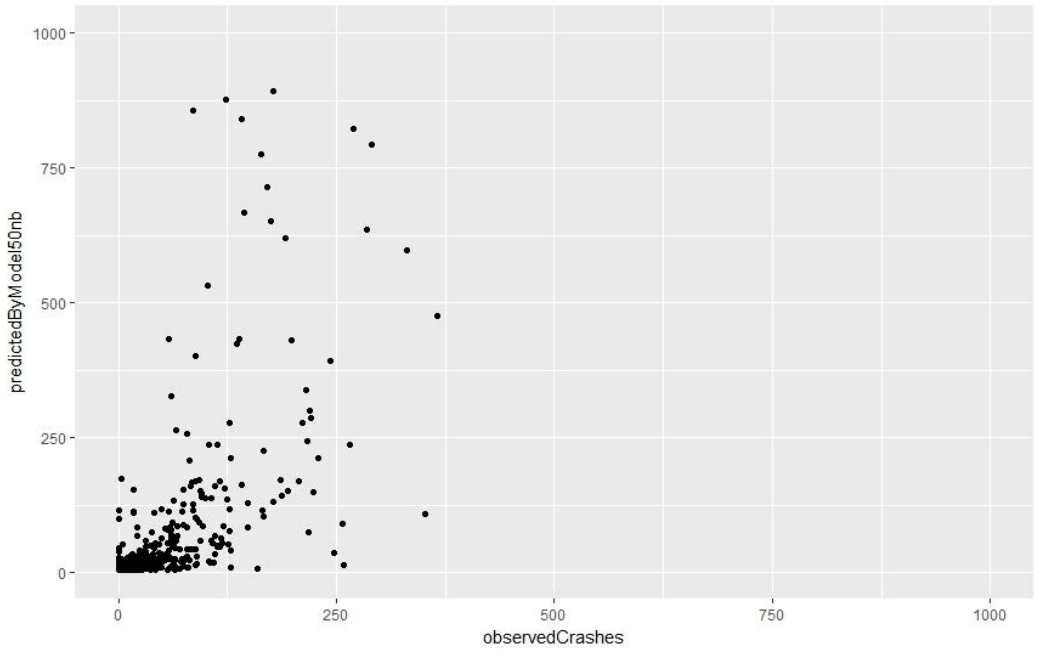


**Figure 20: Negative binomial model: Estimated numbers of crashes in relation to observed numbers of crashes**




It can be clearly seen, that the negative binomial regression has some values which are far off the chart. After filtering out these values, the distribution will look like this:

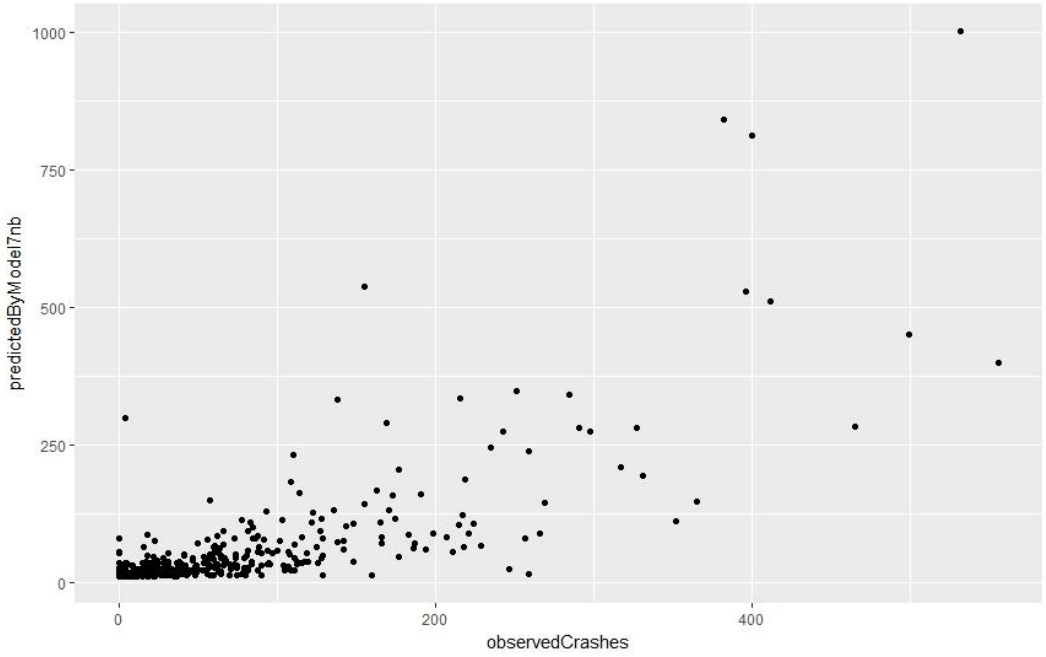
***Figure 21: Negative binomial model: Estimated numbers of crashes in relation to observed numbers of crashes (filtered values)***



After the filtering the data looks much better, but it is obviously still a worse fit than the Possion: the estimated values are much higher in general than in reality.

However, the negative binomial model and the results can be changed with changing the dispersion parameter (theta), which is assumed to be 1 by R if not ordered differently. One can set any value to theta, and as the dispersion parameter changes, the distribution and results will also change. Interestingly, if instead of 1, we set the dispersion parameter to 10, 100, 1000 or 10.000, the results will be closer and closer to reality and to the results of the Poisson distribution. The distribution after using 10.000 as theta can be observed on Figure 22.

***Figure 22: Negative binomial model: Estimated numbers of crashes in relation to observed numbers of crashes (theta = 10.000)***




The results of the negative binomial model with theta set to 10.000 are very similar to the results of the Poisson-distribution. The reason for this is the following: with Poisson-distribution, the variance and the mean are the same. With negative binomial distribution this is not true, but as the dispersion parameter gets larger and larger, the variance converges to the value of the mean, thus turning the negative binomial into a Poisson-distribution (Ford, 2016).

Therefore, it makes no sense for us to use negative binomial distribution, since with unchanged dispersion parameter it does not fit the data, and with increased dispersion parameter it is basically the same as the Poisson-distribution. For this reason, in the next chapter only the results of the Poisson-models will be presented.

# 5. Model estimation and results

In this section the results of the final Poisson-models will be described and explained, since the negative binomial models did not prove to be a good fit for the data. Before proceeding to the final model containing both the land uses and road categories though, models containing only one of the two variable-categories will be examined.

$$\log(Y) = a + \sum b\ x\ l_{road}$$

When only the road network properties are considered, the Poisson-model provides the following result:

**Table 10: Results of Poisson-model with only road categories**

| | Estimate | Std. Error | Pr(>\|z\|) | Signif. |
|---|---|---|---|---|
| (Intercept) | 2.349041 | 0.010281 | <2e-16 | *** |
| MOTORWAY | 0.136916 | 0.002943 | <2e-16 | *** |
| PRIMARY | 0.09099 | 0.002179 | <2e-16 | *** |
| RESIDENTIAL | 0.178239 | 0.002001 | <2e-16 | *** |
| SECONDARY | 0.216296 | 0.003894 | <2e-16 | *** |
| SERVICE | 0.1261 | 0.003811 | <2e-16 | *** |
| TERTIARY | 0.170394 | 0.005491 | <2e-16 | *** |

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

As seen in the last column of Table 10, in the case of the Poisson-model, all variables are significant. Furthermore, based on the "estimate" values, secondary roads are the most prone to attract crashes, followed by residential and tertiary roads. This does not mean however, that these road types are the ones where most crashes happen. Instead, it means that the zone that has the highest length of secondary roads will have the highest number of crashes (not considering other road types).

As a next step, the same analysis has been done for the land use categories:

$$\log(Y) = a + \sum c\ x\ A_{land\ use}$$

**Table 11: Results of Poisson-model with only land use categories**

| | Estimate | Std. Error | Pr(>\|z\|) | Signif. |
|---|---|---|---|---|
| (Intercept) | 2.772827 | 0.008416 | <2e-16 | *** |
| industrial | 4.219497 | 0.148539 | <2e-16 | *** |
| residential | 6.771479 | 0.058158 | <2e-16 | *** |
| commercial | 17.4487 | 0.199016 | <2e-16 | *** |
| recreational | 7.852244 | 0.147798 | <2e-16 | *** |

Again, all variables are significant (just to make sure, first also the category "other" was included in the model, but it did not prove to be significant, just as expected). According to the "estimate" values, commercial land use has by far the highest chance of attracting crashes, followed by recreational and residential land uses, but already with less than 50% chance in comparison to commercial.

Finally, the road network and land use categories will both be included in one model:

$$\log(Y) = a + \sum b\ x\ l_{road} + \sum c\ x\ A_{land\ use}$$

With the following results:

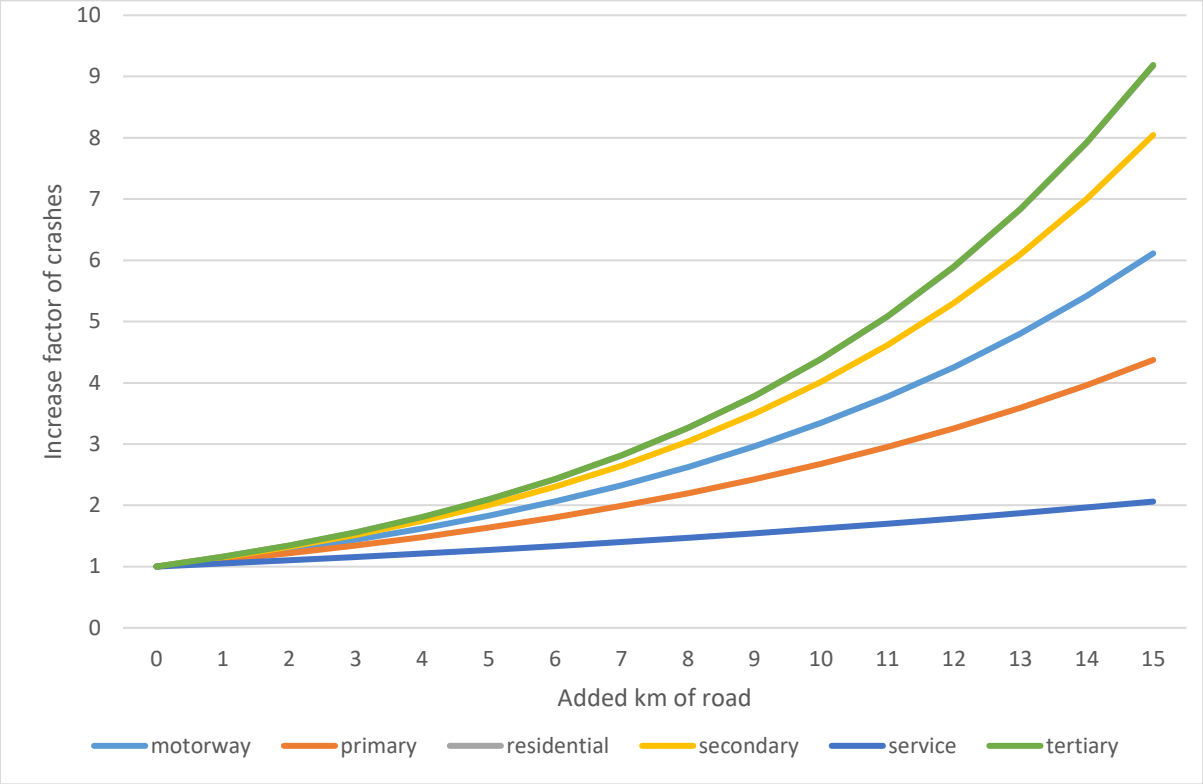
**Table 12: Final results of Poisson-model using both road and land use categories**

| | Estimate | Std. Error | Pr(>\|z\|) | Signif. |
|---|---|---|---|---|
| (Intercept) | 2.256118 | 0.010866 | < 2e-16 | *** |
| MOTORWAY | 0.120694 | 0.002992 | < 2e-16 | *** |
| PRIMARY | 0.098362 | 0.00244 | < 2e-16 | *** |
| RESIDENTIAL | 0.147786 | 0.002304 | < 2e-16 | *** |
| SECONDARY | 0.139008 | 0.004108 | < 2e-16 | *** |
| SERVICE | 0.048227 | 0.004055 | < 2e-16 | *** |
| TERTIARY | 0.147869 | 0.005793 | < 2e-16 | *** |
| industrial | 2.134627 | 0.172651 | < 2e-16 | *** |
| residential | 3.378497 | 0.070639 | < 2e-16 | *** |
| commercial | 1.089956 | 0.272373 | 6.29E-05 | *** |
| recreational | 3.221822 | 0.155879 | < 2e-16 | *** |

When considering the effects of road and land use variables together, the results are somewhat different from the models that considered only one of the two categories. According to the joint model, residential and tertiary roads are both almost equally responsible for attracting the highest number of crashes. These categories are followed very closely by secondary roads, then motorways, and the other categories are falling behind. The order of the road categories and the

constants might have changed somewhat in comparison to the previous model, but the main characteristics are the same: residential, secondary and tertiary roads are the leading group just like before. The effect of adding X km from each types of roads to the increase of yearly crash numbers is presented on Figure 23.

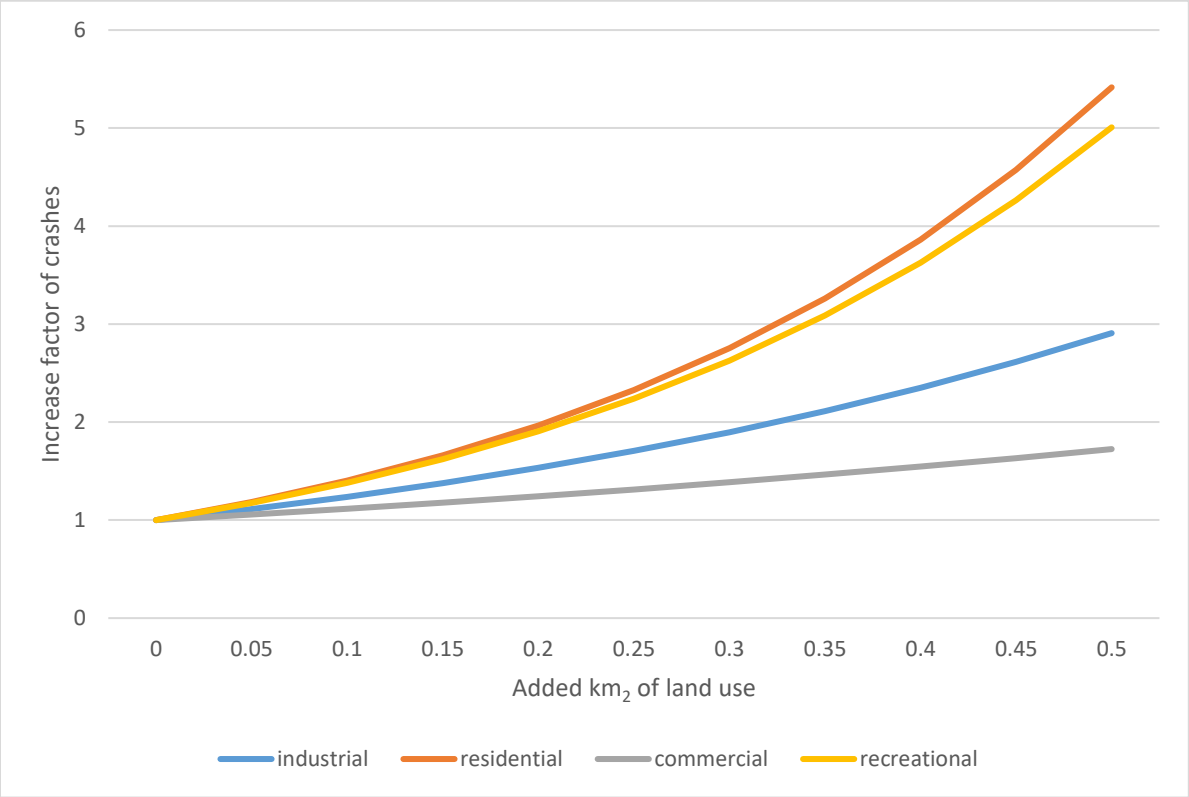
**Figure 23: Increase in the yearly number of aggregated crashes by added km of new roads**




It is important to point out that these numbers are only valid within zones, meaning that Figure 23 shows how many more crashes will happen in a zone, if to that zone X km of roads are added.

Let us proceed to the land uses. In this case, the result has changed much more than in the case of road categories. When considering them separately, commercial had a significant lead on recreational and residential land uses (as in which land use attracts more crashes). According to the joint model however, residential and recreational land uses are leading the group with almost the same values, then industrial comes and commercial falls behind to the last place. These results are shown in detail on Figure 24.

**Figure 24: Increase in the yearly number of aggregated crashes by added $km_2$ of new land uses**




Again, this result is only valid when considering land use and road properties together. Residential land use on its own will not increase the crashes as shown on this chart. However, the new land use most likely comes with such new roads or other characteristics, that will have an effect on the crash numbers. This fact shows that the land uses might be much more strongly affected by the road variables than previously assumed. This observation is also confirmed by the 3 final models' AIC-values:

**Table 13: AIC-values of the final models**

| Model Nr. | Variables | AIC |
|---|---|---|
| Road | road | 30147 |
| Land | land use | 43559 |
| Final | road + land use | 26595 |

As Table 13 shows, the AIC has already dropped to 30147 when only road properties were considered. When the same procedure was done with only the land uses, it increased to 43559 again, showing that the land uses on their own might not describe the situation as well as the road properties. And in the case of the final model, the AIC decreased further in comparison to both models, indicating that the model that considers land uses together with road variables is

still better than the one only considering road properties, but the land uses might not have the same weight in the estimation as the road categories.

## 5.1. Interpretation of results

For the interpretation of the results, first it is important to know that Hong Kong is a special case in many senses. The best example for this is the overwhelming majority of the residential land use (37.1%), while retail and commerce together barely reaches 2.5%. The reason for this is that the land uses are much different in Hong Kong than in any other western country. In the case of most other countries, residential is a low-density land use with mostly family homes. This however does not exist in Hong Kong, or only in a statistically insignificant percentage. The majority of residential areas is very high density indeed, therefore traffic behaves there differently than in other countries would within the same land use. Also because of the scarcity of space, Hong Kong cannot afford to have areas entirely devoted to retail: all retail areas are basically mixed use with retail functions on the lower few floors and residential apartments above those. This is the reason why there is practically no exclusive retail (0.11% of land uses) in Hong Kong.

The next example is related to the previous one: exactly because of the high density, tertiary roads represent completely different types of roads than they would in other countries. Hong Kong's tertiary roads are narrow, comparatively low traffic roads, but because of the extreme high density this low traffic will still be much higher than in some other cities even secondary roads would deal with. In some countries, tertiary roads might still be dirt roads in a low-density area with basically no traffic, however in Hong Kong these roads are up to the newest standards, they are located in the busiest areas and are heavily used.

Another example is the special situation of parks and recreation in general. Recreational land use in Hong Kong mostly consist of parks, and those are also different in Hong Kong than elsewhere. Because of its immense density, Hong Kong has a high number of parks that can provide some peace to people tired of the hustle and bustle of the busy city. However, again because of the density, these parks are relatively small and are surrounded very closely with other, completely different land uses, like residential and commercial land uses in the shape of skyscrapers. Many people living and working in the surrounding skyscrapers will see the nearby park as an escape from their stressful daily life, therefore these parks will attract a high number of crowd, which can shed on a light on why is this land use within the ones attracting the most crashes.

After stating the previous facts, it will be easier to understand why the results turned out to be the way they did. Residential and recreational land uses are the most prone to attract crashes, because of their special situation. Commercial land use does not have high results when

considered together with road categories, but separately it suddenly attracts the highest numbers of crashes. The reason for this is that all commercial land uses are located in the city center at the busiest areas. While residential land use is also widely present in that area, that one is not limited to the city center. Therefore when land uses are the only variables, commercial land use will be associated with the highest numbers of crashes. However, when roads are also considered, the effects of the major roads to crash occurrences turns out to be much higher than that of any land uses, therefore the importance of commercial land use decreases significantly.

In the case of road types, the high rate of crash attraction experienced with secondary roads does not require long explanations, many other studies have found similar results in other regions. What comes as a surprise is the similar rate of tertiary roads. However, as explained earlier, these roads are much more similar to secondary roads in Hong Kong than in other countries, and because most of them are located in the highest density areas, they will be associated with the highest crash numbers.

## 5.2. Comparison to previous research

One of the main purposes of this thesis is to compare the results to results provided by previous research mostly using more complicated models, and determine if simpler models also provide good enough results to use them in Transportation Safety Planning. In the followings the findings of four of the most relevant studies will be introduced shortly and compared to the findings of current study.

Pulugurtha et al. (2013) have found that commercial land use is correlated to others, therefore they did not consider it through the development of the models. It was found that mixed use, residential, business and office land uses were positively correlated with the crash occurrences, while single-family residential land use was negatively correlated with them. Our analysis had similar results: residential, commercial and industrial land uses all proved to be positively correlated with crash occurrences for some extent (note: single-family residential land use is basically non-existent in Hong Kong).

Kim & Yamashita (2002) have investigated the role of land uses in crash occurrences in Hawaii. However, contrary to present analysis, they have chosen a more microscopic approach and they assigned each crash to the land use it was found the closest to. Of course this method also had its limitations, since in many cases major roads have two different land uses on their two sides. They have also used much more detailed land use categories than this analysis. In spite of all this, their findings were still similar to ours: they found that residential and commercial land uses account for around 2/3 of the crashes. In our case when only looking at the land uses, the results are very similar, but when considering road properties too, the results are distorted by the higher

significance of road properties. However, as Kim & Yamashita also mentions, the land use categories in many cases do not cover the reality exactly, and also their findings should not be used out of the context of Hawaii. These are the reasons why the results might differ somewhat from what this analysis has found.

Songpatanasilp et. al (2015) performed a very similar analysis to ours using land use and road properties and a 1 km x 1 km mesh grid, but on the example of Tokyo. They have found that the zero-inflated versions of GLM models represented a better fit than the conventional versions, but the results they provided were very similar. According to their findings, crashes occur most frequently around commercial areas, and least frequently in low-rise residential areas. As Hong Kong does not have low-rise residential areas, it might be hard to compare these results, but their basic findings (higher density = higher chance of crashes) are in line with our findings.

Finally, Guo et al. (2015) have also performed a similar analysis to ours and also using Hong Kong as an example. However, instead of road properties, they were analyzing street patterns together with land use properties and some other variables. The difference is that their model used much more data than ours: apart from crash, land use and road pattern data, they have also obtained speed data from 480 GPS-equipped taxis traveling on the roads of Hong Kong and also vehicle hours. Furthermore, they have divided the analysis to 6 sections by time of day. In their analysis they consider 4 land use categories: residential, commercial, mixed-land and others. They have found that in comparison to residential areas, other land uses have a significantly lower likelihood of attracting crashes. This is also consistent with our results, since their analysis did not have recreational land use in a separate category.

# 6. Conclusion and discussion

The goal of this thesis was to test if crash estimation with only using two variables (land use and road properties) is a viable option in TSP, and the results show that it is.

We started by plotting the crashes in ArcGIS and preparing a grid network in order to be able to analyze the data by grid cells. Then the land use and road network layers were assigned to the grid layer together with the crashes, which provided us with the information of how many $km^2$ of different land uses and how many km of different roads are related to how many and how serious crashes.

During the analysis Poisson and negative binomial models were used, but the negative binomial model did not seem good fit for the data, therefore at the end the results of the Poisson-model were presented.

The results showed the followings: when only considering land uses, commercial land use is the most prone to attract crashes into a grid cell, followed by residential and recreational land uses. When only considering roads, the same is true for secondary roads followed by residential and tertiary ones. When considering roads and land uses together, roads seem to have a more significant effect on crash occurrences than land uses; in this case also the same three road categories are the most prone to attract crashes. However, in this case residential and recreational land uses seem to be more important in crash occurrences than others.

The reason why the results turned out to be this way is related to the special circumstances experienced in Hong Kong and to the fact that the land use and road categories contain somewhat different elements there than in other parts of the world. But even with this, the results proved to be more or less similar to the results of other, more complicated researches around the world, like the work of Pulugurtha et al. (2013), Kim & Yamashita (2002), Songpatanasilp et. al (2015) and Guo et al. (2015).

It is important to mention that this research has its own limitations. Because of the special conditions of Hong Kong, these results are not representative for any other location. Therefore the models introduced in this analysis should not be used for other locations without adapting them to the local conditions. The results of this analysis might also be affected by the fact that the land use data did not cover the entirety of Hong Kong. Even though the white spots were mostly located in uninhabited areas, there still might be a chance that also in central areas the land use of some blocks were missing, and this could have affected the final results.

It is also true that the Poisson-distribution does not handle overdistribution well in some cases. In our case this was not a problem, but for each dataset it has to be examined which model represents the better fit, and the one providing better results have to be used. Furthermore, in

our case the model cannot handle cells without any roads or land uses. The cells without any roads have been excluded from the analysis already from the beginning, given that normally where there is no road, there cannot be a crash either. However, it is important to mention that when using this model for the purposes of TSP, the model for the estimation can only be applied for cells containing at least one road.

As the research shows however, it is indeed a viable option to estimate crash numbers only based on land use and road properties, where especially the second one is strongly correlated with the crash occurrences. Even though the results might not be as precise as when using more variables, the estimation is still good enough to show trends. This can be a good opportunity for safety professionals and city planners in less developed countries to perform crash estimation in regions where they have not done this before because of the lack of the data.

Further research should be done in order to determine which regression fits this kind of data the best if the models only consider a low amount of variables. Furthermore, the relationship between these variables and crashes should also be explored in more detail, with putting an emphasis on how important each variable is. Also, research should be done on the effect of local characteristics to the results of the estimation, in order to find out if different variables have different relevancy levels in different locations.

# List of References

Abdel-Aty, M., Lee, J., Siddiqui, C. & Choi, K. (2013). Geographical unit based analysis in the context of transportation safety planning. *Transportation Research Part A: Policy and Practice, 49, 62-75*

Aguero-Valverde, J. & Jovantis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention, 38(3), 618-625*

Association for Safe International Road Travel (2017). Road Crash Statistics. Retrieved from: http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics

Belgian Science Policy Office (n.d). Impact of spatial planning on sustainable traffic safety; Belgian situation analysis. Retrieved from: http://www.belspo.be/belspo/organisation/publ/pub_ostc/mobil/rMD20s_en.pdf

Boland, R. (2017). What Are the Business Hours in Hong Kong? Retrieved from: https://www.tripsavvy.com/business-hours-in-hong-kong-1535496

Boulieri, A., Liverani, S., De Hoogh, K. & Blangiardo, M. (2017). A space-time multivariate Bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society, Statistics in Society: Series A, 180, 119-139*

De Guevara, F. L., Washington, S. & Oh, J. (2004). Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. *Transportation Research Record: Journal of the Transportation Research Board, 1897, DOI: 10.3141/1897-25*

Dissanayake, D., Aryaija, J. & Priyantha Wedagama, D. M. (2009). Modelling the effects of land use and temporal factors on child pedestrian casualties. *Accident Analysis and Prevention, 41(5), 1016-1024.*

Ford, C. (2016). Getting started with Negative Binomial Regression Modeling. Retrieved from: http://data.library.virginia.edu/getting-started-with-negative-binomial-regression-modeling/

GovHK (2017). Hong Kong – the Facts. Retrieved from: https://www.gov.hk/en/about/abouthk/facts.htm

Guo, Q., Pei, X. Yao, D.Y. & Wong, S.C. (2015). Role of street patterns in zone-based traffic safety analysis. *Journal of Central South University, 2015, 22(6), 2416-2422.*

Hadayeghi, A., Shalaby, A. S. & Persaud, B. N. (2003). Macrolevel Accident Prediction Models for Evaluating Safety of Urban Transportation Systems. *Transportation Research Record: Journal of the Transportation Research Board, 1840, DOI: 10.3141/1840-10*

Hadayeghi, A., Shalaby, A. S. & Persaud, B. N. (2007). Safety Prediction Models: Proactive Tool for Safety Evaluation in Urban Transportation Planning Applications. *Transportation Research Record: Journal of the Transportation Research Board, 2019, DOI: 10.3141/2019-27*

Hadayeghi, A., Shalaby, A. S. & Persaud, B. N. (2010). Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. *Accident Analysis & Prevention, 42(2), 676-688*

Hong Kong Observatory (2015). Climate of Hong Kong. Retrieved from: http://www.weather.gov.hk/cis/climahk_e.htm

Huang, H., Abdel-Aty, M. & Darwiche, A. (2010). Country-Level Crash Risk Analysis in Florida. Bayesian Spatial Modeling. *Transportation Research Record: Journal of the Transportation Research Board, 2148, 27-37, DOI: 10.3141/2148-04*

Huang, H., Song, B., Xu, P., Zeng. Q., Lee, J. & Abdel-Aty, M. (2016). Macro and micro models for zonal crash prediction with application in hot zones identification. *Journal of Transport Geography, 54, 248-256*

Ivan, J. N., Wang, C. & Bernardo, N. R. (2000). Explaining two-lane highway crash rates using land use and hourly exposure. *Accident Analysis & Prevention, 32(6), 787-795.*

Khan, G., Qin, X. & Noyce, D. A (2008). Spatial Analysis of Weather Crash Patterns. *Journal of Transportation Engineering, 134(5).*

Kim, K. & Yamashita, E. (2002). Motor Vehicle Crashes and Land Use: Empirical Analysis from Hawaii. *Transportation Research Record: Journal of the Transportation Research Board, 1784, DOI: 10.3141/1784-10*

Kusselson, S. B. (2013). Investigating how land use patterns affect traffic accident rates near frontage road cross sections: a case study on Interstate 610 in Houston, Texas. Retrieved from: https://shareok.org/bitstream/handle/11244/14948/Kusselson_okstate_0664M_13038.pdf?sequence=1

Lee, J. Abdel-Aty, M. & Jiang, X. (2014). Development of zone system for macro-level traffic safety analysis. *Journal of Transport Geography, 38, 13-21*

Marshall, W. E. & Garrick, N. W. (2010). Street network types and road safety: A study of 24 California cities. *URBAN DESIGN International, 15(3), 133-147*

Moffatt, M. (2017). An Introduction to Akaike's Information Criterion (AIC). Retrieved from: https://www.thoughtco.com/introduction-to-akaikes-information-criterion-1145956

Oris, W. N. (2011). Spatial Analysis of Fatal Autombile Crashes in Kentucky. Retrieved from http://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=2122&context=theses

Quddus, M. A. (2008) Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis & Prevention, 40(4), 1486-1497.*

Pulugurtha, S. S., Duddu, V. R. & Kotagiri, Y. (2013). Traffic analysis zone level crash estimation models based on land use characteristics. *Accident Analysis & Prevention, 50, 678-687*

Strauss, T. & Lentz, J. (2009). Spatial Scale of Clustering of Motor Vehicle Crash Types and Appropriate Countermeasures. Retrieved from: http://www.intrans.iastate.edu/reports/Crash_patterns.pdf

Siddiqui, C. K. A. (2012). Macroscopic crash analysis and its implications for transportation safety planning. Retrieved from http://etd.fcla.edu/CF/CFE0004191/Siddiqui_Chowdhury_KA_201205_PhD.pdf

Songpatanasilp, P., Yamada, H., Horanont, T. & Shibasaki R. (2015). Traffic accidents risk analysis based on road and land use factors using GLMs and zero-inflated models. Retrieved from http://web.mit.edu/cron/project/CUPUM2015/proceedings/Content/modeling/320_songpatanasilp_h.pdf

Thomas, P., Morris, A., Talbot, R. & Fagerlind, H. (2013). Identifying the causes of road crashes in Europe. *Annals of Advances in Automotive Medicine, 57, 13-22*

Transport Department – The Government of the Hong Kong Special Administrative Region (2017). Transport figures. Retrieved from: http://www.td.gov.hk/en/transport_in_hong_kong/transport_figures/index.html

US Department of Transportation, Federal Highway Administration (2011). Highway Safety Improvement Program Manual, 3.0 Planning: Countermeasure Identification. Retrieved from: https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec3.cfm

US Department of Transportation, Federal Highway Administration (2015). Highway Safety Improvement Program Manual, 2.0 Planning: Problem Identification. Retrieved from https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec2.cfm

US Department of Transportation, Federal Highway Administration (2017). Transportation Safety Planning (TSP). Retrieved from: https://safety.fhwa.dot.gov/tsp/

Weisstein, E. W. (2017). Bayesian Analysis. Retrieved from: http://mathworld.wolfram.com/BayesianAnalysis.html

World Weather and Climate Information (2016). Averge Monthly Rainy Days in Hong Kong. Retrieved from: https://weather-and-climate.com/average-monthly-Rainy-days,Hong-Kong,Hong-Kong

Yau, E. (2016). Uncertain future for Hong Kong's red and minibuses, born of 1967 riots. *South China Morning Post: Travel & Leisure, 20.06.2016.* Retrieved from: http://www.scmp.com/lifestyle/travel-leisure/article/1976256/uncertain-future-hong-kongs-red-and-green-minibuses-born

Yazdani-Charati, J., Siamian, H. & Ahmadi-Basiri, E. (2014). Spatial Analysis and Geographic Variation of Fatal and Injury Crashes in Mazandaran Province from 2006 to 2010. *Materia Socio Medica, 26(3), 177-181., doi: 10.5455/msm.2014.26.177-181*

## List of Abbreviations

| | |
|---|---|
| AIC | Akaike Information Criterion |
| BG | Block Group |
| CT | Census Tract |
| GLM | Generalized Linear Model |
| GWPR | Geographically Weighted Poisson Regression |
| HKSAR | Hong Kong Special Administrative Region |
| KSI | Killed or Seriously Injured |
| LRTP | Long Range Transportation Plan |
| TAZ | Traffic Analysis Zone |
| TSAZ | Traffic Safety Analysis Zone |
| TSP | Transportation Safety Planning |
| VKT | Vehicle Kilometers Travelled |
| VMT | Vehicle Miles Travelled |

# 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, October 10th, 2017

Robert Kiss