UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
CAMINOS, CANALES Y PUERTOS

Spatial Analysis of Traffic Accidents using GIS. The Case of Banda Aceh, Indonesia

DOCTORAL THESIS

Romi Satria

Civil Engineer

M.Eng. Civil Engineer

Madrid, 2020

DEPARTAMENTO DE INGENIERÍA DEL TRANSPORTE, TERRITORIO Y URBANISMO
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE CAMINOS, CANALES Y PUERTOS

Spatial Analysis of Traffic Accidents using GIS. The Case of Banda Aceh, Indonesia

Author:

Romi Satria

Civil Engineer

M.Eng. Civil Engineer

Supervisor:

María Castro Malpica

Dra. Ingeniera de Caminos, Canales y Puertos

Madrid, 2020

POLITÉCNICA

Tribunal nombrado por el Magfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día ____ de _____ de 2020.

Presidente: _____

Vocal: _____

Vocal: _____

Vocal: _____

Secretario: _____

Suplente: _____

Suplente: _____

Realizado el acto de defensa y lectura de la Tesis el día ____ de _____ de 2020 en la *E.T.S. de Ingenieros de Canales, Caminos y Puertos* de la U.P.M.

Acuerda otorgarle la calificación de:

Madrid, a ____ de _____ de 2020.

EL PRESIDENTE

LOS VOCALES

EL SECRETARIO

Acknowledgements

First, I would like to express my sincere gratitude to my supervisor, professor Maria Castro, for her continuous support, patience, motivation and enthusiasm she has provided me over these years. Her deep knowledge guided me throughout all the steps of this thesis. I could have not imagined a better advisor and mentor for my PhD study. I owe much of the success achieved to her.

My sincere thanks go to professor Becky P. Loo, who kindly welcomed me at the University of Hong Kong for a research visit and whom I have learnt from tremendously. I am also deeply thankful for the opportunity to work, and be part of, a diverse and exciting research group. I would also like to thank to professor Jonathan Aguero-Valverde for his support, guidance and time as well as for sharing his expertise.

The financial support of the Aceh government is gratefully acknowledged. Without their support, the development of this thesis would have not been possible. Thanks also to the Department of traffic police in Aceh province and the Department of public work for providing road and accident data used in this thesis.

I express my warm thanks to my colleagues and friends, for their valuable suggestions and constructive criticism and support over these years.

Last but by no means least, I am grateful to my parents for their continuous encouragement and support during my studies.

Resumen

Como en la mayoría de los países del mundo, los accidentes de tráfico son un problema importante en Indonesia. Actualmente, su coste representa entre el 2,8 y el 3,1% del PIB de Indonesia. En la provincia de Aceh, situada al norte de Sumatra, el número de accidentes de tráfico registrados en la red nacional de carreteras ha aumentado constantemente desde 2001. En 2014 y 2015, el número de accidentes y muertes asociadas en la provincia fue superior a la media nacional. Ha sido una prioridad del gobierno de Indonesia reducir el número de accidentes en sus carreteras.

La mayoría de las metodologías desarrolladas para identificar concentraciones de accidentes de tráfico (puntos negros o tramos negros), no investigan los factores locales involucrados. Reducir eficientemente el número de accidentes requiere más información que simplemente conocer dónde ocurren. De hecho, sólo si se comprenden los factores locales que más contribuyen a los accidentes de tráfico, los responsables de la toma de decisiones podrán tomar medidas correctoras eficaces.

La tesis desarrolla una metodología para identificar los accidentes de tráfico e identificar los factores que contribuyen a ellos. En total, se analizan dos carreteras convencionales que representan más de 1.050 kilómetros.

En primer lugar, se analizan las diferencias entre la identificación de puntos negros y tramos negros (zonas de concentración de accidentes). Los resultados muestran que los puntos negros más peligrosos identificados se encuentran predominantemente dentro de tramos negros. Además, estos puntos negros y tramos negros se encuentran en áreas urbanas, y principalmente en la carretera del Este (que es la más transitada de la provincia). Los distritos más densamente poblados también parecen presentar el mayor número de puntos negros y tramos negros.

A continuación, utilizando un enfoque espacial bayesiano, se identifican los factores que contribuyen a los accidentes de tráfico y se desarrollan modelos de predicción de accidentes (*Safety Performance Functions*). Entre las variables analizadas, la IMD (intensidad media diaria de tráfico) presenta coeficientes significativos positivos en todos los casos (accidentes totales, accidentes mortales, accidentes con lesiones graves, accidentes con lesiones leves y accidentes con daños materiales). Los coeficientes correspondientes al uso del suelo y al trazado en planta son negativamente significativos en los accidentes totales, pero positivamente significativos en los mortales. Esto indica que la tasa de accidentes totales disminuye en entornos rurales y urbanos-rurales con trazados ondulados y sinuosos, mientras que aumenta la tasa de accidentes mortales en las mismas condiciones de entorno.

Por último, se sugiere una serie de medidas de mejora de la seguridad basadas en los resultados de los análisis. Estas medidas se adaptan al contexto local y tienen en cuenta

la peligrosidad y las características de los tramos de concentración de accidentes. Por ejemplo, se ha comprobado que las zonas urbanas se caracterizan por una elevada tasa de accidentes. Las medidas correctoras sugeridas incluyen la implementación de controles de velocidad más frecuentes y patrullas de tráfico para animar a los usuarios de la carretera a cumplir con las normas de tráfico. En las zonas rurales, se sugiere que las medidas de seguridad vial se centren en las zonas identificadas con altas tasas de accidentes. Hay que tener en cuenta que, aunque los tramos de carreteras de zonas rurales registran un menor volumen de tráfico diario, la atención médica suele ser menos accesible.

La metodología en dos pasos propuesta puede aplicarse con datos e información de calidad limitada y proporcionar resultados que las autoridades puedan usar para empezar a aplicar medidas correctoras. Este aspecto de la metodología es particularmente importante en Indonesia, así como en toda la región de Asia Sudoriental, donde la disponibilidad y el acceso a los datos siguen siendo difíciles. En conjunto, este estudio proporciona a las agencias de transporte de la provincia información específica sobre dónde deben implementarse las mejoras para aumentar las condiciones de seguridad en las carreteras.

Abstract

As in most countries in the world, road crashes are a major issue in Indonesia. Currently, they cost approximately 2.8% to 3.1% of Indonesia's GDP. In the province of Aceh, located in Northern Sumatra, the number of traffic crashes recorded on the national road system has been continuously rising since 2001. In 2014 and 2015, the number of crashes and associated fatalities in the province were higher than the national average. It has been an Indonesian government priority to reduce the number of crashes on its roads.

Most methodologies developed to identify concentrations of road crashes (hot spots or hot zones) do not investigate the local factors involved. Efficiently reducing the number of accidents requires more information than simply knowing where they occur. Indeed, only by understanding what local factors contribute most strongly to road crashes will decision-makers be able to develop effective countermeasures.

This thesis develops a methodology to locate road crashes and factors contributing to them. In total, two highways representing more than 1,050 km of road are analysed.

First, the differences between identifying hot spots and identifying hot zones are analysed. The results show that the most dangerous hot spots identified are predominantly located inside hot zones. In addition, these hot spots and hot zones are located in urban areas and primarily on the east road. The latter is the busiest highway in the province. The most densely populated districts also appear to present the highest number of hot spots and hot zones.

Next, using a Bayesian spatial approach, factors contributing to road crashes are identified and safety performance functions are developed. Amongst the variables analysed, the AADT (Annual average daily traffic) coefficient returned positively significant values in all scenarios (total crashes, fatal crashes, major injury crashes, minor injury crashes and PDO (property damage only) crashes). The coefficients for land use and horizontal alignment were negatively significant in the total crashes scenario but positively significant in the fatal crashes scenario. This indicated that the total crash rate was lower on windy and meandering roads in rural and urban-rural environments, whereas the fatal crash rate was higher in the same environmental conditions.

Lastly, a series of countermeasures are suggested based on the results in the analyses. These countermeasures, tailored to the local context, consider the dangerousness and features of hot zones. For example, it was found that urban areas are characterised by high crash rates. Suggested countermeasures include implementing more frequent speed controls and traffic patrols to encourage road users to comply with traffic regulations. In rural areas, it is suggested that road safety measures should target areas

identified with high crash rates. Additionally, while sections of highways in rural areas record lower daily traffic volumes, medical services tend to be less accessible.

The proposed two-step methodology can be run with limited high-quality data and information and provides results authorities can use to start implementing appropriate countermeasures. This aspect of methodology is particularly important in Indonesia, as in the broader South-East Asian region, where data availability and data access remain a challenge. Overall, this study provides transportation agencies of the province with specific information on where improvements should be implemented to increase road safety conditions.

Errata for PhD thesis

Spatial Analysis of Traffic Accidents using GIS. The Case of Banda Aceh, Indonesia

Romi Satria

- Page 59 : Table 5.1 for hot zone column in Aceh Timur district value 7 :
Change with 10
- Page 102 in paragraph table 5.4 :The AADT coefficient in this scenario was 0.6352, and the mean values of the land use variables 'urban and rural' and 'rural' were 0.3138 and -0.1542 :
mistake is rural , and it should be Urban
- Page 102 In paragraph table 5.4 : The estimates of the horizontal alignment 'meandering' and 'windy' variables were -0.3948 and -0.8052 : mistake are the values, it should be The estimates of the horizontal alignment 'meandering' and 'windy' variables were -0.178 and -0.3948
- Page 102 In paragraph table 5.5: The land use variable 'urban and rural' also appears to be a significant parameter in the spatial model, with a coefficient of 0.3138 for the land use variable: mistake are the values: The land use variable 'urban - rural' also appears negatively significant parameter in the major injury crash model, with a coefficient of -0.1514 for the variable.
- Page 109: In paragraph 'The results for all goodness [...]': Amongst all scenarios, the major injury crashes scenario predicted crashes most accurately, with a difference in DIC values of more than 59 points among all the models. :
Change the sentence with: Amongst all scenarios, the major injury crashes scenario predicted crashes most accurately, with the lowest DIC values among all the models
- Page 109: In paragraph 'The results for all goodness [...]': Similarly, the deviance and WAIC values are lower in the spatial model of major injury crashes (2620.11 and 2584.86, respectively). The higher LPML value for the major injury model (-4536.82)
Change the values: Similarly, the deviance and WAIC values are lower in the spatial model of major injury crashes (744.18 and 1,513.704, respectively). The higher LPML value for the major injury model (-4,972.685)
- Page 109 : Table 5.8 : major injury crashes in LPML column : -4,940.098, change with -4,972.685
- Page 109 : Table 5.8 : Minor injury crashes in LPML column : 4,972.685, change with -4,940.098
- Page 122 : in paragraph in this section : The countermeasures put forward are based on the results obtained from the three methodologies
Change with: The countermeasures put forward are based on the results obtained from the combined approach methodologies

Table of contents

1. INTRODUCTION	1
1.1. Motivation	3
1.2. Statement of the research problem	4
1.3. Objectives	5
1.4. Methodology	5
2. LITERATURE REVIEW	9
2.1. Introduction	9
2.2. Non-spatial approaches to analysing road accidents	10
2.2.1. Poisson regression model	11
2.2.2. Tobit model	12
2.2.3. Fuzzy logic	16
2.2.4. Negative binomial (Poisson-gamma) regression model	18
2.2.5. Poisson lognormal model	21
2.2.6. Zero-inflated model	22
2.2.7. Conway-Maxwell-Poisson model	24
2.2.8. Random effects model	25
2.2.9. Bayesian network	26
2.2.10. Empirical Bayes	30
2.2.11. Generalized additive model	32
2.3. Spatial analysis of accidents	33
2.3.1. Moran's index (MI) statistic	36
2.3.2. Getis-Ord	39
2.3.3. Kernel density estimation (KDE)	41
2.3.4. Network Kernel Density Estimation	44
2.3.5. Kriging	48
2.3.6. Full Bayes hierarchical model with spatial effects	51
2.3.7. Safety performance function (SPF) modelling	55
2.3.8. Spatial random-parameters Tobit model	56
2.3.9. Hot zone	57
2.4. Conclusions	58
3. DATA	61
3.1. Introduction	61
3.2. Accident data	64

3.3.	Road data	69
4.	METHODOLOGY	77
4.1.	Introduction	77
4.2.	Selection of the models.....	77
4.3.	Data preparation	78
4.3.1.	Cleaning of the data	78
4.3.2.	GIS formatting of the data	79
4.4.	Network kernel density analysis	80
4.4.1.	Background.....	80
4.4.2.	Network Kernel Density Estimation (NetKDE)	80
4.4.3.	General framework of NetKDE.....	81
4.5.	Hot zones identification	83
4.5.1.	Background.....	83
4.5.2.	Steps in hot zone identification.....	84
4.6.	Bayesian spatial model.....	85
4.6.1.	Background.....	85
4.6.2.	Development of the models.....	86
4.6.3.	Goodness of fit measures.....	89
5.	RESULTS AND DISCUSSION	93
5.1.	Introduction	93
5.2.	NetKDE approach	93
5.2.1.	Model parameters.....	93
5.2.2.	Results and discussion.....	94
5.3.	Hot zone approach	95
5.3.1.	Model parameters.....	95
5.3.2.	Results and discussion.....	95
5.4.	Bayesian spatial approach.....	100
5.4.1.	Model parameters.....	100
5.4.2.	Spatial model.....	101
5.4.3.	Distribution maps.....	109
5.5.	Comparison of the methods	115
5.5.1.	Comparison of results	115
5.5.2.	Discussion.....	117
5.6.	Proposed combined approach	118

5.7. Accident countermeasures	121
6. CONCLUSIONS AND FUTURE RESEARCH	127
6.1. Conclusions	127
6.2. Future research	130
7. REFERENCES	133
8. Annex A: PICTURES OF SELECTED HIGHWAYS LOCATIONS.....	151
9. Annex B: ADDITIONAL MAPS OF THE NetKDE METHOD	157
10. Annex C: ADDITIONAL MAPS OF THE HOT ZONE METHODOLOGY	163
11. Annex D: SAFETY PERFORMANCE FUNCTIONS (BAYESIAN SPATIAL APPROACH).....	169
12. Annex E: ADDITIONAL MAPS OF THE BAYESIAN SPATIAL METHOD	173

List of tables

Table 2-1: Global road traffic injury mortality (Source: WHO 2003)	10
Table 2-2: Spatial statistical methodologies to analyse accidents using GIS	35
Table 3-1: Definition and description of the variables associated with each road accident record	64
Table 3-2: Description of road inventory variables.....	69
Table 3-3: Accident density per district	73
Table 3-4: Land use proportions along highways in each district.....	74
Table 5-1: Hot zones per district on the east and west roads	96
Table 5-2: Top 20 hot zones by number of fatalities on the east road.....	99
Table 5-3: Results of the spatial model of the total number of crashes scenario for the case study	104
Table 5-4: Results of the spatial model of the fatal crash frequency scenario for the case study	105
Table 5-5: Results of the spatial model of major injury crash frequency scenario for the case study.....	106
Table 5-6: Results of the spatial model of the minor crash frequency scenario for the case study	107
Table 5-7: Results of the spatial model of the PDO of crash frequency scenario for the case study	108
Table 5-8: Summary of goodness-of-fit measures for the spatial models for the case study ..	109
Table 5-9: Examples of countermeasures to address road accident hot zones and mitigate contributing factors.....	124

List of figures

Figure 2-1: Example of Bayesian network topology.....	28
Figure 2-2: Graphical output from the spatial autocorrelation tool (Global Moran's I) (Source: ESRI).....	36
Figure 2-3: Principle of the Moran index	37
Figure 2-4: Before (a) and after (b) running a Getis-Ord tool	40
Figure 2-5: Principle of kernel density	42
Figure 2-6: Principle of network kernel density estimation (NKDE)	45
Figure 2-7: The different components of the SANET toolbox (Source: Okabe et al. 2009)	46
Figure 3-1: Political map of Indonesia (Source: Worldometer)	62
Figure 3-2: Locations of districts in Aceh province (Sumatra island)	63
Figure 3-3: Population density in the different districts of Aceh province	63
Figure 3-4: Distribution of road accidents along the two main highways of the province of Aceh	65
Figure 3-5: Distribution of analyzed records of road accidents on the west road (WR) and the east road (ER)	66
Figure 3-6: Distribution of accidents by severity within each district on the west road (WR) and the east road (ER).....	67
Figure 3-7: Distribution of accidents by lighting conditions within each district on the west road (WR) and the east road (ER).....	68
Figure 3-8: Distribution of accidents by weather conditions within each district on the west road (WR) and the east road (ER).....	68
Figure 3-9: Distribution of the road types in Aceh.....	70
Figure 3-10: Location of the two main highways of the province of Aceh	71
Figure 3-11: Cross-sectional diagram of the East and West highways (Source: Indonesian Ministry of Public Work, Aceh Province).....	72
Figure 3-12: Distribution of the 'urban', urban-rural' and 'rural' land uses on the two main highways.....	75
Figure 4-1: Flowchart for identifying hot zones (Loo, 2009).....	85
Figure 5-1: Location of the hot spots identified on the east and west roads using SANET. The inset zooms in on the district of Lhokseumawe.....	94
Figure 5-2: Location of the hot zones on the east road.....	97
Figure 5-3: Location of the top 20 hot zones by number of fatalities	98
Figure 5-4: Map representing the distribution of the parameters land use 'urban' and land 'urban-rural', both found significant in two scenarios (total crash and fatal injury).....	111
Figure 5-5: Map representing the distribution of the parameter access density, found significant in four scenarios (total crash, major injury, minor injury and PDO)	112
Figure 5-6: Map representing the distribution of the parameter AADT, found significant in four scenarios (fatal injury, major injury, minor injury and PDO)	113
Figure 5-7: Map representing the distribution of the parameter land use 'urban-rural', found significant in the minor injury scenario.....	114
Figure 5-8: Comparison of the locations of hot spots and hot zones identified with different methodologies	116

Figure 5-9: A practitioner's step-by-step guide to undertake an accident study using the combined approach	120
Figure 5-10: Flow chart for the development and planning of road traffic accident countermeasures	122
Figure 8-1: View of a section of the East road in a mountainous area, Aceh Besar district	153
Figure 8-2: View of a section of the East road in a mountainous area	153
Figure 8-3: View of a section of the East road in a rural environment (district of Bireun)	154
Figure 8-4: View of a road section of the East road in a rural environment	154
Figure 8-5: View of a section of the West road in a rural environment	155
Figure 8-6: View of a section of the West road in an urban environment	155
Figure 8-7: View of a section of the West road in the province of Aceh	156
Figure 9-1: Hot spot distribution with NetKDE, with zoom in on Lhokseumawe (East road) ...	159
Figure 9-2: Hazardous sections identified with NetKDE, with zoom in on Lhokseumawe (East road)	160
Figure 9-3: Hot spot distribution with NetKDE, with zoom in on Manggeng (West road)	161
Figure 9-4: Hazardous sections identified with NetKDE, with zoom in on Manggeng (West road)	162
Figure 10-1: Localization of the hot zones identified on the East Road	165
Figure 10-2: Localization of the hot zones identified in Lhokseumawe (East Road)	166
Figure 10-3: Comparison of the distribution of the hot zones (Hot zone method) and hot spots (NetKDE) in Lhokseumawe (East Road)	167
Figure 12-1: Total crash scenario: representation of the significant parameter 'land use' (urban, and urban-rural)	175
Figure 12-2: Total crash scenario: representation of the significant parameter 'land use' (urban, and urban-rural), with the distribution of road accidents	176
Figure 12-3: Total crash scenario: representation of the significant parameter 'access density'	177
Figure 12-4: Total crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents	178
Figure 12-5: Total crash scenario: representation of all significant parameters, without the distribution of road accidents	179
Figure 12-6: Fatal crash scenario: representation of the significant parameter 'land use' (urban-rural)	180
Figure 12-7: Fatal crash scenario: representation of the significant parameter 'land use' (urban-rural), with the distribution of road accidents	181
Figure 12-8: Fatal crash scenario: representation of the significant parameter 'AADT'	182
Figure 12-9: Fatal crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents	183
Figure 12-10: Fatal crash scenario: representation of all significant parameters, without the distribution of road accidents	184
Figure 12-11: Major injury crash scenario: representation of the significant parameter 'access density'	185
Figure 12-12: Major injury crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents	186

Figure 12-13: Major injury crash scenario: representation of the significant parameter 'AADT'	187
Figure 12-14: Major injury crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents	188
Figure 12-15: Minor injury crash scenario: representation of the significant parameter 'land use' (urban-rural)	189
Figure 12-16: Minor injury crash scenario: representation of the significant parameter 'land use' (urban-rural), with the distribution of road accidents	190
Figure 12-17: Minor injury crash scenario: representation of the significant parameter 'access density'	191
Figure 12-18: Minor injury crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents	192
Figure 12-19: Minor injury crash scenario: representation of the significant parameter 'AADT'	193
Figure 12-20: Minor injury crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents	194
Figure 12-21: PDO crash scenario: representation of the significant parameter 'access density'	195
Figure 12-22: PDO crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents	196
Figure 12-23: PDO crash scenario: representation of the significant parameter 'AADT'	197
Figure 12-24: PDO crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents	198

List of abbreviations

AADT: Annual Average Daily Traffic
AIC: Akaike information criterion
AR: Actual crash rate
ASEAN: Association of Southeast Asian Nations
BIC: Bayesian Information Criterion
BN: Bayesian Network
BSU: Basic spatial unit
BYM: Besag York Mollie
CAR: Conditional Auto Regressive
CMF: Crash modification factor
CMP: Conway-Maxwell-Poisson
CPO: Conditional Predictive Ordinate
CR: Critical crash rate
D: Distance
DALYs: Disability adjusted life years
DIC: Deviance Information Criteria
EB: Empirical Bayes
ESCAP: Economic and Social Commission for Asian and the Pacific
FB: Full Bayesian
FIS: Fuzzy Inference System
ESRI: Environmental Systems Research Institute
GA: Generalized additive
GEE: Generalized estimating equation
GI: Getis Ord G_i^*
GIS: Geographic Information System
GMRF: Gaussian Markov Random Field
GNP: Gross National Product
GDP: Gross Domestic Product
GPS: Global Positioning System
HC: Hot cluster
HS: Hot spot
HSM: Highway Safety Manual
HZ: Hot zone
INLA: Integrated Nested Laplace Approximation
KD: Kernel Density
KDE: Kernel Density Estimation
LPML: Log Pseudo Marginal Likelihood
MCMC: Markov Chain Monte Carlo
MI: Moran Index

MLE: Maximum Likelihood Estimation
MVPLN: Multivariate Poisson-Lognormal
NB: Negative Binomial
NB-GE: Negative Binomial-Generalized Exponential
NB-L: Negative Binomial-Lindley
NKDE: Network Kernel Density Estimation
PDO: Property Damage Only
QGIS: Quantum Geographic Information System
RE: Random Effect
RENB: Random Effect Negative Binomial
RGR: Ratio of longitudinal grade and curve radius
SANET: Spatial Analysis on a Network
SPF: Safety Performance Function
TAZ: Traffic Zonal Analysis
TC: Traffic Control
TF: Traffic Flow
UN: United Nations
UNPLN: Univariate Poisson-Lognormal
UTM: Universal Transverse Mercator
WHO: World Health Organization
WAIC: Watanabe-Akaike information criterion
ZI: Zero Inflated
ZIP: Zero Inflated Poisson

1. INTRODUCTION

Every year, the World Health Organization (WHO) publishes a new edition of the *Global status report on roads*, which includes updated figures about road traffic deaths, both globally and broken down by country. Year after year, these figures remain very high, with an estimated 1.35 million annual deaths on roads. This is equivalent to about 3,700 persons killed on roads every day. In addition to these fatalities, every year approximately 50 million people are injured or become disabled as a result of road crashes (WHO, 2018). The economic costs associated with these road crashes range from 2% to 5% of each country's gross domestic product (GDP) (WHO, 2018). Countries of the Global South (less developed countries of South-East Asia, Africa and Latin America) often bear the greatest burden.

Today, road traffic injuries are the eighth leading cause of death across all age groups, and they are the leading cause of death for children and young adults (i.e. age groups of 5 to 29 years) (WHO, 2018). Pedestrians, cyclists and motorcyclists represent a substantial fraction of these deaths. In South-East Asia, riders of motorized two- and three-wheelers represent the majority of the fatalities recorded in the region (43%) (ESCAP, 2010; WHO, 2018). In Thailand, for example, the rate of fatalities related to motorcycles is 24.3 per 100,000; in Myanmar it is 12.9 per 100,000. Too often, and particularly in South-East Asia, these major categories of road users remain neglected when road systems are designed.

While progress in reducing the number of deaths on roads has been observed globally over the past few years, it is not uniform across regions and countries. Africa and South-East Asia record higher rates of road traffic deaths than other regions. Moreover, while middle-to-high economy countries have shown progress towards reducing the number of deaths on roads, the situation in countries with low-to-middle economy has not improved since 2013. In fact, the road traffic accident rates are higher in low-income countries: these countries record 1% of the world's motor vehicles but 13% of the global deaths (WHO, 2018).

This global crisis led to the development of two specific targets in the United Nations (UN) Sustainable Development Goals set up in 2012, which were developed to meet urgent environmental, political and economic challenges globally. The first related target, target 3.6, reads 'By 2020, halve the number of global deaths and injuries from road traffic accidents'. The second target, target 11.2, reads 'By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older persons'. However, and despite some progress observed in some parts of the world, these targets (and particularly target 3.6) are still far from being achieved.

In many countries, infrastructure, vehicle and road design are inadequate, and traffic planning, implementation and operation procedures are poor. Important efforts are required to develop and implement comprehensive measures to improve road traffic safety, curb the number of deaths occurring on roads every year and reduce the economic burden associated with road crashes. These measures are often classified in three main categories: engineering, enforcement and education. Measures related to engineering include improving junction designs and pedestrian walkways and constructing tunnels and bridges to reduce pedestrian–vehicle conflicts. Enforcement measures include better enforcement of speed limits and more frequent vehicle controls and safety inspections. Finally, measures related to education involve, for example, developing road safety campaigns, enhancing the education of young students and children, and designing targeted awareness-raising campaigns for high-risk road user groups. There exists a wealth of literature on various general to more specific measures, mainly coming from developed countries. These best practices could, in many instances, be applied in other countries.

South-East Asian countries, along with many countries in Africa, record higher numbers of deaths on roads compared with other regions in the world (WHO, 2018). Within the Association of Southeast Asian Nations (ASEAN) region, these rates vary across countries. Some countries, such as Singapore, record some of the lowest road fatality rates in the world, whereas countries such as Thailand and Malaysia record very high fatality rates. In low-income countries, such as Myanmar or Cambodia, lower per capita rates of motor vehicles are associated with lower fatality rates. Nonetheless, significant efforts are needed to reduce the number of traffic accidents and improve road safety outcomes in the whole region. These efforts are all the more important because, with these countries' increasing wealth, more people will have access to motor vehicles in the coming decades.

Indonesia is the largest and most populous country in South-East Asia. The 2016 census indicated a population of approximately 261,115,456 inhabitants, with just over half of the population living in urban areas. Almost 90% of the country's 503,604 km of roads

is paved. The road system in Indonesia is classified into four road type categories: national roads (including highways), provincial roads, district roads, and village roads. National roads include all arterial and collector roads in the primary networks that connect the provincial capitals in addition to national strategic roads and highways. Provincial roads include provincial strategic roads and collector roads in the primary road system that connects provincial capitals to district or city capitals or inter-capitals. District roads represent all local roads in the primary road system that connect capital cities with capital districts, inter-district capitals, regent capitals with local activity centres, inter-local activity centres. They also include strategic district roads and public roads in a district's secondary road network. Finally, village roads include all public roads that connect the region on the village as well as environment roads. In 2016, 31,282 road traffic fatalities were reported across all road systems in Indonesia, however, few systematic statistics are available that provide consistent information about traffic crashes on the different road types. Riders of two- and three-wheeled motorized vehicles, bus passengers and pedestrians are the three largest groups represented among traffic. In 2016, cost from road traffic crashes represented 3% of the country's GDP. While the UN Decade of Action for Road Safety has been an opportunity to decrease the number of fatalities on roads, these numbers have trended upward in Indonesia since 2005. Indeed, the number of fatal accidents has been steadily increasing, and the number of minor and major injuries increased by more than 500% between 2001 and 2015. Reducing the number of road crashes and implementing effective countermeasures are therefore high priorities in the country.

1.1. Motivation

Road crashes, and their associated social and economic costs, are a global concern. Improving road safety, curbing the number of crashes on roads, and establishing appropriate countermeasures are therefore priorities in most countries in the world. Indonesia is no exception. In this South-East Asian country, the number of road crashes and the number of fatalities and injuries have been continuously increasing for almost two decades. The economic burden of these crashes on the national economy is estimated to be about 2.9% to 3.1% of the GDP. With a national population steadily increasing, and with almost 25% of the population under 15, the number of persons estimated to access motor vehicles over the next few decades is significant.

Limiting the number of crashes on Indonesian roads will only be possible with better knowledge of the main factors contributing to these crashes within this national context. However, little research has been done on traffic safety in Indonesia, on understanding where and how accidents occur, or on identifying the contributing factors. More globally, there exists a major knowledge gap in these areas between developed countries and countries of the Global South. A large amount of literature is available on

road safety in Europe, Australia or the United States. This has contributed to significant progress in developing new methodologies for road safety and sharing best practices across the world. However, the Western context is often not suited to countries of the Global South. There exists a substantial information gap about road safety in South-East Asia. There are few publications about the status of road safety in Asian countries, and few of these are in English.

Therefore, it is hoped that the present thesis sheds light on trends and characteristics of road crashes in the province of Aceh, Indonesia, and identifies some key factors contributing to these crashes. It is hoped that this information will inform decision-making to develop and implement appropriate measures to reduce the number of fatalities on the main roads of the province.

1.2. Statement of the research problem

Primary roads, including highways, are the busiest roads in the country. Therefore, improving traffic safety on these roads is essential to reducing fatality rates at the national level. Identifying areas on these roads where accidents tend to be concentrated (also called 'black spots') will allow traffic officials to identify, select, and prioritise segments of the road that are the most dangerous. They can then implement countermeasures as well as measures to increase safety in these areas.

Identifying these black spots is at the heart of many safety analyses and traffic improvement programs all around the world. Knowledge of these areas allows decision-makers to maximise the efficiency of their investments in improving traffic safety, which are often constrained by limited national budgets. Various methodologies have been developed to analyse and understand these areas.

Following the first generation of statistical methods to analyse road crashes, the normalisation of the use of geographic information system (GIS) in the field of road traffic safety has been key to developing a new series of spatial methodologies. By including the spatial characteristics of crashes, these methodologies, including those used to analyse black spots, enable the mapping and visualisation of crashes, allowing analysis of them in the context of other local factors and the development of more effective countermeasures.

In Indonesia, however, most existing studies about traffic accidents have been of a statistical nature, relying on available traffic crash databases and volume count techniques. Spatial analyses are seldom used. In particular, there is an important knowledge gap as to the relationships between road design and traffic accidents. This is precisely what this thesis investigates.

1.3. Objectives

The main objective of this thesis is to gain a better understanding of the relationships between traffic accidents and road design on the primary roads of the province of Aceh. Highway roads are the source of heavy traffic and are where the majority of road crashes occur in the province. Understanding how these crashes occur is essential to developing appropriate countermeasures.

From this objective, specific goals were set up within the scope of this thesis. These are as follows:

1. To identify hot spots of road traffic accidents.
2. To analyse the distribution and characteristics of traffic crashes.
3. To analyse local factors contributing to traffic accidents.
4. To propose accident countermeasures.

To achieve the above-defined goals, the following sub-goals are set:

1. Undertake a state-of-the-art review of key methodologies used to analyse road accidents, focusing on the differences between spatial and non-spatial methodologies.
2. Create a spatial dataset, based on available official information, of accidents and road information. A GIS is used to collate, store, digitise, standardise, and quality check this information, which is the central dataset used in the analyses.
3. Test the suitability of different existing spatial tools for the study area and use the results to inform the development of a spatial-statistical methodology most suited to analyse the dataset created in sub-goal 2.
4. Run the appropriate methodologies to analyse the relationships between road crashes and road design within the national and local contexts.
5. Use the results of sub-goal 4 to inform the development of appropriate countermeasures and propose policies to help reduce the number of accidents in the province.

1.4. Methodology

The methodology followed in the elaboration of this doctoral thesis is outlined below:

1. A comprehensive literature review is undertaken and presents some of the main methodologies used to analyse traffic accidents. Both non-spatial and spatial methodologies are reviewed. Traditionally, non-spatial methodologies have been extensively used in the field of traffic safety. Methodologies deriving from the

Poisson family have been amongst the most popular ones. The emergence of GIS has resulted in the development of spatial methodologies that enable the visualization of accident locations on a map. The continuous improvements in computational abilities are now allowing spatial and statistical methodologies to be combined resulting in models that are more sophisticated and precise.

2. The province of Aceh, located in Northern Sumatra, Indonesia, is represented as a case study in this thesis. The analyses, results and countermeasures are developed using road and accident information from the province. This area was chosen because while the total number of traffic accidents has been decreasing in the country since 2012, the number of traffic accidents in the province increased by 70% between 2012 and 2015. Furthermore, in 2014 and 2015 the fatality rate in the province was higher than the national fatality rate. The road spatial dataset was created from quality-checked, up-to-date and government-approved information and is composed a crashes layer and a road layer will serve as a basis for the subsequent analyses. In total, the distribution and characteristics of 2,444 accident records are analysed on 1,053 km of highways.
3. Three methodologies are selected (Network Kernel Density, Hot zone, and Bayesian spatial model) and used to analyse the distribution of traffic accidents, analyse the localisation of hot spots and hazardous road locations, and to identify the main factors contributing to these accidents. These methodologies are selected because they are considered most appropriate to respond to the objectives of this thesis in a South-East Asian context characterized by important gaps in terms of data quality and data availability. The Network Kernel Density method (NetKDE) is used to identify concentrations of crashes (hot spots) on the roads. Network Kernel Density Estimation (NKDE) methods, including NetKDE, represent a new set of kernel density estimation methods based on networks and have been developed to address some of the drawbacks of using a planar kernel for network-based analyses. The hot zones methodology is used to identify hazardous road locations that consist of two or more than one contiguous segment. The rationale behind the concept of hot zone is that some contiguous road segments may be more dangerous than a hot spot alone. To date, the hot zone methodology has mostly been applied in large-scale analyses (for example, network-level analyses). Finally, the Bayesian spatial model is used to identify factors that contribute most to road crashes and to determine appropriate countermeasures that need to be adopted to improve road safety.
4. A combined approach is then developed from these methodologies. More precisely, a combination of the hot zones method and the Bayesian model is proposed. The Bayesian approach is suitable for investigating local factors contributing to road crashes. The hot zone approach considers the spatial dependency of road segments, and thus enables the identification of hazardous road locations that encompass contiguous segments. Combining these two approaches, which have different but complementary purposes, will not only identify the most hazardous sections of a

road but will also identify the factors contributing to road crashes in these locations and weight them by importance so that decision-makers can quickly identify such factors for a given road. The approach is useful because it can be run at different scales, from small road sections to bigger networks, and can handle small data samples and lower-quality datasets. More specifically, the benefits of the combined approach, adopted in this thesis, is that it provides new knowledge about the distribution of crashes on highways in Aceh and enables the development of appropriate countermeasures. The broader benefits of this combined approach are that it can be applied to support decision-making and improve road safety in other countries in South-East Asia, which face similar data availability issues and road fatalities challenges. With the analyses of the results, a series of countermeasures are suggested.

5. Finally, based on the overall findings of this thesis, some conclusions are drawn, and future lines of research are suggested.

It is hoped that the present thesis will not only provide much needed baseline information allowing Indonesian officials in Aceh to start taking appropriate actions to limit the number of accidents on the province's national roads but also constitute a scientific contribution in the field of road crash and traffic safety research at the global level.

Indonesian officials are encouraged to use the findings of this thesis to request additional and more consistent monitoring on these roads and extend the monitoring to other road types (provincial, district and village roads). These latter road types also absorb a large fraction of the daily traffic, yet very little is known about the pattern of crashes on these roads.

Additionally, it is hoped that the findings of this thesis provide an example to other countries in the region in terms of traffic safety and monitoring. Data availability is a common issue across most countries in South-East Asia. The methodology developed here required basic data and information yet provides results for decision-makers to act on.

2. LITERATURE REVIEW

2.1. Introduction

Every day around the world, over 3,000 people die on roads. Low- and middle-income countries account for about 85% of these deaths and for 90% of the annual disability-adjusted life years (DALYs) lost because of road traffic injuries. Globally, road traffic crashes are a leading cause of death among young people and the main cause of death among those aged 15–29 years. Projections show that between 2000 and 2020, road traffic deaths will decline by about 30% in high-income countries but will increase substantially in low- and middle-income countries (WHO, 2018). Without appropriate actions, by 2020 road traffic injuries are predicted to be the third leading contributor to the global burden of disease and injury.

The economic cost of road crashes is estimated to be 1% of gross national product (GNP) in low-income countries, 1.5% in middle-income countries and 2% in high-income countries. The global cost is estimated to be US\$ 518 billion per year (WHO, 2018). Low- and middle-income countries account for US\$ 65 billion of this cost, more than they receive in development assistance. Road traffic injuries place a heavy burden not only on global and national economies but also at the household level. Many families are driven deeply into poverty by the loss of breadwinners and the added burden of caring for members disabled by road traffic injuries.

By contrast, very little money is invested in preventing road crashes. In fact, research has shown that funds spent on research and developments are focused on several health concerns, including road safety. Comparatively little is spent on implementation, even though many interventions that would prevent crashes and injuries are well known, well tested, cost-effective and publicly acceptable.

Traditionally, road safety has been assumed to be the responsibility of the transport sector. In the early 1960s, many developed countries set up traffic safety agencies, usually located within the government's transport department. In general, the public health sector was slow to become involved. Nevertheless, road traffic injuries are a

major public health issue and not just an offshoot of traffic mobility. The global health sector would greatly benefit from better road injury prevention in terms of fewer hospital admissions and a reduced severity of injuries. The health sector would also gain if – with safer conditions on the roads guaranteed for pedestrians and cyclists – more people were to adopt a healthier lifestyle of walking or cycling, without fearing for their safety.

Data compiled by the WHO show an overall global road traffic injury mortality rate of 19.0 per 100,000 persons (table 2.1). Low- and middle-income countries had a rate slightly greater than the global average, while for high-income countries the rate was considerably lower. Most road traffic deaths – 90% – occurred in low- and middle-income countries. Only 10% of road traffic deaths occurred in high-income countries.

Table 2-1: Global road traffic injury mortality (Source: WHO 2003)

Estimated global road traffic injury-related deaths	Number	Rate per 100,000 population	Proportion of total (%)
Low-income and middle-income countries	1,065,988	20.2	90
High-income countries	117,504	12.6	10
Total	1,183,492	19.0	100

2.2. Non-spatial approaches to analysing road accidents

Over the past few decades, the rising number of road accidents globally has led to in the number of studies on the subject. Researchers have highlighted three categories of factors) to be the main causes of road accidents: personal or human behavioural factors, road and environmental factors, and vehicle factors.

The collection and analysis of traffic accident data and information are the main elements contributing to accident studies. Data collection should include all the following parameters: (a) general information about date, time and number of persons involved in the accidents and classification of accidents as fatal, serious or minor; (b) the coordinates and a detailed description of the location of each accident; (c) road and traffic conditions, including for instance, details of the road geometry, characteristics of the surface and traffic data; and (d) financial losses incurred due to property damage, personal injuries and fatalities.

Among alternative methodologies that have evolved over the years for studying accidents, frequencies of crashes have been a feature of various methods and models,

such as Poisson regression, Tobit, negative binomial regression, Poisson lognormal, zero-inflated Poisson, random effects, Conway Maxwell Poisson and generalized additive models. The details of these methods are discussed below, and analyses are provided with a review of each method's strengths and weaknesses.

2.2.1. Poisson regression model

Poisson regression models have been widely used in recent decades as an introductory method of modelling highway crash predictions because they can easily handle the nature of the crash frequency data. These are often described as random events, discrete event, and non-negative integers, and often their distributions are found to be skewed and closer to a Poisson distribution than to other distributions such as the normal distribution.

The Poisson distribution is used to develop collision models because of its discrete event, non-negative and random events. Let Y_i denote the number of collisions at road segment i ($i = 1, 2, 3, \dots, n$). This assumes that the collisions at road segment n are independent events' (equation 2.1).

$$Y_i | \theta_i \sim \text{Poisson}(\theta_i) \quad (2.1)$$

where, θ_i is the Poisson parameter. As shown in equation 2.2, the probability of road segment i having y_i collisions is given by

$$\Pr \{ Y_i = y_i | \theta_i \} = \frac{e^{-\theta_i} \theta_i^{y_i}}{y_i!} \quad (2.2)$$

The Poisson parameter θ_i is generally defined as an exponential function of road segment characteristics, such as exposure, traffic and geometry (Miaou and Lord, 2003), and is usually expressed as shown in equation 2.3.

$$\theta_i = \exp(X_i \beta) \quad (2.3)$$

where

X_i' is a row vector of covariates representing segment-specific attributes and

β is a vector of regression parameters to be estimated from the data.

In a Poisson regression model, it is assumed that the mean and variance of the count variables are constrained to be equal, such that

$$E(Y_i) = \text{Var}(Y_i) = \theta_i$$

Greibe (2003) applied a Poisson regression model to predict the accidents on urban junctions and roads and then to identify factors contributing to road safety and conduct

a safety network analysis for local road authorities. He found that modelling accidents for road links is less complicated than for junctions, probably due to a more uniform accident pattern and a simpler traffic flow exposure or due to a lack of adequate explanatory variables for junctions.

Oh et al. (2006) used a Poisson model to assess factors correlated to railroad crossing crashes. He found that the proximity of crossings to commercial areas and the distance of the train detector from crossings were correlated with greater numbers of accidents. The unique contributions of the paper are the application of the gamma probability model to deal with under-dispersion and the insights obtained regarding railroad-crossing-related vehicle crashes.

Mannering and Bhat (2014) also used a Poisson regression approach to count data as a basis for initial research efforts to identify factors influencing crash frequencies so that effective crash-mitigation designs and policies could be determined. As research progressed, the limitations of the simple Poisson regression model quickly became obvious, and Poisson variants became the dominant methodological approach.

In addition, Lord and Mannering (2010) wrote of the advantage and disadvantages of this method. The advantage is that poisson model is easy to estimate for most basic models, while the disadvantages are that Poisson models cannot handle over- and under-dispersion, they can be adversely affected by low sample means, and they can produce biased results in small sample means.

Hoseipur et al. (2018) studied an MVPLN (multivariate Poisson log normal) spatial model to estimate crash counts for four multi-vehicle collision types, including head-on, rear-end, angle, and sideswipe collisions. They examined and developed the performance of the MVPLN spatial model, a two-stage model and a univariate Poisson lognormal (UNPLN) model. They found that including spatial heterogeneity in the multivariate model significantly improved the model fit, as indicated by the deviance information criterion (DIC). The correlation between crash types is high and positive, implying that the occurrence of a specific collision type is highly associated with the occurrence of other crash types on the same road segment.

2.2.2. Tobit model

The Tobit model is a regression model with a dependent variable which is censored in some way (Tobin, 1958). Anastasopoulos et al. (2008) applied Tobit regression by viewing vehicle accident rates directly (instead of frequencies) as a continuous variable that is left-censored at zero. The findings reveal that many factors relating to pavement condition, roadway geometrics and traffic characteristics significantly affect vehicle accident rates.

Subsequently, Anastasopoulos et al. (2012) estimated a multivariate Tobit model of accident-injury-severity rates that addresses the possibility of differential censoring across injury-severity levels, while also accounting for the possible contemporaneous error correlation resulting from commonly shared unobserved characteristics across roadway segments.

Farah et al. (2009) developed a Tobit model to explain the minimum time to collision that was based on the analysis of drivers' passing decisions on two-lane rural highways using an interactive driving simulator. The authors found that when the lead vehicle speed increases, it is more difficult for the passing vehicle to complete the passing manoeuvre because a longer distance in the opposing lane is required to complete the passing manoeuvre. The result is a shorter time to collision at the end of passing.

The developed model gives insight into the factors affecting the time to collision. It can be applied to estimate the risk level on existing roads or newly designed roads. For example, existing roads with relatively low and risky times to collision could be improved. These types of improvements could include widening of lanes or increasing radii to improve sight distance.

The Tobit regression for modelling crash rates is expressed as equations 2.4, 2.5 and 2.6.

$$Y_{it}^* = \beta_0 + \sum_{m=0}^M \beta_m X_{it}^m + \varepsilon_{it} \quad (2.4)$$

$$Y_{it} = \begin{cases} Y_{it}^*, & \text{if } Y_{it}^* > 0 \\ 0, & \text{if } Y_{it}^* \leq 0 \end{cases}, l = 1, 2, \dots, N, t = 1, 2, \dots, T \quad (2.5)$$

In the above equations, Y_{it} and X_{it}^m are the observed values of the crash rate, and m is the covariate at site i during period t . M , N and T are the numbers of covariates, observed and covariates, respectively. β_0 is the constant, while β_m is the estimate of the coefficient of the covariate m . Y_{it}^* is a latent variable observed only when positive, and ε_{it} denotes the unstructured error which is assumed to follow independently a normal distribution with zero mean and standard deviation σ ($\sigma > 0$):

$$\varepsilon_{it} \sim \text{normal} (0, \sigma^2) \quad (2.6)$$

Debnath et al. (2014) developed a Tobit model using a regression technique to assess the probability and the magnitude of non-compliance with speed limits at various locations in work zones. Speed data were transformed into two groups, continuous for non-compliant drivers and left-censored for compliant, for a Tobit model framework. The model used speed data from three long-term highway work zones in Queensland,

Australia. The authors found that motorists' speeds were significantly affected by the composition and speeding characteristics of surrounding traffic.

Yu et al. (2015) applied a Tobit model to examine the complex effects of weather-related factors on crash occurrence based on safety performance functions. The study highlighted the interaction between roadway geometry, weather and safety risk. It was found that weather-related factors such as visibility and precipitation had complex effects on crash occurrence rate. Indeed, it was shown that their effects not only vary across segments but also are highly correlated with other geometric characteristic variables.

Zeng et al. (2017a) used a Tobit model to analyse crash rates by injury severity. The analysis demonstrated the superiority of the multivariate random parameters of the Tobit model and showed a significant correlation between crash rates and certain risk factors. Notably, like other random-parameters models, the proposed multivariate random-parameters Tobit model may suffer from transferability issues, since the individual parameter vector associated with each observation is unique.

Chand and Dixit (2018) applied a Tobit model to analyse fractal theory for crash rate prediction. Their study made two substantial contributions: i) the application of fractal theory and ii) the estimation of random parameters and latent-class Tobit models for crash rate modelling. Congestion indication, in terms of the Hurst exponent, was found to be significant in both the high- and low-priority segments, suggesting the importance of capturing congestion information in crash prediction models.

Finally, Hezaveh and Cherry (2018) developed a Tobit model to investigate neighbourhood-level factors affecting seat belt use. The results highlighted the potential of using police crash reports as a source for examining seat belt use at the neighbourhood level. Individuals' home addresses extracted from police reports could be used to identify areas with lower seat belt use rates, which could be useful in designing safety campaigns such as 'Click It or Ticket' to efficiently reach individuals who are more prone to lower seat belt use. This method could be more effective than blanket campaigns that tend to show small population-level effects. There is also a need for developing a methodology that enables researchers and safety practitioners to identify seat belt non-use hot spots.

2.2.2.1. Multivariate Tobit model

The Tobit regression, first proposed by James Tobin (1958), is an appropriate method for the analysis of censored data (Anastasopoulos et al., 2008). To accommodate the possible correlation between crash rates at various severity levels, Anastasopoulos et al. (2012b) advocated a multivariate Tobit model. Using a left-censored threshold of zero,

the multivariate Tobit regression for the joint modelling of the crash rate by injury severity is expressed as follows equations 2.7, 2.8 and 2.9.

$$Y_{it}^{k*} = \beta_{k0} + \sum_{m=1}^M \beta_{km} X_{it}^m + \varepsilon_{it}^k \quad (2.7)$$

$$Y_{it}^k = \begin{cases} Y_{it}^{ks}, & \text{if } Y_{it}^{ks} > 0 \\ 0, & \text{if } Y_{it}^{ks} \leq 0 \end{cases}, i = 1, 2, \dots, N, t = 1, 2, \dots, T, k = 1, 2, \dots, k, \quad (2.8)$$

where

Y_{it}^{k*} and Y_{it}^k represent the unobservable and observed crash rate, respectively, at injury severity level k and site i during period t .

N , T and K are the numbers of observed sites, periods and categorized injury severity levels, respectively.

$x_{it}^1, x_{it}^2, \dots, x_{it}^M$ are the observed values of M risk factors at site i during period t ,

while $\beta_{k1}, \beta_{k2}, \dots, \beta_{km}$ are the estimable coefficients corresponding to injury severity level k .

β_{k0} is the constant, and ε_{it}^k denotes the random error term which is assumed to follow a multi-normal distribution with zero mean: that is,

$$\varepsilon_{it} \sim N_k(0, \Sigma), \varepsilon_{it} = \begin{pmatrix} \varepsilon_{it}^1 \\ \varepsilon_{it}^2 \\ \vdots \\ \varepsilon_{it}^k \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} \end{pmatrix}. \quad (2.9)$$

In this equation $kk(k = 1, 2, \dots, K)$ represents the variance of error term ε_{it}^k and σ_{k_1, k_2} ($k_1 \neq k_2$) denotes the covariance between $\varepsilon_{it}^{k_1}$ and $\varepsilon_{it}^{k_2}$.

2.2.2.2. Multivariate random-parameters Tobit model

Several studies have found that heterogeneous effects of certain factors may be influenced by observations of the crash rates (Anastasopoulos et al., 2012a; Caliendo et al., 2015; Ma et al., 2015a; Yu et al., 2015). A random-parameters Tobit model is the common approach applied to deal with this issue. It is reasonable to consider that this phenomenon may also appear in crash-injury-severity rate analysis. Thus, to provide the underlying unobserved heterogeneity in the multivariate Tobit model, the coefficients $(\beta_{k0}, \beta_{k1}, \dots, \beta_{km})$ in equation 2.10 are set to be random parameters $(\beta_{it}^{k0}, \beta_{it}^{k1}, \dots, \beta_{it}^{km})$:

$$Y_{it}^{k*} = \beta_{it}^{k0} + \sum_{m=1}^M \beta_{it}^{km} \times_{it}^m + \varepsilon_{it}^k. \quad (2.10)$$

Because $\beta_{it}^{1m}, \beta_{it}^{2m}, \dots, \beta_{it}^{km}$ may be also correlated, they are assumed to be multi-normally distributed as equation 2.11.

$$\beta_{it}^m \sim N_k(\beta_m, \Sigma_m), \beta_{it}^m = \begin{pmatrix} \beta_{it}^{1m} \\ \beta_{it}^{2m} \\ \dots \\ \beta_{it}^{km} \end{pmatrix}, \Sigma_m = \begin{pmatrix} \sigma_{11}^m & \sigma_{12}^m & \dots & \sigma_{1k}^m \\ \sigma_{21}^m & \sigma_{22}^m & \dots & \sigma_{2k}^m \\ \dots & \dots & \dots & \dots \\ \sigma_{k1}^m & \sigma_{k2}^m & \dots & \sigma_{kk}^m \end{pmatrix}, \quad (2.11)$$

in which β_m and Σ_m are the mean vector and variance-covariance matrix of β_{it}^m , respectively. Notably, if the covariance of two random parameters is statistically insignificant (95% credibility level), they are assumed to follow independent normal distributions (Barua et al., 2016). If the variance of a certain random parameter is not statistically significant (95% credibility level), the random parameter is simplified to be fixed across observations (Anastasopoulos et al., 2012a).

2.2.3. Fuzzy logic

A fuzzy set is defined as the extension of a crisp (classical) set which allows its elements only full membership or no membership (Zadeh, 1965). Fuzzy logic is determined by different levels of truth and by interpretation of linguistic terms. It is appropriate for analysing data likely to be nonspecific (MathWorks, 1999). The two input variables of a fuzzy inference system (FIS) are the distance (D) from the crash point to a candidate segment and the angular difference ($\Delta\theta$) between the intended vehicle direction and the direction of a candidate link. D is defined as the perpendicular distance from the reported location point to the candidate link, if the reported location's projection point is between the two nodes of the link; otherwise it is the minimum of the distances between the crash location and the link's nodes.

Fuzzy set theory extends this concept by defining partial membership. A fuzzy set A on a universe of discourse U is characterized by a membership $\mu_A(x)$ that takes values in the interval [0, 1].

Tayfur (2003) stated that fuzzy logic theory has its main challenge in terms of the rejection of any object from a universal set. The functions may depend on the experts' knowledge. However, triangular and trapezoidal functions are practically preferred as simple linear functions.

Zadeh (1989) listed some of the essential characteristics of fuzzy logic as follows:

- Exact reasoning is viewed as a limiting case of approximate reasoning.
- Everything is a matter of degree.

- Any logical system can be fuzzified.
- Knowledge is interpreted as a collection of elastic or, equivalently, fuzzy constraints on a collection of variables.
- Inference is viewed as a process of propagation of elastic constraints.

The initial process of a fuzzy model is called fuzzification. In this process, all subsets of variables are constructed. If there are no data, intuition and experience can be used in the fuzzification process. By simply looking at the distribution of data of each variable, the obvious clusters can be seen and fuzzified.

If data are available, the dataset is initially portioned into two sets: calibration (training) and verification (testing). The calibration set is used for fuzzification and constructing the fuzzy rules. The verification set is used for testing the accuracy of the model.

Yin et al. (2002) used fuzzy logic to predict the traffic flow in urban areas in real-time conditions across several sites in Hong Kong. The findings of the developed model were significant compared the observed actual data.

Mure and Demichela (2009) developed a model based on fuzzy logic to quantify the accident risk for various industrial activities and to identify effective interventions for reducing the risk.

Madhavan and Cai (2007) suggested a fuzzy-logic controller to adapted anomalies in a signalized intersection because of simplicity and efficiency in traffic conditions. The simulation was performed using the MATLAB toolbox for fuzzy logic, which enables the quick creation of fuzzy rules and allows changes to be made easily. The researchers developed a system of fuzzy rules based on expert knowledge of historical traffic data and accidents that can improve the management of road risk.

Meng et al. (2009) applied a fuzzy method to assess the frequency of road accidents under varying traffic and road conditions. Annual average daily traffic (AADT) and traffic load were used as the key influencing factors in the model. The researchers found that a model using fuzzy inference was reasonable to apply in the study area.

Saranavan et al. (2014) applied a fuzzy model to predict the risk of accidents in a network using the factors road, driver and pedestrian. They developed a new framework for accident risk studies. It was found that improving road conditions could decrease the risk.

Gaber et al. (2016) applied a fuzzy model to predict the accident rate in Aswan using data from 2011 to 2015. They found that a fuzzy logic approach was an appropriate model to deal with uncertain phenomena. The results provided a correlation coefficient of more than 88% when model predictions were compared with accident data. The

analysis was performed to identify the effects of the input factors and to determine the feasibility of applying the proposed model.

Imprealo et al. (2014) applied fuzzy logic for high accuracy mapping. They used crash databases for crash modelling, crash mapping, hazardous road segment analysis and other analysis to reduce the number of crashes. Their method is based on a transformed map-matching candidate segment selection technique and a fuzzy inference system that was developed based on characteristics of the study area (the Strategic Road Network of the United Kingdom) and the limitations of the crash database information (STATS19 reports).

2.2.4. Negative binomial (Poisson-gamma) regression model

The negative binomial (or Poisson-gamma) regression model is the most commonly used model in crash frequency modelling and was introduced as an alternative to the Poisson regression model to account for possible over-dispersion in crash data counts. To address over-dispersion for unobserved or unmeasured heterogeneity is assumed as equations 2.12 and 2.13.

$$\theta_i = \mu_i \exp(\varepsilon_i) \quad (2.12)$$

$$\ln(\theta_i) = \ln \mu_i + \varepsilon_i, \quad (2.13)$$

where

μ_i is an exponential function of segment-specific attributes, such as exposure, traffic and geometric characteristics as displayed equation 2.14.

$$\ln(\mu_i) = \beta_0 + \sum_{m=1}^m \beta_m \times_{mi}, \quad (2.14)$$

where

$\times_{mi}$ denotes the matrix of covariates (relevant geometric and non-geometric road attributes);

m ($m = 1, 2, 3, \dots, M$) is the number of variables;

$\times_{1i} = \ln(L_i)$ and $\times_{2i} = \ln(V_i)$; L_i and V_i represent the length and traffic volume, respectively; and

β_0 is the intercept; and

β_0 denotes the vector of regression coefficients.

The term $\exp \varepsilon_i$ represents a multiplicative random effect due to unobserved heterogeneity (also known as unstructured errors), which follows a gamma distribution

with an inverse dispersion parameter k (also known as the shape parameter) as in equations 2.15 and 2.16.

$$\exp(\varepsilon_i) \mid k \sim \text{Gamma}(k, k) \quad (2.15)$$

The dispersion (or over-dispersion) parameter is usually referred to as

$$\alpha = \frac{1}{k}. \quad (2.16)$$

The probability density function of the negative binomial model is given by equation 2.17.

$$\Pr\{Y_1 = y_1 \mid \mu_1, k\} = \frac{\Gamma(y_1 + k)}{y_1! \Gamma(k)} \left(\frac{k}{k + \mu_1}\right) \left(\frac{\mu_1}{k + \mu_1}\right) \quad (2.17)$$

Under the negative binomial model, the mean and variance are given by equations 2.18 and 2.19.

$$E(Y_1) = \mu_i \quad (2.18)$$

$$\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{k} \quad (2.19)$$

Abdel and Radwan (2000) applied a negative binomial model to assess the frequency of accident occurrence and involvement. They used three years of accident data from a principal arterial in central Florida. It was found that factors contributing significantly to the frequency of accidents were AADT; degree of horizontal curvature; lane, shoulder and median widths; urban vs. rural location; and the section's length. The results showed that heavy traffic volume, speeding, narrow lane width, more lanes, urban location, narrow shoulder width and reduced median width increased the likelihood for accident involvement.

Lord and Mannering (2010) found that the negative binomial model is an extension of the Poisson model to overcome possible over-dispersion in the data. The negative binomial model assumes that the Poisson parameter follows a gamma probability distribution. The model resulted in a closed-form equation, and the mathematics to manipulate the relationship between the mean and the variance structures is relatively simple.

In addition, Mannering and Bhat (2014) stated that the negative binomial model (or Poisson-gamma) became widely used because it can handle over-dispersed data, in which the mean of the frequencies is much greater than the variance and It is also due to crash-frequency databases, which are often found to have many observations with no observed crashes.

Mohammadi et al. (2015) studied negative binomials to investigate the effects of temporal correlation in crash frequency models at the highway level. They used

generalized estimating equation (GEE) and maximum likelihood estimation (MLE) methods to assess the standard errors and the chi-squared values of the variables estimated. They also evaluated the cumulative residual plots for the two models.

Furthermore, Vangala et al. (2015) developed an application of the negative binomial generalized exponential model for analysing traffic crash data to handle at assets characterized by a large number of zero counts that were over-dispersed. Their findings show that for datasets characterized by a sizable over-dispersion and containing a large number of zeros, the Negative binomial generalized exponential performs similarly to the Negative binomial-L but significantly outclasses the traditional Negative binomial model.

Mothafe et al. (2017) applied a negative binomial crash sum model to assess time-invariant heterogeneity on three consecutive years of crash data (2005–2007). Four categories (in addition to total crashes) of crash types were considered in the study: rear end, sideswipe, fixed objects and all other types. The empirical results showed that when time effects are insignificant in short panels such as the one used in this study, the three-year crash sum model is a computationally simpler alternative to a panel model for modelling time-invariant heterogeneity and imposes fewer data requirements such as annual measurements.

Ma et al. (2017) used negative binomials to explore the correlation between crash frequency and potential influencing factors. The accident data were based on a 50-km expressway in China. Longitudinal grade, road width and ratio of longitudinal grade to curve radius, were found to contribute significantly to crash frequency. From the elasticity analysis perspective, wider road width and higher AADT will lead to higher crash frequency. Finally, the fixed-length segment method is superior to the homogeneous longitudinal grade segment method, providing a smaller relative prediction error and cumulative standardized residual. Since the studied expressway is located in a straight and flat area, the fixed-length method is recommended.

The advantages are that it easily overestimates and can explain the dispersion. The weakness is its inability to handle under-dispersed data and dispersion-parameter estimation problems when the data are characterized by low sample mean values and small sample sizes.

Raihan et al. (2019) developed crash modification factors for bicycle crashes for different roadway segments and intersection facility types in urban areas. They found that lane width, speed limit and grass on the median had positive impacts on reducing bicycle crashes. Conversely, the presence of sidewalk and sidewalk barriers were found to increase bicycle crash probabilities. Increased bicycle activity was found to reduce bicycle crash probabilities on segments, whereas increased bicycle activity resulted in higher probabilities at intersections. Bus stops were also found to increase bicycle crash

probabilities at intersections, whereas protected signal controls had a positive impact on bicycle safety. This research provided insight into how various factors affect bicycle safety, a topic that is seldom considered by researchers and practitioners.

2.2.5. Poisson lognormal model

The Poisson lognormal model is an alternative model in accident studies and has been used and discussed by many researchers. The assumption of this model is that accidents occur following a Poisson distribution with a mean expected accident frequency that is log normally distributed. The Poisson lognormal regression model is obtained by the assumption equation 2.20.

$$\exp(\varepsilon_i) \mid \sigma_m^2 \sim \text{Lognormal}(0, \mu_m^2), \quad (2.20)$$

where σ_u^2 denotes the extra Poisson variance or variance for heterogeneous effects (also known as within-site [extra] variation).

The Poisson lognormal model has been used by some researchers as an alternative to the negative binomial model for modelling of crash data (Aguero-Valverde and Jovanis, 2008; Lord and Miranda-Moreno, 2008; Miaou et al., 2003).

Wang et al. (2015) applied the Poisson lognormal model to investigate the relationship of variables to safety and to examine the relationships among geometric design features, land use, traffic volume, average travel speed and urban arterial safety. They found crash frequencies increased with increases in traffic volume, and they observed that higher average speeds were associated with higher crash frequencies during peak periods but not during off peak periods.

Wang et al. (2017) used the Poisson lognormal model to integrate crash frequency and real-time safety analyses of expressway data. The integrated model linked the crash frequency and real-time models by adding the logarithm of the estimated expected crash frequency in the real-time model. The non-integrated model independently estimated crash frequency and real-time crash risk. The integrated model outperformed the non-integrated model, providing much better results for both the crash frequency and real-time models.

Cai et al. (2017) developed a Poisson lognormal model based on the three zonal systems with and without consideration of spatial autocorrelation. Based on the estimated models, they computed predicted crash counts for the three zonal systems. This study focused on comparing zonal systems for crash modelling and transportation safety planning. However, only three zonal systems were adopted for the validation of the proposed comparison method. Extending the current approach to compare other zonal systems (e.g. census blocks and counties) could be meaningful. In addition, it is possible

that the trip distance might be related to the size of appropriate geographic units for crash modelling.

Yu and Abdel-Aty (2013) studied four approaches to formulate Bayesian informative priors for Poisson lognormal models of crashes on a freeway section in Colorado. One approach was the two-stage Bayesian updating approach in which past data is modelled while non-informative priors are assigned to all independent variables and the inverse dispersion parameter. The output means and variances of the variables' posterior distributions act as the informative priors for modelling the training data. Overall, the researchers concluded that the performances of models with informative priors are superior to those with non-informative priors.

Lord and Mannering (2010) stated that although the Poisson lognormal potentially offers more flexibility than the negative binomial/Poisson-gamma, it does have limitations because it cannot handle under-dispersion, and it can be detrimentally influenced by a low sample average and a small-sample size bias (less influenced than Poisson-gamma).

2.2.6. Zero-inflated model

The first zero-inflated model is zero of Poisson model, which concerns a random event containing excess zero-count data in unit time. Zero-inflated Poisson regression is applied to find the relationship between dependent and independent variables when there are many zero values in the dependent variable. This relationship is a mixture between a Poisson model and a logistic model. Let the number of crashes, Y , be distributed according to a zero-inflated Poisson distribution that can be represented as equations 2.21, 2.22 and 2.23.

$$Y = V(1 - B), \quad (2.21)$$

where B is a Bernoulli(p) random variable and V independently of B has a Poisson (θ) distribution. Then the probability mass function of the zero-inflated distribution, here represented as ZIP(p, θ) is given by

$$P(Y = 0) = p + (1 - p)e^{-\theta} \quad (2.22)$$

$$P(Y = k) = (1 - p) \frac{\theta^k e^{-\theta}}{k!} \quad K = 1, 2, \dots, \quad (2.23)$$

where p is the probability of the outcome belonging to a degenerated distribution at zero and is the Poisson rate.

Lord et al. (2005) applied a zero-inflated model to make a suitable crash model. It is distinguished by a dual-state process of regression models, where the observed data can be either a perfect or imperfect state with a mean.

Lord et al. (2007) argued that the zero-inflated model has enhanced statistical fit compared with traditional Poisson and Poisson-gamma models. Although the adoption of such model forms did not actually seem logical to some authors, one might argue that the model has enhanced statistical fit for modelling crash data due to a preponderance of zeros.

Shankar et al. (2003) presented an empirical inquiry into using a zero-inflated model in the predictive modelling of crashes involving pedestrians and motorized traffic on roadways. This study concluded that the Zero inflated Poisson model is most suitable for analysing pedestrian crash contexts. A combination of exposure variables such as average daily traffic, traffic control factors such as traffic signal spacing, network supply factors such as illumination, and the presence of centre turn lanes was found to have a statistically significant impact on pedestrian-vehicle crash probabilities.

Anastasopoulos's (2017) proposed modelling approach accounts for unobserved factors that may differ analytically across segments with and without observed or reported accident injury severities, consequently addressing unobserved zero accident state and non-zero accident state heterogeneity using a zero-inflated model. Zero-inflated count data modelling approaches address the excessive number of zeros inherent in the two sets of dependent variables (accident injury-severity rates and frequencies), which are by nature continuous and discrete count data, respectively, that are left-censored with a clustering at zero.

Lord and Mannering (2010) stated that zero-inflated models have been developed to handle data characterized by a significant number of zeros or by more zeros than one would expect in a traditional Poisson model. Zero-inflated models operate on the principle that the excess zero density that cannot be accommodated by a traditional count structure is accounted for by a splitting regime that models a crash-free versus a crash-prone propensity of a roadway segment.

The zero-inflated model has been popular among transportation safety analysts (Shankar et al., 1995; Carson and Mannering, 2001; Lee and Mannering, 2002; Kumara and Chin, 2003; Shankar et al., 2003). Regardless of its broad applicability to a variety of situations in which the observed data are characterized by high zero densities, Lord et al. (2007) argued that because the zero or safe state has a long-term mean equal to zero, this model cannot properly reflect the crash-data generating process.

Fountas and Anastasopoulos (2018) used a zero-inflated model to analyse accident injury severity outcomes. In injury-severity analysis, an inherent limitation of the traditional ordered probit model arises from the assumption that all zero-injury

accidents are governed by a homogeneous underlying non-accident-generation mechanism. To statistically account for the possible presence of such regimes, which likely reflect the mechanisms of a zero-injury and an ordered injury-severity state, they employed a zero-inflated ordered probit framework. The inflation of the zero-injury observations was accommodated through a joint estimation of a binary probit model and an ordered probit model.

2.2.7. Conway-Maxwell-Poisson model

Conway and Maxwell (1962) introduced the Conway-Maxwell-Poisson distribution (CMP), which is a generalization of the Poisson distribution used for modelling queues and service rates. Additionally the CMP is a generalization of the Poisson distribution that allows it to model both under- and over-dispersed data. The CMP is defined to be the distribution with probability mass function equations 2.24 and 2.25.

$$P(y_i) = \frac{1}{Z(\lambda_i, v_i)} \frac{\lambda_i^{y_i}}{(y_i)_{v_i}} \text{ for } y_i = 0, 1, 2, \dots \quad (2.24)$$

Note that except for the normalization factor $Z(\lambda_i, v_i)$, which is equal to

$$\sum_{n=0}^{\infty} \frac{\lambda_i^n}{(n!)_{v_i}}, \quad (2.25)$$

the distribution is very similar to the Poisson distribution with an extra parameter, i , which can take any non-negative value. The CMP overcomes the requirement that mean and variance be equal by allowing flexibility in modelling the tail behaviour of the distribution. If $i = 1$ and 'If $i < 1$ ', the distribution has longer tails than the Poisson distribution and can model over-dispersed data.

A special case in this situation is when $= 0$ and $< 1 = 0$ and < 1 , for which case the distribution is a geometric distribution, an extreme over-dispersion. If > 1 , the distributions have shorter tails than the Poisson distributions and can model under-dispersed data. Another special case in this situation is when $\rightarrow \infty$ and < 1 , a case in which the distribution is a Bernoulli distribution, an extreme under-dispersion, and the data can only take the values 0 and 1.

Lord et al. (2008) used the CMP to analyse motor vehicle crash data to link the number of crashes to the entering flows at intersections or in segments. The CMP can handle both under-dispersed and over-dispersed data and several common probability density functions in special cases (for example, the geometric distribution, the Bernoulli distribution and the Poisson distribution). This flexibility greatly expands the types of

problems for which the CMP can be used to model crash frequency data. Mannering and Bhat (2014) discussed the variety of other count-data models that have been considered over the years in the CMP.

Furthermore, Lu and Toliver (2016) studied the CMP to investigate and compare various highway rail grade crossing accident-frequency models that can handle the under-dispersion issue. The CMP model generates parameter estimates for explanatory variables that are quite different from those generated by the Poisson regression model, indicating that even though the models provide consistent relative relationships between contributory variables and crash likelihood, the absolute parameter estimates can be significantly affected by model type.

The advantage of the CMP model is its strength with data characterized by under-dispersion. The model's weakness is that it can be negatively influenced by a low sample mean or a small-sample bias, and to date, there have been no multivariate applications of the approach.

2.2.8. Random effects model

Random effects (RE) models assume the data are drawn from a hierarchy of different populations. RE models are used when the common unobserved effects are assumed to be distributed over the spatial or temporal units according to some distribution and shared unobserved effects are assumed to be uncorrelated with explanatory variables.

Chin and Quddus (2003) applied an RE models to identify the elements that affect intersection safety and to establish the suitability of the model using several goodness-of-fit statistics. The RE model was applied to identify the geometric elements, traffic factors and traffic control measures that may influence total accident frequency at signalized intersections in Singapore. It was found that total approach volumes, right-turn volumes, the presence of an uncontrolled left-turn lane, median widths above 2 m, the presence of bus stops, intersection sight distance together with the presence of a surveillance camera, and the number of phases per cycle were associated with higher total accident occurrence.

Caliendo et al. (2014) stated that RE models become necessary when spatial correlations (data collected from the same geographic region) or temporal correlations (data collected on the same observational unit over successive time periods) are suspected to be present. Random effects models are especially appropriate for capturing spatial or temporal correlations.

Chen and Tarko (2014) used an RE model to investigate traffic safety in highway work zones using data from a survey of project engineers and existing datasets. They organized observations in monthly clusters that corresponded to individual work zones,

and they estimated a two-level random-parameters negative binomial model that reflected the structure of the observations. This study concluded that work zone inventory datasets should be maintained by all state transportation agencies for identifying safety issues and potential improvements, and disaggregated data and advanced statistical models should be used to account for time-varying effects and to achieve the best estimates possible.

Ma et al. (2017) used an RE model to investigate the relationship between crash frequency and potential influencing factors in a dataset of accident events occurring on a 50-km expressway in China. Three explanatory variables – longitudinal grade, road width, and ratio of longitudinal grade to curve radius – were found to significantly affect crash frequency. In addition, the authors found that the predicted crash frequency from the RE model was closer to the actual crash frequency.

Lord and Mannering (2010) stated that the advantage of RE models is that they handle temporal and spatial correlation, and the disadvantage is that they may not be easily transferable to other datasets.

2.2.9. Bayesian network

The Bayesian network (BN) is the backbone of the crash prediction method. They are directed a cyclic graph causal influence, where the nodes represent random variables and arcs represent the (usually causal) relationships between variables. The two main components of a BN are causal network models (topology) and conditional probability tables.

Kjaerulff and Madsen (2008) suggested a causal relationship model as a directed acyclic graph. The variable is expressed as the knot in the model and can have a number of states, so the measurement scale choice is the discretion of analyst causal relationship between variables and that are described by an arc between nodes.

Jensen (2001) stated that conditional probability tables describe prior knowledge about the problem domain and explicitly specify causal dependency in terms of a conditional probability distribution. Conditional probability table parameterization is often the most demanding task in the development of a BN, because the number of probabilities can be in the hundreds or even thousands (Druzdzel and Van-der-Gaag, 2000).

A BN encodes a joint probability distribution over a set of n variables $X = \{X_1, \dots, X_n\}$ and avoids the combinatorial explosion problem. Therefore, denote by X_i a random variable and by P_i the set of parent nodes of X_i . Then the joint distribution of X can be expressed as the product of the conditional distributions of each variable given its parents, where x represents an instantiation of X , p_i an instantiation of P_i , and x_i the state of X_i (equation 2.26).

$$P(\times) = \prod_{i=1}^n p(\times_i | \pi_i) \quad (2.26)$$

The conditional probabilities described by the previous equation are presented in some conditional probability tables. When the topology and conditional probability tables have been completed, Bayes' theorem can be used to diagnose a cause given an effect or the chain rule can be used to predict an effect given a number of causes. The theorem is shown in equation 2.27.

$$p(x_i/x_j) = \frac{p(x_j/x_i)p(x_i)}{p(x_j)}, \quad (2.27)$$

where

$p(x_i/x_j)$ = posterior (unknown) probability of x_i given x_j ; $p(x_j/x_i)$ = prediction term for x_j given x_i ; $p(x_i)$ = prior (input) probability of x_i ; $p(x_j)$ = input probability of x_j or, less formally, shown in equation 2.28.

$$\text{Posterior Probability} = \frac{\text{Likelihood.Prior Probability}}{\text{Evidence}} \quad (2.28)$$

The example in figure 2.1 shows two influences on accident risk (AR): namely, traffic flow (TF) and traffic control (TC). Denote by W the AR, M the TF, and S the TC. Their corresponding states are described by w , m and s , respectively. The variables can have any number of states, so the choice of measurement scale is left to the analyst's discretion. Denote by nW , nM and nS the number of states for W , M and S , respectively. In the following sections, the variables are assumed to take three discrete states ($nW = nM = nS = 3$): namely, high (h), medium (m) and low (l). Based on the above example, to diagnose (bottom-up) the probability that traffic flow is m_j given that accident risk was found to be w_k , the Bayes' rule is used in equation 2.29.

$$p(m_j|w_k) = \frac{p(w_k|m_j)p(m_j)}{p(w_k)} \quad (2.29)$$

In predictive reasoning the chain rule is applied to calculate the likelihood that accident risk is w_k , given evidence of traffic flow is m_j and traffic control is s_i , as shown in equation 2.30.

Figure 2-1: Example of Bayesian network topology

$$p(s_i, w_k, m_j) = p(s_i | s_i, m_j) p(m_j) \quad (2.30)$$

Input evidence values are propagated down the network, updating the values of other nodes as explained above. The network predicts the probability of the inquiring variable(s) being in a particular state(s), given the combination(s) of evidence entered.

The BN model is very computationally intensive when the number of topology and state variables increase. Propagation algorithms of recent evidence, however, utilize the topological properties graphical model to reduce the computational complexity (Detcher and Pearl, 1988; Detcher, 2009).

One of the BNs' restrictions is that they must conform to a strict hierarchy for cycle and generates a probability propagation and nonterminating recursive algorithm. This has led to some compromises in modelling, which are partially mitigated by inserting additional input nodes into the model of the cyclic effect, although this increases the complexity of the network and of process control for the algorithm.

Once all the data are obtained, the next step is to find a way to tie them together to determine how they affect the occurrence of accidents. One way to do this is by using a BN model.

Norsys (2000), Pearl (1988), and Grossman and Domingos (2004) have argued that a BN is the right way to simulate a world which is flexible and adaptable to the size of one's knowledge. The network is also a graphical model for interpretation under ambiguity, where the nodes represent variables and variables represent a direct relationship between the nodes.

In addition, Neapolitan (1990) and Pearl (2000) suggested that the networking Bayesian models represent the quantitative potential of the relationship between variables, enabling them to be updated as new information is provided. Furthermore, Sireli et al.

(2002) stated that this model can handle the complicated relationship between traffic accidents and their causes, in addition to predicting how a contributing factor correlated.

A Bayesian network is based on autonomy and the probability of automation, and presents a model of the effect of the contribution or contributions then used to process and analyse the complete dataset with the network (Heckerman et al., 1997). Variables that contribute to each set of conditional probabilities in Bayesian models are used to estimate the probability of traffic accidents. Shuttleworth (2009) said a Bayesian probability model is a method to use probability to predict the likelihood of future event. This probability is usually assessed as a percentage and assesses the level of trust that one has in an idea. The Bayesian probability model is used by most researchers in measuring the level of independence of a given outcome. Bruyninckx (2002) and Myllymaki (2005) stated that when there is uncertainty, the Bayesian model is one of the most important models; it is valuable and allows for quick action. When the tool used is Bayesian probability, results obtained from the model are always reliable regardless of how the information is processed.

One of the main elements of the Bayesian model is that it allows modellers to use common sense and real-world knowledge to eliminate unnecessary complexity. Once all the variables in the model are set, the variables that cause changes in the system are deliberately associated with those variables that influence of the model (Heckerman and Breese, 1994). Myllymaki (2005) showed that the Bayesian network can be built expressly stating how the random variables of a problem of mutual range are dependent. Each joint (node) in the Bayesian system accords with a variable, and the relationship between joints describes a direct dependence between variables. Each node, in essence, has its own distinct conditional probability in the network.

Myllymaki (2005) stated that the calculation of a Bayesian network model can be used to estimate the results that can be obtained from an action and It is easy to update and serves as a model for the system and is changed easily when adjustments are made.

Chen et al. (2018) used a Bayesian network to analyse the time required to handle hazardous material transportation accidents. The authors estimated accident handling times using a Bayesian network, and they found that the network could be used to estimate the probability of different handling time periods and that it was possible to order accident types by level of difficulty. Furthermore, Dadashova et al. (2014) applied a Bayesian approach as a model selection criteria to establish a practical road accident model that can provide better interpretation and prediction performance. The model estimation was carried out using Bayesian inference, and the best model was selected based on goodness-of-fit measures. Their findings from the variable selection process indicated that the factors explaining fatal road accidents were mainly exposure, economic factors and surveillance and legislative measures.

Cheng et al. (2018) empirically evaluated multivariate spatial crash frequency models under a full Bayesian framework to accommodate complex correlation structures. They demonstrated the robustness of models based on the similar signs and closeness of coefficients for the posterior estimates of parameters. The variability in performance of other distance models suggested that caution must be exercised in the choice of exponents. The correlation analysis revealed the presence of positive correlations among the criteria based on training errors, and on cross-validation. Their finding reinforced the importance of selecting the optimum weight matrix for spatial correlation, as a more complex structure may not lead to expected advantages in model performance.

2.2.10. Empirical Bayes

Hauer (1997) conducted a before-after observation using the empirical Bayes (EB) method, a well-accepted approach for evaluating the safety effects of treatments because of the power of statistics. The effectiveness of the treatment of safety is calculated by comparing the number of accidents observed with the number of accidents estimated in the period there after.

Generally, there are two assumptions for EB methods: (a) the frequency of accidents at any site follows a Poisson distribution, and the average population distributions follow a gamma system, and (b) various parameters change from year to year, the same for at all reference sites. In the before-after EB method, the number of accidents that are expected without treatment ($N_{expected,B}$) can be estimated by using equations 2.31 and 2.32.

$$N_{expected,B} = W(N_{predicted,B}) + (1 - w)(N_{observed,B}) \quad (2.31)$$

$$W = \frac{1}{1 + k \times N_{predicted,B}}, \quad (2.32)$$

where

$N_{expected,B}$ = expected crash frequency in the before period,

$N_{predicted,B}$ = predicted crash frequency estimated using the safety performance function (SPF) in the before period,

$N_{observed,B}$ = observed crash frequency in the before period, and

w = a weight factor estimated using the over-dispersion parameter from the negative binomial model (the SPF) and the predicted crash frequency in the before period for the treated site.

The expected crash frequency in the after period can be calculated as a product of the expected crash frequency in the before period and the ratio of the predicted crash frequency in the after period to the predicted crash frequency in the before period. The crash modification factor (CMF) can be estimated as shown in equations 2.33 and 2.34 (Gross et al., 2010; Persaud and Lyon, 2007).

$$CMF = \frac{\left[\frac{N_{observed,A}}{N_{expected,A}} \right]}{1 + \frac{Var(N_{expected,A})}{N_{expected,A}^2}} \quad (2.33)$$

$$Var(N_{expected,A}) = N_{expected,A} \times \frac{N_{predicted,A}}{N_{predicted,B}} \times (1 - w), \quad (2.34)$$

where

$N_{expected}$ = expected crash frequency in the after period,

$N_{predicted}$ = predicted crash frequency estimated using the SPF in the after period, and

$Var(N_{expected,A})$ = the variance of the expected crash frequency in the after period as shown in equation 2.35.

$$Var(CMF) = \frac{CMF^2 \left[\left(\frac{1}{N_{observed,A}} \right) + \left(\frac{Var(N_{expected,A})}{N_{expected,A}^2} \right) \right]}{\left[1 + \frac{Var(N_{expected,A})}{N_{expected,A}^2} \right]^2}, \quad (2.35)$$

where

$Var(CMF)$ = variance of crash modification factor.

Elvik et al. (2017) used EB before evaluating the effect of the safety of new highways in Østfold, Norway in 1996–2002 and 2009–2015. The street was rebuilt from a two-lane divided road into a four-lane divided road. Casualties expected by the model or injured road users were estimated for the period 2000–2005 and adjusted for the period 1996–2002. The adjustment made depended on the assumption that the relationship between the independent variables and the number of accidents or injured road users remained stable over time. So far as this assumption can be tested, it is supported. A new empirical before-after evaluation of the impact on road safety of new highways in Norway found that the number of road users killed or seriously injured was reduced by about 75%.

Wang et al. (2017) used an EB method to develop a modification factor for accidents. The optimal safety performance factors for each category of the accident were

identified based on the results of factorial experiments. The safety performances of different functions were applied to the five categories of impact, and it was found that each category of accident had a different form of SPF. The results showed that signalization significantly decreased angle and left-turn crashes, and CMF had the highest precision.

The main advantage of the EB method is that it can explain the observed changes in the frequency of accidents in the periods before and after, the effect of regression to the mean and changes in the characteristics of the traffic and highways.

2.2.11. Generalized additive model

The generalized additive (GA) model is part of a generalized linear model. Unlike other linear models, this model is able to describe the characteristics of a traffic accident by assuming that the relationship between the explanatory variables used is not linear. In other words, the GA model can describe nonlinear relationships among explanatory variables and statistical interpretation of results.

Lord and Mannering (2010) stated that although the GA model can be more flexible than traditional count models, it has limitations. First, because GA models include more parameters, the estimation process can become very complex, especially when the default values are not used. Second, because GA models use spline functions, they are more difficult to interpret than traditional count models. Third, results of GA models and traditional models are likely to be similar if the explanatory variables are exogenous and the dependent variable has a linear or exponential relationship with the explanatory variables.

Today most of these conventional methods (Poisson regression, Tobit, negative binomial regression, Poisson lognormal, zero-inflated Poisson, RE, CMP and GA models) are not entirely appropriate for accidents studies. As mentioned, the Poisson method assumes that the mean and the variance are equal. However, this assumption is not satisfied in many accident datasets, where heterogeneity across sites is common (Mitra and Washington, 2007).

Furthermore, negative binomial models, CMP models and RE models became widely used because they can handle over-dispersed data, even though they are unable to handle under-dispersed data and dispersion-parameter estimation problems when the data are characterized by low sample-mean values and small sample sizes.

Similarly, GA models and RE models have some difficulties in terms of transferable datasets. Recently, new methods have been used in analysing traffic accidents and have provided satisfactory results. Examples include methods based on decision trees

(Montella et al., 2012; Da Cruz et al., 2017) or fuzzy logic with neural networks (Samadzadegan et al., 2005). Spatial methods also offer a different approach.

2.3. Spatial analysis of accidents

Unlike traditional methods, spatial analysis may, for example, be used to identify the location where a crash occurred, to assess specific distribution patterns, and to identify hot spots through map visualization and the use of statistical techniques. Geographic information systems are powerful tools for spatially analysing accidents.

The use of GISs has rapidly grown in many key sectors, such as engineering, planning, transport and natural resource management. GISs enable users to collect, store, manage and analyse geographic data and display them on a map and to produce useful information by exploring spatial relationships between and within datasets.

The strength of GISs is that, unlike normal databases, they allow users to compare and manipulate data based on their spatial relationships. GIS stores information as thematic layers, all linked by their geographic coordinates. Almost any event that happens, happens somewhere, and it is often important to know where (Longley et al., 2005).

In road accident and hot spot analysis, for example, a GIS is a key tool in minimizing accident hazards by identifying accident locations so that remedial measures can be planned. In addition to enable the visualization of information, it can drive better decision-making by revealing relationships between spatial and attribute data, allowing better decisions about location and improving communication by offering maps and visualisations that greatly assist in explanations and story-telling.

In the field of road accident analysis, many different methodologies using GIS tools have been applied over the last two decades. Kim and Nitz (1995) used a spatial analysis of the nearest-neighbour method to assess motor vehicles in Honolulu at every hour of the day, weekdays and weekends. Quddus (2008) developed a series of relationships between area-wide traffic casualties and contributing factors associated with ward characteristics. He used both non-spatial models (such as negative binomial models) and spatial models and identified the similarities and differences among these relationships.

Gundogdu (2010) identified causal factors and preventive to measure a road crashes reduction. Truong and Somenhali (2011) used spatial analysis to find pedestrian-vehicle crash hot spots and unsafe bus stops.

Effati et al. (2014) developed a geospatial neuro-fuzzy approach for modelling hazardous zones in regional transportation corridors. They used spatial analysis to manipulate spatial data and visualize the relationships between spatial and non-spatial variables related to hazardous zones.

In recent years, many researchers have used a combination of a GIS and statistical analysis to assess road accidents. For example, some studies have used GIS tools using Moran's index (MI) to analyse road accidents (Manepalli and Bham, 2016; Erdogan, 2009; Blazquez and Celis, 2013; Moons et al., 2009; Tortum and Atalay, 2015).

Kernel density (KD) is another method used to analyse spatial patterns of crash data. Erdogan et al. (2008) developed a traffic accident analysis approach to identifying the hot spots and causes of accidents in Turkey. Additionally, Banos and Huguenin-Richard (2000) used KD for traffic accidents and mapped the distribution of child pedestrian accidents. Similarly, Blazquez and Celis (2013) and Plug et al. (2011) used KD to analyse traffic accidents spatially and temporally.

Several other studies have also identified spatial clusters of accidents using KD (Anderson, 2009; Pulugurtha et al., 2007; Çela et al., 2013; Hashimoto et al., 2016; Steenberghen et al., 2004). Finally, the Getis-Ord method has also been used to identify road accident hot spots (Khan et al., 2008). Table 2.2 provides the main published works that have used these spatial and statistical methodologies.

The collection of accurate and comprehensive road accident data as well as contextual road information (e.g. weather conditions and land use) has enabled the development of new and more accurate models, including spatial models.

Past statistical approaches in the field tended to focus more on the temporal aspects (trends over time) without capturing the spatial elements of accidents in a comprehensive manner. The development of work in spatial prediction and spatial autoregressive models marked an increasing use of spatial statistics. This was linked to the increasing availability of more accurate spatial data as well as to improvements in computational powers, enabling the development of more complex models. These include, for example, the use of Bayesian methods with the development of Markov-chain Monte Carlo algorithms, which allow greater flexibility for fitting spatially structured statistical models.

Song et al. (2006) studied spatial modelling and used several Bayesian multivariates to estimate crash rates from different kinds of crashes. Multivariate CAR models are considered to account for spatial effects. Zeng and Huang (2014) proposed a Bayesian spatial joint model of crash prediction, including both road segments and intersections located in an urban road network and it was done using the spatial correlations among heterogeneous types of entities. Vandenbulcke et al. (2013) studied a spatial case-control to predict cycling accident risk for an entire network and identified how road infrastructure influences cycling safety in Brussels. They proposed a spatial Bayesian modelling approach that uses a binary dependent variable (accident vs. no accident at location) constructed from a case-control strategy. Xu et al. (2017) studied crash spatial heterogeneity to investigate the spatially varying relationships between crash frequency

and related risk factors. Their results suggested that ignoring spatially structured heterogeneity may result in biased estimates and incorrect inferences, whereas assuming the regression coefficients to be spatially clustered only probably results in over-smoothness.

Table 2-2: Spatial statistical methodologies to analyse accidents using GIS

Methodology	Author(s) and year
Moran's index (MI)	Blazquez and Celis (2016); Erdogan (2009); Manepali and Bham (2013); Moons et al. (2009); Tortum and Atalay (2015); Truong and Somenahalli (2011)
Kernel density	Banos and Huguenin-Richard (2000); Blazquez and Celis (2013); Çela et al. (2013); Erdogan et al. (2008); Hashimoto et al. (2016); Pulugurtha et al. (2007); Plug et al. (2011); Steenberghen et al. (2004); Yu, H. et al. (2014)
Getis-Ord	Erdogan (2009); Khan et al. (2008); Manepali and Bham (2013); Rankavat and Tiwari (2013); Tortum and Atalay (2015); Truong and Somenahalli (2011)
Network kernel density estimation	Xie and Yan (2008); Çela et al. (2013); Mohaymany et al. (2013); Xie and Yan (2013); Yalcin and Duzgun (2015)
Full Bayes hierarchical model with spatial effects	Aguero-Valverde and Jovanis (2006); Li et al. (2007); Aguero and Jovanis (2008); Mitra (2009); Guadamuz-Flores and Aguero-Valverde (2017)
Spatial random-parameters Tobit model	Anastasopoulos et al. (2012); Caliendo et al. (2015); Ma et al. (2015); Yu et al. (2015); Zeng et al. (2017b)
Kriging	Gundogdu (2010); Thakali et al. (2009); Mohamadzadeh Moghaddam and Abolfazl (2016)
Hot zone	Flahaut (2004); Loo (2009); Loo and Yao (2013); Moons et al. (2009)

The main methodologies used for the spatial analysis of road accidents are described in the following section.

2.3.1. Moran's index (MI) statistic

The MI is a statistical tool that measures the spatial dependence of the accident scene (Moran, 1948). The index proposed by Moran shows whether the spatial distribution of currently examined data values exhibits any regularity: that is, whether the data of neighbouring unit areas are similar to each other. The MI is a measure of spatial autocorrelation and looks at the spatial distribution of the data. More precisely, it looks at how dispersed, random or clustered the distribution is. It uses Z-scores and P-values to measure the level of statistical significance, which enables us to accept or reject the null hypothesis. These scores show whether the observed spatial clustering of high or low values is more pronounced than would be expected in a random distribution. For example, a high Z-score and a small P-value indicate a significant hot spot, whereas a low negative Z-score and a small P-value indicate a significant cold spot (figure 2.2).

Figure 2-2: Graphical output from the spatial autocorrelation tool (Global Moran's I) (Source: ESRI)

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} Z_i Z_j}{s_0 \sum_{i=1}^n Z_i^2}, \quad (2.36)$$

where

z_i is the distance of x_i from the average of x , and

z_j is the distance of x_j from the average of x .

MI evaluates the spatial pattern of a cluster as dispersed or randomized and determines the level of concentration. In figure 2.3, the red colour indicates an accident occurred. Positive spatial autocorrelation occurs when the MI is close to +1, indicating the mean values are clustered together. Conversely, negative spatial autocorrelation occurs when

MI is near -1, because dissimilar values are next to each other. A value of 0 typically indicates no autocorrelation.

Figure 2-3: Principle of the Moran index

Erdogan (2009) mentioned that the MI provides a single value for the spatial correlation and examines the group of spatial patterns. Hot spot location closeness between two points is often defined as the inverse of the distance between them. The attribute of the similarity severity index between two points is defined as the difference between each value and the global average value.

The MI is one of the most common statistics used to measure spatial autocorrelation by translating non-spatial correlation for spatial contexts, and usually applied to the data area in the numerical ratio or the value of the interval (O'Sullivan and Unwin, 2010).

The MI is usually applied to the data area, but it can also be used for other geometry data types such as point (Truong and Somenahalli, 2011) and linear features (Moons et al., 2009). Truong and Somenahalli (2011) studied traffic accident hot spots with the severity of each accident measured first and then used as an attribute value for computing the value of the MI.

Spatial autocorrelation (MI) works for the location of a feature or for an attribute value, but it also can be used for both feature locations and feature values simultaneously. Given a set of features and associated attributes, it evaluates whether the pattern expressed is clustered, dispersed or random.

Prasankumar et al. (2011) investigated and compared different kinds of traffic accidents in terms of spatial and temporal aspects. The study was the first of its kind in the Thiruvananthapuram city corporation. The results of the spatial statistics and cluster analysis elicited the spatial and temporal variations in accident highs (hot spots) and lows (cold spots) in the area. The assessment of spatial characteristics of the accident data by the MI method – the derived Z-score, MI and critical value – indicates that

accident datasets, as a whole, are categorized as clustered in nature, and accidents during the non-monsoon season and those associated with educational institutions.

Soltani and Askari (2017) studied the spatial autocorrelation of traffic crashes based on severity. The purpose was to find potential aggregated temporal and spatial patterns of road crashes at the traffic analysis zonal level in urban environments using spatial autocorrelation methods. They found that the hot spots generally appeared on arterial roadways, which tend to have higher car speed and volume and more travel lanes (multi-lane roadways). The analysis showed that the majority of crashes occurred in a number of specific zones, including major arterial roads and specific urban activity centres.

Mitra (2009) used the MI to identify intersections with statistically significant clusters of crashes involving fatalities and injuries. It was found that even though fatal and injurious crashes occurred at many intersections, only certain intersections were dangerous.

Moons et al. (2009) combined the accidents on a 100-meter segments and used the count of each segment as an attribute value to calculate the value of the MI. Additionally, Pirdavani et al. (2014) developed an accident prediction model using geographically weighted regression. This was done by calculating the MI for the selected dependent and explanatory variables. The results revealed the need to consider the spatial correlation when developing accident prediction models. Furthermore, Manepali and Bham (2013) identified those crash-contributing factors that have similar effects on spatial autocorrelation. The MI was used to determine the correlation, while multinomial regression was used to identify the factors.

Xie et al. (2014a) proposed an advanced crash frequency model that can account for the hierarchy and spatial correlation of crash data from signalized intersections in a high-density urban road network. The proposed model can serve as a useful complement to previous safety analysis methods in terms of the clustering characteristics of crash data in high-density road networks. The MI statistic was computed for crash observations, and the results confirmed the clustering characteristics of crashes at neighbouring intersections. This finding indicates the necessity of accounting for spatial effects when modelling crash observations.

Xie et al. (2014b) used the MI to develop an incident duration model which can account for the spatial dependence of duration observations and to investigate the impacts of a hurricane on incident duration. The MI statistics confirmed that durations of neighbouring incidents were spatially correlated. The modelling results show that the spatial lag model incorporating a spatially lagged dependent variable. The significance of spatial autoregressive parameters provides additional evidence for the spatial correlation of incident durations.

The spatial units Soro et al. (2016) measured were the 31 administrative regions of mainland China over the period 2004–2013. The presence of neighbourhood effects was

evidenced through the MI statistic. Models incorporating neighbourhood effects outperform those that did not, and the model incorporating both types of effects was found to best fit the data. The coefficient of the spatial lag was positive and statistically significant, suggesting a similarity of crash rates between neighbouring regions. The coefficient of the error lag was also positive and statistically significant, implying the existence of common unobserved factors affecting crash rates in neighbouring regions.

Famili et al. (2018) identified pedestrian crash hot spot clusters using ordinary kriging. They developed several criteria to identify such hot spots. These criteria will help decision-makers to appropriately allocate resources, including government safety funds. Their findings showed that ordinary kriging performed linear predictions for each segment with a weighted average of neighbourhood data points.

2.3.2. Getis-Ord

Getis and Ord (1992) showed the extent to which a location is surrounded by a group of high and low values. They identified statistical areas in which the average values found in close proximity to each other were higher than average or were lower than average. Getis-Ord statistics is a tool for identifying statistically significant hot spots, represented by cluster of high values, and cold spots, represented by clusters of low values. It also uses Z-scores and P-values to measure the statistical significance of the data and inform on whether to accept or reject the null hypothesis. The higher (positive) the Z-score is, the more intense the clustering of high values is, whereas a very low Z-score (negative) indicates intense clustering of low values as shown in equation 2.37.

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{\sqrt{\frac{n \sum_{j=1}^n w_{i,j}^2 - (\sum_{j=1}^n w_{i,j})^2}{n-1}}}, \quad (2.37)$$

where

x_j is the value of the j th element, w_i ,

j is the spatial neighbouring matrix element between i and j , and

n is the number of elements.

Figure 2.4(a) represents a given distribution of road accidents (identified by point features) on a road network. Running Getis-Ord statistics on such a dataset identifies statistically significant hot spots (clusters of high values) and cold spots (clusters of low values) through the calculation of Z-scores and P-values that measure the statistical significance. The G standard is counted from the single Z-score for each location in the

study area. Figure 2.4(b) shows the results of running Getis-Ord with the identification of hot spots (high Z-score and low P-value, represented by red point features) and cold spots (low negative Z-score and low P-value, represented by yellow point features). Getis-Ord (GI) statistics is a family of statistics that have a number of attributes that make them attractive for measuring the dependence of spatially distributed variables, especially when used in conjunction with the MI. They deepen the knowledge of the processes that cause spatial dependency and improve the detection of local dependency 'pockets' that may not appear when using global statistics (Getis and Ord, 1992). For example, Erdogan et al. (2015) used this method to detect local pockets of dependence. A high GI value indicates a spatial cluster of high values, whereas a low GI value indicates a spatial cluster of low values around the given area.

Gundogdu (2010) developed methods to obtain maps to determine traffic hot spots in Konya, Turkey, by applying GI analysis supported by a GIS. The main intention of this study was to emphasize the importance of using criteria other than total accident number to illustrate intersection hot spots and to constitute a model of accident severity and variety. It is anticipated that the results obtained from this highway accident data will guide improvement of the route segments. The authors found that hot spots determined by mortality criteria were seen on straight routes because of the high speed on those sections. In accidents on such routes, someone is typically killed. Although the number of accidents with injuries is lower than the number with mortalities, the factors are more or less the same.

Aghajani et al. (2017) proposed GI-based spatial statistical methods to identify and model accident hot spots. They investigated the use of localization patterns and hot spot distribution with the help of temporal information. In this analysis, a GI was found for every accident by considering the weights from combining the numbers of accidents, fatalities and injuries. Hot spot analysis (by GI) statistically identifies significant spatial clusters of high or low values (hot or cold spots).

Figure 2-4: Before (a) and after (b) running a Getis-Ord tool

Khan et al. (2008) and Kingham et al. (2011) highlighted the use of statistics when analysing collisions in relation to the school bus route and the weather. Khan et al. (2008) used GI to explore the spatial pattern of weather-related crashes. The study revealed that the crash patterns have a significant correlation with weather conditions, particularly crashes related to rain, fog and snow conditions.

Wang et al. (2018) studied spatial distribution patterns of customer counts and satisfaction with commercial facilities based on social network review data, a new method for quantitative analysis using the network-constrained local Getis-Ord's general G^* as an indicator. The results provided a quantitative reference to support decisions about the locations of commercial facilities and to help in identifying problems in commercial facility services to improve the quality of life among urban residents.

Blazquez et al. (2018) used the Getis-Ord GI^* statistic to test for statistically significant crashes for each year of the studied period from the perspective of local spatial association. They found that the clustering of crash attributes with high values (i.e. hot spots) occurred on highway locations with vertical curves and on cloudy days predominantly in the northern R5 (route number), and that the largest number of recurrent hot spots occurred on sunny days along the southern R5.

2.3.3. Kernel density estimation (KDE)

KDE is one of the important spatial analysis tools in commercially available GIS software. The estimation of kernel density involves placing a symmetrical surface over each point, evaluating the distance from the point to a reference location based on a mathematical function, and then summing the value for all the surfaces for that reference location. This procedure is repeated for successive points. This allows us to place a kernel over each observation, and summing these individual kernels gives us the density estimation for the distribution of accident points shown in equation 2.38.

$$f(x, y) = \frac{1}{nh^2} \sum_{i=1}^n k\left(\frac{d_i}{h}\right), \quad (2.38)$$

where

$f(x, y)$ is the density estimate at the location (x, y) ,

n is the number of observations,

h is the bandwidth or kernel size,

K is the kernel function, and

d_i is the distance between the location (x, y) and the location of the observation.

The effect of placing kernels over the points is to create a smooth and continuous surface. When applying this method, there are many decisions that need to be made

regarding the kernel shape, bandwidth and cell size. Using the kernel density function allows a surface to be created which will visualize the locations of those areas (or cells) with high or low numbers of collisions. Spatial density measurement is more accurate than using cross data across space.

This method is known as KDE because around each point (at which indicator is observed) a circular area (the kernel) of defined bandwidth is created. This takes the value of the indicator at the point observed according to some appropriate function. Summing all these values at all places, including those at which the indicator variable is recorded, gives a surface of density estimates.

Figure 2.5 illustrates the method of KDE. Road accidents are identified by point features and roads by line features. KDE is used to calculate the density of accidents within a specific radius (bandwidth). It uses a number from a kernel function to fit a smooth tapered surface to each accident location. Banos and Huguenin (2000) reported the case of using KDE for traffic accident data to chart the distribution of child pedestrian accidents.

Figure 2-5: Principle of kernel density

KDE measures the density of cells in a raster using a known sample point and estimates every point known as the kernel function. Similarly, some studies have spatially identified a group of accidents via KDE (Anderson, 2009; Pulugurtha et al., 2007; Schneider et al., 2004). Other studies have used KDE for spatially and temporally analysing traffic accidents (Blazquez and Celis, 2013; Plug et al., 2011). KDE has also been compared to other methods (Erdogan et al., 2008; Yu et al., 2014).

Loo et al. (2011) analysed the spatial distribution of road crashes with a GIS-based network-constrained KDE method. They found that the bandwidth has a great impact on the spatial distribution of density estimates. Using a narrow bandwidth that reveals more local effects identifies hot spots within urban and suburban areas. Using this method, authorities can gain more information on hazardous road locations, conduct investigations and propose improvement measures such as installing pedestrian refuges or speed cameras.

Ramp et al. (2005) used KDE to evaluate the distribution and patterning of wildlife fatality hot spots along the Snowy Mountain Highway in southern New South Wales. They suggested that KDE is an appropriate approach to identifying hot spots when presence data exists for all target roads. Where the intention is to extrapolate to additional or larger areas with similar characteristics, are necessary to develop predictive models.

Anderson (2008) examined road accident hot spots using a GIS and KDE to find spatial patterns of injury-related road accidents and to identify high-density accident hot spots. They concluded that the KDE tool enabled an overarching visualisation and manipulation of the accidents based on density, which was used in turn to create the basic spatial unit for the hot spot clustering method.

Hashimoto et al. (2009) investigated the relations between traffic accidents and city characteristics, such as population, road factors and spatial factors. They developed a model to estimate traffic accident density using KDE techniques and found a significant effect for public facilities and healthcare facilities in many models. It can be inferred that these facilities attract many pedestrians and vehicles and that traffic accidents occur frequently around these facilities. The results of this study objectively indicate areas in which area-wide traffic calming should be preferentially implemented. Additionally, the academic contribution of this study is that KDE models achieve Improvement compared to KDE models used in rural areas in applicability among cities.

Randkavat and Tiwari (2013) proposed a KDE method to investigate the potential of utilizing GIS to identify pedestrian accident-prone locations. The kernel method divides the entire study area into a predetermined number of cells and applies a circular neighbourhood around each crash. An individual value is assigned to each cell, and this is the value of the kernel function at the cell centre. The final density of each cell is calculated by adding its individual cell values. The radius of the circular neighbourhood affects the resulting density map. The greater the radius, the flatter the resulting kernel. The researchers found that the percentage of all pedestrian accidents was almost the same on all days of the week.

Shafabakhsh et al. (2017) applied a KDE method to four clustering analyses to better understand traffic accidents patterns in a complex urban network. They stated that the most important criterion for estimating the most suitable density level is bandwidth. Selecting bandwidth affects the output of hot spots. For instance, larger bandwidth shows hot spot areas in larger form. In this analysis, five categories were identified based on the standard deviation, which considers prioritize investigations to solve hot spot safety problems.

There are many advantages of KDE analysis as opposed to hot spot and clustering statistical techniques such as K-means. The main advantage of this method lies in its

ability to determine the spread of the risk of accidents. The risk of accidents can be defined as the area around the defined cluster in which there is an increased likelihood for an accident to occur based on the spatial dependency. Using these densities, arbitrary spatial units of analysis can be defined and homogeneous throughout the entire area to allow comparisons and a possible taxonomy.

Ouni and Belloumi (2018) used KDE to analyse the spatio-temporal pattern of collision hot spots for vulnerable road users (VRUs) and related risk factors for injury severity. They used VRU injuries and fatalities to describe the spatial pattern of collisions according to different temporal scales and investigated the influence of personal and environmental factors. In the results, spatial clustering patterns for each subtype of VRU collision according to a temporal scale were clearly observed under the distance of 38–40 km, with the exception of daytime VRU collisions, which showed a random tendency. All time-based subtypes of VRU collisions also were found to be clustered along the national highways and regional highways, especially in the regions.

2.3.4. Network Kernel Density Estimation

Use of network kernel density estimation (NKDE) as a method for adapting KDE as a network function has increased recently. The fundamental idea of NKDE is that instead of a planar area of influence, each point's influence extends a chosen distance along the network as shown in equation 2.39.

$$\lambda(s) = \sum_{i=1}^n \frac{1}{r} k\left(\frac{d_{is}}{r}\right), \quad (2.39)$$

where

$\lambda(s)$ is the density in the location s ,

r is the search radius (bandwidth) of KDE (only points with radius r are used to estimate $\lambda(s)$), and

k is the weight of point i , located at a distance of location s .

Kaygisiz et al. (2015) used NKDE to utilize an accident hot spot to analyse the effects of various measures taken by authorities to influence the behaviour of drivers. Any changes in the road and the environment influence the driver's choices and therefore the safety of the road itself. The NKDE method to identify hot spots adopted a K-function, and the nearest-neighbour distance method was used to determine the significance of hot spots.

Harirforoush and Bellalite (2016) proposed two integrated methods to identify traffic hot spots in a road network. The first step includes a spatial analysis method called

estimation of kernel density network. The second step is a screening method using a network of critical accident rates and is explained in the *Highway Safety Manual (HSM)*. This method was checked using three years of traffic accident data (2011–2013) from Sherbrooke, Canada. In this case, Kernel density network (KDN) was used with traffic accident data to graphically display sites with a high accident density. The results of the study confirm that the proposed approach can help the traffic authority rapidly identify the most dangerous locations in the road network.

Figure 2.6 shows the basic ideas of NKDE. Unlike a regular kernel, which uses Euclidean distance over a planar area to estimate density values, NKDE constrains the kernel and the bandwidth in a network, the bandwidth describing the shortest road distance along the path. In figure 2.6, accidents are represented by white circles and roads by lines. The kernel of each accident point defined by a specific distance (bandwidth) represents the shortest distance from the points along the road (green line).

Figure 2-6: Principle of network kernel density estimation (NKDE)

In implementing NKDE, Xie and Yan (2008) suggested using a linear segment (called a *lixel*) of roads as the basic unit for aggregating accidents and calculating density and for visualization. The authors found that segments of shorter length are more capable of showing the local variations among segments. The linear segments with geocoded accidents are called source segments. The NKDE density values are computed for each segment and its neighbours. For a segment falling within the search bandwidths of multiple source segments (including a segment that itself functions as a source segment), its density is the cumulative value of the densities computed from all relevant source segments. The result of NKDE is a linear spatial dataset with density values computed and assigned to each linear segment, or *lixel*.

In 2009, Okabe et al. (2009) from the University of Tokyo developed a GIS toolbox called Spatial Analysis on a Network (SANET). The aim of this toolbox is to allow GIS practitioners in the field to easily undertake network spatial analysis. More specifically,

the 'kernel density estimation' tool available in SANET is used to implement the NKDE method in the calculation of traffic accidents.

SANET consists of three components (figure 2.7):

1. the network-computation program,
2. an interface, and
3. a GIS viewer.

The network-computation program deals with computation used in SANET. This program is self-contained but can be connected to other outside programs through import and export files.

Figure 2-7: The different components of the SANET toolbox (Source: Okabe et al. 2009)

SANET uses a different formula, which is shown in equation 2.40.

$$K_y(x) = \left\{ \begin{array}{ll} k(x), & \text{for } -h \leq x \leq 2d - h \\ k(x) - \frac{n-2}{n} k(2d - x), & \text{for } 2d - h \leq x \leq d \\ \frac{2}{n} k(x), & \text{for } d \leq x \leq h \end{array} \right\}, \quad (2.40)$$

where

$k(x)$ = the basic kernel function,

y = the centre of the kernel function,

x = a point located on the network,

h = the bandwidth (meters),

n = the node degree, and

d = the shortest path from y to x (meters).

The density of the point of interest is calculated according to equation 2.41.

$$D(O) = \int_{-h}^{2d-h} k(-y)d + \int_{2d}^d \left[k(-y) - \left(\frac{2-n}{n} \right) k(-2d+y) \right] dy + \int_d^h (n-1) \left(\frac{2}{n} \right) k(-y) dy \quad (2.41)$$

From a broad perspective, NKDE represents an emerging effort to extend the applications of standard spatial statistical methods to analysing spatial point events in a network space. Xie and Yan (2013) developed NKDE to estimate the density of traffic accidents strictly over a network space. They stated that NKDE has proven to be useful in accident analysis over a network space and helps mitigate potential instability in observed accident counts.

Eckley and Curtin (2013) proposed using SANET to test the significance of spatio temporal clusters in the case of a network consisting of many line segments. The test case consisted of 586 traffic collision tests that were scattered on the main road network in Franklin County. The minimum road distance between two neighbours was 0 meters, while the nearest neighbour distance was 7,470 meters, and the average network distance to the nearest neighbour was 344 meters.

Young and Park (2013) said the NKDE technique shows how to take into account how the level of the collision damage affects the identification of hot zones. The researchers compare the hot zone results obtained using a standardized *HSM* approach with the results obtained from the proposed two-step screening approach to spatial data analysis using a KDE restricted network technique. The findings indicate that a proposed two-step screening network method using a restricted NKDE can successfully reduce the effects of regression to mean (RTM) and identify hot zones where appropriate safety improvements should be identified and investigated. Since the two-step screening method takes into account the density of Equivalent property damage only (EPDO) collisions on the adjacent highway, and combines several individual segments into spatially related hot zones.

Ivan and Tesla (2015) used SANET as a spatial analysis tool applied to local data on traffic accidents. The tool enabled fast location of certain segments and intersections with above average numbers of traffic accidents. The researchers evaluated the intensity of traffic accidents on a road network which did not depend on the distribution of segments of varying lengths.

To determine traffic accidents at intersections in the city of Cluj-Napoca, Benedek et al. (2016) used an NKDE method that was implemented by means of kernel density estimates in the SANET toolbox. NKDE was used to calculate the incidence density of type points along the network. In this case, the events analysed were traffic accidents in Cluj-Napoca, while the linear units were represented by the same city road network. The collision location given by the GPS coordinates was recorded by equipment installed in a traffic police car. A total of 1,466 accidents occurred along approximately 662 km of road over four years (2010–2013).

Khlaïd et al. (2018) used SANET to analyse a network-constrained spatio-temporal hot spot mapping of crimes. SANET tools were used to compute NKDE and used its results as input values to calculate Getis-Ord GI* statistics. The results demonstrated that street crimes were highly concentrated in the central part of the city.

2.3.5. Kriging

The main idea behind kriging is that the predicted outputs are weighted averages of sample data, and the weights are determined in such a way that they are unique to each predicted point and a function of the separation distance (lag) between the observed location and the location to be predicted. In other words, kriging provides an estimation for unknown locations based on a set of available observations by characterizing and quantifying spatial variability of the area of interest. The expression of a general kriging model follows equation 2.42.

$$\hat{Z}(X) = m(X) + \sum_{i=1}^n \lambda_i [Z(X_i) - m(X_i)] , \quad (2.42)$$

where $m(x)$ and $m(x_i)$ are expected values of the random variables $Z(x)$ and $Z(x_i)$, and λ_i is a kriging weight assigned to datum $Z(x_i)$ for estimation of a crash frequency at any location x . The random field, $Z(x)$, can be decomposed into two components – namely, residual component $R(x)$ and a trend component $m(x)$ – and can be expressed as $Z(x) = R(x) + m(x)$.

This is a method for interpolating the variable, which are weighted averages of sample data. Based on available data, this tool estimates values at unknown locations by building a methodological basis for spatial inference of quantities in unobserved locations. Gundogdu (2014) applied kriging to determine new safe routes and zones for traffic accidents in Konya, Turkey. The study results differ from results of other studies in the literature that used GA applications for traffic accidents. In addition, the study

differs from others in that it uses not only the total number of accidents but also additional traffic accident criteria.

Selby and Kockelman (2013) used kriging to predict traffic levels in unmeasured locations. The study explores the application of two distinctive spatial methods for prediction of average daily traffic counts across the Texas road network. Results based on Euclidean distances are compared to those based on network distances, and allow for strategic spatial interpolation of count values while controlling for each roadway's functional classification, lane count, speed limit and other site attributes. Kriging methods reduce count prediction errors by controlling for local attributes and recognizing distance-based correlation structures that exploit information found in nearby residuals.

Thakali et al. (2015) compared the outcomes of two geostatistical-based approaches, including kriging, for identifying crash hot spots in a road network. The authors argued that kriging has advantages in handling spatially autocorrelated datasets and is superior in its ability to pinpoint hot spots. Furthermore, the list of hot spots identified by kriging was found to be moderately different from those found with the other method, indicating the importance of selecting the right geostatistical method for hot spot identification.

Mohamadzadeh Moghaddam (2016) identified clusters of injured pedestrians using ordinary kriging analysis. The ordinary kriging was used to make linear predictions for each segment with weighted averages of neighbourhood data points. The results were categorized into ten levels based on standard deviations arranged in order of increasing risk, with each level representing 10% of the total buffered area.

Famili et al. (2018) identified clusters of pedestrian crash hot spots using ordinary kriging. They developed criteria to identify these hot spots to help decision-makers to allocate resources appropriately, including government safety funds. The authors' findings showed that ordinary kriging performed linear prediction for each segment with weighted averages of neighbourhood data points.

2.3.5.1. Universal kriging

Universal kriging, $\mu(s)$, can be a deterministic function of any form. A simple assumption is to use a linear function where $\mu(s) = X\beta$ (where X contains explanatory variables, such as number of lanes and facility type). In contrast, $\varepsilon(s)$ reflects unobserved variation (e.g. local land use patterns and transit service levels).

For purposes of prediction, kriging is performed on the $Z(s)$ values. The sums of interpolated random components $\varepsilon(s)$ and estimated $\mu(s)$ values provide the $Z(s)$ value estimates. The following section explains how the random component $\varepsilon(s)$ is estimated via kriging.

As noted above, weak stationarity ensures the following equation 2.42.

$$\gamma(h) = \frac{1}{2} \text{var} [Z(s+h) - Z(s)] , \quad (2.42)$$

where

$\text{var} [Z(s+h) - Z(s)]$ is the variance (over all sites) between counts at sites s and $s+h$.

The first step is to select an appropriate ‘curve’ or semivariogram model to best fit this relationship (γ vs. h) for a given dataset. There are several commonly used models, including exponential, spherical and Gaussian models. These specifications are as in equations 2.43, 2.44 and 2.45.

- Exponential

$$\gamma(h) = \left\{ c_0 + c_1 \left[1 - \exp\left(-\frac{h}{a}\right) \right] \text{ if } h > 0 \right\} \quad (2.43)$$

- Spherical

$$\gamma(h) = \begin{cases} c_0 + c_1 \left[1.5 \frac{h}{a} - 0.5 \left(\frac{h}{a} \right)^3 \right] & \text{if } 0 < h < a \\ c_0 + c_1 & \\ 0 & \text{if } h > a \end{cases} \quad (2.44)$$

- Gaussian

$$\gamma(h) = \begin{cases} c_0 + c_1 \left[1 - \exp\left(-\frac{h}{a}\right)^2 \right] & \text{if } h > 0 \\ 0 & \end{cases} \quad (2.45)$$

These models all rely on three parameters that describe their shape while quantifying the level of spatial autocorrelation in the data. c_0 is called the ‘nugget effect’, and it reflects discontinuity at the variogram’s origin, caused by factors such as sampling error and short scale variability. (In theory, $\gamma(h=0)$ should be zero.) Here, a is called the ‘range’, and this scale factor determines the threshold distance at which $\gamma(h)$ stabilizes (i.e. flattens). $c_0 + c_1$ is the maximum $\gamma(h)$ value, called the ‘sill’, with c_1 referred to as the ‘partial sill’ (Cressie, 1993).

2.3.5.2. Ordinary kriging

In practice, in most cases the μ mean and covariance function $C(h)$ of the underlying random function $Z(x)$ is unknown. Thus, unfortunately, simple kriging prediction does not apply because it requires information about μ and $C(h)$ (see Cressie 1993 p. 359). For this reason, some predictions by ordinary kriging geostatistical methods are introduced, which unlike simple kriging do not assume knowledge of the mean and its covariant. For this reason, ordinary kriging is the most common kriging method in

practice, and the goal is to predict the value of the random variable $Z(x)$ at the point x_0 in an unsampled geographic region.

Cressie (1993) and Wackernagel (2003) stated that ordinary kriging refers to spatial prediction under the following two assumptions: the global, constant mean $\mu \in \mathbb{R}$ of the random function $Z(x)$ is unknown, and the data come from an intrinsically stationary random function $Z(x)$ with known variogram function $\gamma(h)$, as shown in equations 2.46 and 2.47.

$$\gamma(h) = \frac{1}{2} \text{var}(Z(x+h) - Z(x)) = \frac{1}{2} E[(Z(x+h) - Z(x))^2] \quad (2.46)$$

The ordinary kriging predictor $Z(x_0)$ of the value at x_0 is the linear combination of $Z(x)$ evaluated at each sample $x_i; i = 1; \dots; n$

$$Z_{\omega}^*(x_0) := \sum_{i=1}^n \omega_i Z(x_i) = \omega^T Z, \quad (2.47)$$

where $\omega = (\omega_1, \dots, \omega_n)^T \in \mathbb{R}^n$ provides the unknown weights $\omega_i \in \mathbb{R}$ corresponding with the influence of the variable $Z(x_i)$ in the computation of $Z(x_0)$.

2.3.6. Full Bayes hierarchical model with spatial effects

Bayesian spatial modelling applies a Bayesian method to spatial models: for instance, spatial autoregressive and conditional models. It considers the distributions of data and the unknown coefficient estimators. This approach offers a solid basis due to unknown and/or uncertain parameters.

Bayesian spatial statistical modelling has been widely used in recent years. Mitra (2009), for example, developed a model to investigate the influence of spatial factors on intersection crashes with injuries. Li et al. (2007) studied spatial-temporal patterns of relative crash risks based on a Bayesian approach used to identify and rank those roadway segments with a potentially high risk for crashes.

Mc Nab (2004) studied Bayesian spatial modelling to show how Bayesian techniques can be applied to assess potential risk factors measured at the group level (e.g. area). An integrated modelling framework is presented that allows a thorough investigation of the relationship of the degree of injury to regional characteristics, the residual variation, and spatial autocorrelation. Studies showed large regional variations in motor vehicle accidents in men up to 24 years old in British Columbia, Canada, in 1990–1999, and that adjusting the right risk factors would eliminate almost all of the observed variation. The socioeconomic influence in motor vehicle accidents is visible in young men of all ages, and injury is more common in people with lower socioeconomic status.

Several advantages are offered by full Bayes hierarchical models in certain cases of crash data analysis. They are also a more flexible approach. One important characteristic of the models is their explicit use of prior information, which has a central part in the scientific method and has a natural implementation in Bayesian analysis. The models also take full account of the uncertainty related to estimation parameters. They provide exact measures of uncertainty on the posterior distributions of these parameters.

Sacchi et al. (2016) used a full Bayes method to compare the effects of spatial and temporal dynamics in evaluating the effectiveness of road safety treatments. The analysed countermeasure was the installation of traffic signals at unsignalized urban or suburban intersections in British Columbia, Canada. A full Bayes approach was chosen as the statistical framework to develop the model. The results show that the variation of the zone is the main component of total accident variability, and spatial effects are reduced by clustering intersection together.

Aguero and Jovanis (2008) applied full Bayes hierarchical models to assess the spatial correlation of crash frequency at the segment level. They proposed different conditional correlation structures by varying the weights in the neighbouring structure of the CAR model. They also suggested the use of the inverse of the order (as defined by proximity) as weight in the CAR model. The simplest neighbouring structure was the best in terms of the goodness-of-fit measures.

Yanmaz-Tuzel and Ozbay (2010) developed a step-by-step methodology to implementing a full Bayes model for analysis before and after road safety countermeasures. The estimation results indicate that for urban roads considered in this study, the greatest road safety improvements were observed after repairs to the vertical and horizontal alignment of the road, followed by an increase in lane width and the installation of median barriers.

Huang et al. (2010) proposed a spatial Bayesian model to consider variations in accident risk in a district in Florida to explicitly control the exposure variable and the daily mileage of vehicle trips. CAR had been previously determined to accommodate spatial autocorrelation in the adjacent districts. The results showed no significant difference in the effect of risk factors on safety in all crashes compared to the effect in severe crashes. Districts with higher traffic intensity and higher population density and urbanization were associated with a higher risk of accidents. Unlike arterials, freeways seemed safer with regard to the risks of accidents given either vehicle trips. An increase in truck traffic volume tended to result in more severe crashes.

Aguero-Valverde and Jovanis (2011) investigated spatial correlation in multi-level collision frequency models for different types of urban and rural road segments. The study applied a full Bayesian hierarchical approach with CAR distribution for the spatial correlation terms. The findings revealed that spatial correlation substantially increases

random effects. They indicated that 70–90% of the variation explained by random effects resulted from spatial correlation. This suggests that spatial models offer a significant advantage, because poor estimates that result from small sample sizes and low sample means are frequent issues in highway collision analysis.

Barua et al. (2014) investigated the inclusion of spatial correlation in the data model that calculates the multivariate severity of the collision. The model was developed for severe crashes (injuries and fatalities) and no injury crashes using three years of collision data from the city of Richmond and downtown Vancouver. The proposed model was estimated within a full Bayesian context through the simulation of a Markov-chain Monte Carlo (MCMC) method. The results showed a significant positive correlation between various attributes of the road and severity. For the Richmond dataset, the proportion of variance for the spatial correlation was smaller than the proportion of variance for heterogeneous effects.

Zeng and Huang (2014) proposed a Bayesian spatial joint model of crash prediction, including both road segments and intersections in an urban road network, through which the spatial correlations between heterogeneous types of entities could be considered. The results showed that the joint and CAR models outperformed the Poisson and negative binomial models in model-fitting and predictive performance, which indicates that the consideration of spatial correlations between adjacent entities is reasonable.

Saheed et al. (2016) used a full Bayes approach to precisely determine the factors that affect the severity in winter, which shows the relationship between the accident and the severities of the injury. The results of the model developed showed a significant correlation of accidents in the collection of research data. These findings justify the use of multivariate speciation on multivariate error address associated with a statistical analysis of the exact things that might happen as a result of accidents.

Dong et al. (2016) proposed a spatio-temporal interaction approach to the identification of hot spots by applying full Bayesian techniques in the context of macroscopic security analysis. Their findings identified hazardous zones at the macro level, and these findings resulted in proactive safety planning. As urban areas are developed, it is expected authorities will formulate strategies to manage traffic safety that consider the characteristics of the zone level.

Xu et al. (2017) investigated the various spatial relationships between the frequency of accidents and risk factors. A Bayesian variable spatial coefficient model was introduced as an alternative methodology to simultaneously consider the spatial heterogeneity of structured and unstructured regression coefficients in predicting the frequency of accidents.

Moreover, the spatial correlations between intersections and their feeding segments were found to be more significant than those solely between adjacent intersections or adjacent segments, a finding which supports the employment of the joint model as an alternative in road network level safety modelling.

Guadamuz-Flores and Aguero-Valverde (2017) studied spatial correlation structures of crash frequency at railway crossings, particularly CAR and joint specification using Gaussian kriging models. They found that for more complex model definitions, including spatial random effects, Poisson lognormal models showed more stable performance.

Baka et al. (2018) used a Bayesian framework to investigate the relationship between bicycle crash frequency and contributing factors to account for the effects of spatial dependence across census block groups. The class of CAR models were employed within the hierarchical Bayesian framework. The study explored and compared the performance of two CAR models – namely, the Besag’s model and the Leroux’s model – in crash prediction. The Besag’s models, which differ from the Leroux’s models in the structure of how spatial autocorrelations are specified, were found to fit the data better. A 95% Bayesian credible interval was selected to identify the variables that had significant impact on bicycle crashes.

Fu et al. (2019) investigated spatial autocorrelation and spill effects in freeway crash frequency data. The results of Bayesian estimation conducted in WinBUGS software showed that significant spatial autocorrelation and spill-over effects simultaneously existed in the freeway crash-frequency data. The lower value of the DIC and the more significant exogenous variables for the hybrid model compared to the other alternatives indicate the strength of accounting for both spatial models. In addition, the authors mentioned that the vertical grade and horizontal curvature of adjacent freeway segments had significant effects on crash frequency of the targeted segments, manifesting a significant spatial spill-over effect in the freeway crash data.

2.3.6.1. The Integrated nested Laplace approximation

The integrated nested Laplace approximation (INLA) approach has been developed to approximate Bayesian inference of latent Gaussian models. It is seen as a faster alternative to the MCMC approach for fitting Bayesian models, providing results in minutes and seconds where the MCMC method would require hours and days to fit the same models.

An associated R package (R-INLA) developed to run Bayesian inference is efficient and effective. It supports hierarchically the Gaussian Markov random field models of the following type of equations 2.48 and 2.49.

$$y_j | n_j, \theta_1 \sim \pi(y_j | n_j, \theta_1), \quad j \in J \quad (2.48)$$

$$n_i = Offset_i + z_i^T \beta + \epsilon_i + \sum_{k=0}^{n_f-1} w_{ki} f_k(c_{ki}), \quad i = 0, \dots, n_n - 1, \quad (2.49)$$

where

J is a subset of $\{0, 1, \dots, n_n-1\}$,

$\pi(y_j|n_j, \theta_1)$ is the likelihood of the observed data assumed to be conditionally independent given the latent parameters n and possibly some additional parameters θ_1 ,

ϵ is a vector of unstructured random effects of length n_n ,

$n = (n_1, n_2, \dots)$ is a vector of predictors,

Offset is an a priori known component to be included in the linear predictor during fitting,

w_k are known weights defined for each observed data point,

$f_k(c_{ki})$ is the effect of a generic covariate k which assumes value c_{ki} for observation i , and

z_i is a vector of n_β covariates assumed to have a linear effect.

2.3.7. Safety performance function (SPF) modelling

The *HSM* published by the American Association of State Highway Transportation Officials provides a description of the methods available to estimate the expected number of crashes for various roadway facilities. The critical component of these methods is the SPF, which is essentially a regression model to correlate quantitatively the expected number of crashes with traffic exposure and the geometric characteristics of the road. The SPF has long been developed for various types of entities, such as rural roads, urban arterials, intersections, freeways and even freeway ramps.

Lord and Persaud (2004) studied the macroscopic representation of physical networks and the application of SPFs, or crash prediction models, which describe the relationship between the number of crashes and the most important independent variables, such as traffic flow or the number of lanes on a road segment. They created a series of predictive models (i.e. SPFs) for predicting the numbers of crashes at nodes and on links using data from Toronto, Ontario, for 1990–1995. The results demonstrated some important pitfalls and limitations in the application of SPFs to computerized transportation networks.

Wang et al. (2013) compared and defined effective road network indices and analysed the relationship between road network structure and traffic safety at the level of the traffic analysis zone. Data for traffic, crashes, roadways, demographics and land use features of Orange and Hillsborough Counties in Florida were examined to assess the safety of road networks. Three indices of structure (closeness centrality, betweenness centrality and meshedness coefficient) were used to determine the extent to which

researchers could classify various road network structure types on a continuum that would allow for the prediction of safety as a function of structural configurations.

Garber and Ehrhart (2000) examined models that used the standard deviation of the speed of vehicles on a roadway. They noted that more crashes occur when there is a higher speed distribution. They summarized that crash rates are higher in the early morning and late day hours because there is higher variation in vehicle speeds at these times, when fewer vehicles are on the road. Mensah and Hauer (1998) used actual causal factors to define the crash modelling problem. They discussed the limitations of modelling crash frequencies based on average values, such as AADT. They asserted that an instantaneous value for traffic flow is a better predictor of crashes than an average daily count is, because crashes are discrete events and are the results of specific causes.

2.3.8. Spatial random-parameters Tobit model

Previous studies have suggested that heterogeneous effects of certain factors may affect crash rate. For this reason, Anastasopoulos et al. (2012) applied a random-parameters Tobit model. They studied accident rates categorized by injury severity (based on the most severely injured vehicle occupant in an accident) and by estimating a multivariate Tobit model of accident injury-severity rates which accounts for both the underlying masking of the censoring sources and the contemporaneous error correlation of the latent variables due to commonly shared unobserved characteristics.

Various random-parameters Tobit model forms have been proposed to take into account unobserved heterogeneity (Anastasopoulos et al., 2012a; Caliendo et al., 2015; Ma et al., 2015a; Yu et al., 2015). Calculating the accident rate with a Tobit model avoids some of the shortcomings associated with Poisson-based accident frequency prediction models such as requirements of the mean and variance of the Poisson distribution and state of zero hypothesized on the model count of zero-inflated (Lord et al., 2005; Lord et al., 2007), which may be the technical advantages of modelling the rate of accidents modelling over accidents frequency modelling.

Zeng et al. (2017a) proposed a multivariate random-parameters Tobit model to analyse the relationship between the rate of accidents and injury severity. They used an accident dataset obtained from the traffic information system managed by the Ministry of Transport in Hong Kong. In short, their empirical analysis demonstrated the superiority of a multivariate random-parameters Tobit model and a significant correlation between the rate of accidents and their severity along with heterogeneous effects of multiple risk factors in crash severity data.

A second study by Zeng et al. (2017) developed a Bayesian spatial random-parameters Tobit model to analyse crash rates on road segments. The findings showed a significant

spatial effect in both spatial models. The results of the DIC suggested that including spatial correlation in the Tobit regression considerably improves model fit, which indicates the reasonableness of considering cross-segment spatial correlation.

2.3.9. Hot zone

Two of the more commonly used spatial methodologies include hot spot and hot zone analyses. The hot spot methodology is a rather straightforward and self-explanatory concept. A hot spot refers to a dangerous road location where the actual crash rate is higher than the critical crash rate (Loo, 2009; Flahaut et al., 2003; Ouni and Belloumi, 2019). Other researchers suggest different parameters to define a hot spot, including the expected number of crashes rather than the actual number. The expected number is determined using empirical Bayes and full Bayes models (El-Basyouny and Sayed, 2013; Huang et al., 2009; Miranda-Moreno and Fu, 2007). A hot spot is mainly a 'point' location such as a road intersection or a small but distinctive geographical feature such as a basic spatial unit (BSU). If the actual crash rate of a BSU is higher than the critical or expected crash rate, the BSU is considered a hot spot. Therefore, hot spots are highly separated and independent events and network contiguity does not need to be considered in identifying them. A first step in a hot spot analysis is road segmentation, which essentially divides the entire road network into smaller BSUs (Flahaut et al., 2003, Flahaut, 2004; Moons et al., 2009). The length of the BSUs can vary and is context-specific (Ouni and Belloumi, 2019), but 100-meter long segments are commonly used in the literature (Elvik, 2008; Flahaut et al., 2003; Loo, 2009; Loo and Yao, 2013). The road segments must be long enough to allow for the identification of hot spots and short enough to reflect changes in the environment that may influence road safety (Loo, 2009). Although the hot spot methodology makes it easier to compare road segments, the case of an actual crash rate exceeding the critical value can also be a result of randomness rather than an example of a 'systematic' focus of traffic crashes, leading to a so-called 'false-positive' scenario (Loo, 2009; Miranda-Moreno et al., 2007). This can be problematic when developing road safety strategies with limited resources in selected priority areas.

The *hot zone* methodology, by contrast, mainly identifies dangerous road locations that include two or more continuous road segments (i.e. BSUs) with actual crash rates higher than a threshold value for each spatial unit (Flahaut, 2004; Loo, 2009; Loo and Yao, 2013; Moons, Brijs and Wets, 2009). The spatial analysis of hot zones takes into account the contiguity of the entire road network, thereby considering the spatial dependence between BSUs. Various spatial statistics of local and global measures are used to illustrate the clustering patterns of road traffic crashes. Global measures usually incorporate quadrant methods and K-functions, whereas local measures include kernel density estimation (KDE) (also planar or network-constrained KDE), K-functions, local MI,

local Getis-Ord G^* and spatial autocorrelation. Although there is no consensus on the best approach to detecting hot zones, the use of local measures can generally generate better results, as global measures capture only the average spatial pattern in an entire region (Ouni and Belloumi, 2019). Some studies challenge the use of the absolute crash rate as the parameter for generating hot zones, as it is potentially susceptible to the RTM bias. Some other parameters have been suggested to reduce this bias, including the empirical Bayes and full Bayes methods (Yang and Loo, 2016). Again, this highlights that the occurrence of traffic rashes can be highly random and even extremely random' (i.e. a Poisson distribution). Overall, spatial autocorrelation strategies are particularly useful and offer explanatory power. Loo (2009) proposes a three-step hot zone generation approach which includes geo-validation of traffic crashes, network segmentation to BSUs and assignment of traffic crashes to those BSUs, and finally the identification of hot zones. When the actual crash rate of contiguous BSUs exceeds the threshold value, these BSUs are considered to be hot zones.

2.4. Conclusions

A large number of methodologies have been developed to analyse road traffic crashes. These methodologies can be classified into two broad categories: non-spatial and spatial methodologies.

Traditionally, non-spatial methodologies have been extensively used in the field of traffic safety. Methodologies deriving from the Poisson family have been amongst the most popular ones. Certain limitations of these methodologies have been highlighted, including the fact that they do not consider the spatial aspects of crashes. For example, while the Poisson lognormal approach has the potential to offer more flexibility than the negative binomial (Poisson-gamma), it cannot handle under-dispersion effectively. This means that it can be influenced by a low sample average and a small-sample size bias.

With the appearance of GIS, many spatial methodologies have emerged. GIS-based and spatial methodologies have enabled accident locations to be visualised on a map. They have also allowed researchers to identify concentrations of accidents and analyse the relationships between accidents and contributing factors. Methodologies to identify the locations of hot spots have been the most popular. Recently however, the concept of hot zones has emerged. A hot zone identifies hazardous locations that include at least two contiguous segments. The hot zone methodology can be seen as an expansion of the hot spot methodology and was developed in response to one of the drawbacks of hot spot methodology and to distinguish true-positive from false-positive sites.

The continuous improvements in computational abilities are now allowing spatial and statistical methodologies to be combined. These combined methodologies include

models that are more sophisticated, complex and resource-intensive but also more precise.

The KDE method is comparatively simple and primarily used to determine road accident hot spots. It can integrate various datasets, including data about the environment or the number of people involved in road accidents. Other methods are sometimes more suitable to analyse road accident hot spots. MI, for example, measures the spatial dependence between the crash site and accident cluster site based on the spatial pattern and provides a powerful tool to study the occurrence of accidents. GI statistics, on the other hand, measure the value of a single spatial correlation and provide a measure of the degree of clustering in spatial patterns. The NKDE methodology has largely been applied as a tool to visualize accident analyses missing quantitative statistical inference assessments. Areas of high and low densities can also be represented simultaneously. Kriging provides estimations at unknown locations based on a set of available observations by characterizing and quantifying the spatial variability of the area of interest. Bayesian spatial modelling considers the distribution of data and the unknown coefficients estimators.

For the scope of this thesis, three methodologies have been retained: NDKE, the hot zone methodology, and the Bayesian spatial methodology. NetKDE is used to identify concentrations of crashes (hot spots) on the two roads. NKDE is one of new KDE methods based on networks that have recently been developed to estimate the density distribution of point events over a one-dimensional network space. NKDE was created to address some of the drawbacks of using a planar kernel for network-based analyses. The hot zones methodology is to use to identify hazardous road locations that consists of two or more than one contiguous segment. The rationale behind the concept of hot zone is that some contiguous road segments may be more dangerous than a hot spot alone. Many researchers are highlighting the benefits of using the hot zone methodology to complement the hot spot methodology. To date, the hot zone methodology has mostly been applied in large-scale analyses (for example, network-level analyses). Finally, the Bayesian spatial model is used to identify factors that contribute most to road crashes. Bayesian models have been widely used by traffic agencies around the world to identify road crash locations and to analyse contributing factors in order to determine appropriate countermeasures to improve road safety. Among their benefits, Bayesian models can capture the variability of estimates on low population samples.

3. DATA

3.1. Introduction

As stated in the introduction, the main objective of this thesis is to gain a better understanding of the relationships between traffic accidents and road design on the primary roads of the province of Aceh. For this, crash and road information from the two main highways of the province for the period 2012 to 2015 were collected, quality checked, and formatted into a common schema. The two datasets (the crash dataset and the road dataset) are presented in the next subsections and are used in the subsequent analyses presented in the next chapters.

It is important to note that gaining access to data in Indonesia, as in many countries of the region, remains a challenge. Unlike Western countries, which often have an extended GIS unit within the relevant traffic agencies and ministry departments, there is no such unit in the relevant agencies of the province of Aceh. According to the traffic police department, mobility practices have changed dramatically over the past 15 years in Aceh, and particularly since the 2004 tsunami disaster that hit the province and destroyed a large part of the infrastructure. Yet, data collection still relies much on traffic crash databases and manual spreadsheets. These are useful for temporal and historical records but are limited for undertaking spatial analyses to understand how the different events or variables relate to each other. To achieve this end, Indonesian officials must start collecting and managing accurate and precise spatial information. However, maintaining such datasets is also very costly, thus the benefits of doing so must be better presented and understood.

Aceh is an independent province in Indonesia, located on the island of Sumatra in Western Indonesia. Indonesia is one of the largest archipelagic countries in the world, with more than 17,000 islands (figure 3.1). Indonesia is also one of the most populous countries in the world. Most of its population is located on the country's five main islands Sumatra, Java, Kalimantan, Sulawesi and Papua. In particular, the island of Java,

which represents 7% of the total land area of the country, is home to 60% of the population. Over the past two decades, the country has experienced a rapid growth in its urban population (+250%). Projections indicate that the urban population in the country is expected to reach 66% by 2035. The important annual population growth, the rural-urban migration and the reclassification of rural areas into urban are the three main factors of urbanisation.

Figure 3-1: Political map of Indonesia (Source: Worldometer)

The province of Aceh counts 17 districts (figure 3.2). In total, the province represents 3.2% of the terrestrial part of Indonesia and, in 2015, its population was accounted for about 2% of the total population. However, the distribution of the population in the province is uneven across districts (figure 3.3). Districts located on the east side of the island (and crossed by the east highway) record a higher population density than districts located on the west side of the island (and crossed by the west highway). In addition, over the past decade, the urbanisation rate has increased more quickly in districts located on the east side of the island than in districts located on the west side of the province. In Aceh, the population is distributed over large areas of the districts in the form of many small cities and villages, rather than concentrated in smaller areas of high densities and high buildings.

Figure 3-2: Locations of districts in Aceh province (Sumatra island)

Figure 3-3: Population density in the different districts of Aceh province

3.2. Accident data

The accident information was obtained from the department of traffic police of the province of Banda Aceh. The raw information was compiled in Bahasa in an Excel spread sheet. Each record contained information about an accident that occurred on either of the two main highways. For each accident, information such as latitude and longitude were captured to record the location. A set of descriptive information was also associated with each accident and provided some context. These descriptive attributes included, for example, information about the severity of the injuries, the type of accident, the day and time of the accident, and the road conditions. A full list of these associated variables, and their definitions, is provided in table 3.1. When the variable information was too incomplete, it was excluded from subsequent analyses.

Table 3-1: Definition and description of the variables associated with each road accident record

Name of the variable	Definition or description	Accepted values	Included in the analysis?
Weather condition	Weather condition at the time of the crash	Bright, foggy, cloudy, rainy	Yes
Light condition	Lighting condition at the time of the crash	Bright, dim, dark	Yes
Crash occurrence	Location of the crash	Latitude, longitude	Yes
Date of crash	Numerical date on which the crash occurred	Not applicable	No
Time of crash	Time of the crash	Not applicable	No
Fatality	A death caused by the accident	Number of persons fatally injured	Yes
Serious injury	A serious injury caused by the accident	Number of persons severely injured	Yes
Minor injury	A minor injury cause by the accident	Number of persons lightly injured	Yes
Type of geometry	Type of geometry of the road where the accident occurred	Straight, curve, tunnel, bridge, zig zag	Yes
Speed limit	Speed limit of road on which the crash occurred	Number in km/h	Yes
Collisions	Type of collision	Not applicable	No

In total, 2,444 road accident records were available for this analysis (figure 3.4). Of this total, 2,035 were located on the east road, cutting across seven districts (Aceh Tamiang, Aceh Timur, Aceh Utara, Aceh Besar, Biruen, Lhoksemawe and Pidie), and 409 were

located on the west road, cutting across six districts (Aceh Besar, Aceh Barat, Aceh Barat Daya, Aceh Jaya, Aceh Selatan and Subulussalam).

Figure 3-4: Distribution of road accidents along the two main highways of the province of Aceh

The graphs below present some information on the distribution and characteristics of road accidents at the provincial scale and a comparison of some of the accident variables between the east road and the west road.

Figure 3.5 shows that more accidents were recorded on the east road than on the west road. Only one district on the west road records more than 100 accidents (Aceh Selatan). In comparison, six out of seven districts on the east road record more than 100 accidents for the same time period, with some districts recording over 350 accidents.

Figure 3-5: Distribution of analyzed records of road accidents on the west road (WR) and the east road (ER)

Figure 3.6 shows the distribution of accidents by severity within each district. In Indonesia, four categories of severity levels are used to classify road accidents: fatal injury, serious injury, minor injury and property damage only. According to a government regulation from 1993, a fatal accident is one in which a victim died, either on the spot or within 30 days due to injuries sustained in the accident. A serious injury is one in which the victim suffered serious injuries, was admitted to a hospital and received treatment for over 30 days; a minor injury is one in which the victim required medical treatment or was admitted to a hospital and received treatment for less than 30 days.

The definition of a fatal accident used in Indonesia aligns with the definition of a fatal accident described in the Vienna Convention ('an accident in which a person died of their injuries at the scene or within thirty days') (United Nations Treaty Collection, 2019). This definition has also been adopted in most countries of the European Union.

The definition of a serious injury used in Indonesia differs from the one used in European countries. In European countries, the definition describes a serious injury as a person injured who was hospitalised for a period of more than 24 hours (Eurostat, 2020). In Indonesia, the threshold used is over 30 days. Given the stricter definition in Indonesia, this means that the statistics reporting on severe injuries in this country refer to people in more serious health conditions as they would have been in the hospital for at least 30 days. This also means that statistics reporting on minor injuries in Indonesia include some people who, in Europe, would be categorized as severely injured. Therefore, the statistics (both minor and severe injuries) reported in Indonesia could in fact reflect a higher degree of severity than those same statistics reported in Europe. These

differences in reporting (due to differences in adopted definitions) may affect road safety assessments and indicators.

It can be seen from the graph in figure 3.6 that minor injuries are the dominant type of accident within each district, except in Pidie, where major injuries were dominant, and in Subulussalam, where fatal injuries were dominant. More generally, while the trend in the number of major injuries in Indonesia has been decreasing, trends in fatalities and minor injuries in Aceh increased continuously between 2012 and 2015. The number of major injuries recorded in the province has been decreasing since 2013.

Figure 3-6: Distribution of accidents by severity within each district on the west road (WR) and the east road (ER)

Figure 3.7 shows the distribution of accidents by lighting conditions within each district. Most accidents happen during bright light conditions. Fewer accidents occur under dim light or dark conditions.

Figure 3-7: Distribution of accidents by lighting conditions within each district on the west road (WR) and the east road (ER)

Figure 3.8 shows the distribution of accidents by weather conditions within each district. In each district, many accidents happen during bright weather conditions. All districts record some accidents under rain conditions, but these represented a higher proportion of all accidents in west road districts.

Figure 3-8: Distribution of accidents by weather conditions within each district on the west road (WR) and the east road (ER)

3.3. Road data

The road information was obtained from the Department of Public Work of Banda Aceh's Road Management System. It was provided in shape file format, which is a format used for analyses in GIS software such as ESRI's ArcGIS. The spatial layer recorded partial or full information about the main roads of the province. The associated attribute table included information such as district, number of segments, segment length, average annual daily traffic (AADT), pavement width, type of geometry and land use code. Table 3.2 describes the different road variables. A windy (or winding) alignment for example is a part of the road that twists and turns and is very sinuous. A meandering alignment is a part of the road that is curving or rambling. In general, windy roads have sharper curves than meandering roads.

Table 3-2: Description of road inventory variables

Variable	Accepted values	Unit	Included in the models?	Description
Land use	Rural; Rural and urban; Urban	Not applicable	Yes	Land use type covering the area
Alignment horizontal	2 (Straight) 3 (Meandering) 5 (Windy)	Meter	Yes	A series of description of the horizontal alignment of road segments
Alignment vertical	1 (Flat) 4 (Rolling)	Meter	Yes	A series of description of the vertical alignment of road segments
Annual average daily traffic (AADT)	Values vary by segment	Number of vehicles	Yes	The total volume of vehicle traffic on the road
Width of ditch	1	Meter	No	Width of ditch
Width of shoulder	1	Meter	No	Width of shoulder
Width of surface	7	Meter	No	A measurement of width of surface
Median	0	Meter	No	Width of road median
Lane	2	Not applicable	No	Number of lanes on the road
Access density	Values vary by segment	Number of accidents per segment	Yes	Density of access in each segment calculated as the

					total number of access roads to the segment divided by the total length in km of the segment
Length	Values vary by segment	Meter and km	Yes	Length in meters (length_m) and in km (length_km) of the road segment	

In Indonesia as in Aceh, the road network is classified into four road type categories: national roads (including highways), provincial roads, district roads, and village roads (figure 3.9). National roads include all arterial and collector roads in the primary networks that connect the provincial capitals in addition to national strategic roads and highways. Provincial roads include provincial strategic roads and collector roads in the primary road system that connects provincial capitals to district or city capitals or inter-capitals. District roads represent all local roads in the primary road system that connect capital cities with capital districts, inter-district capitals, regent capitals with local activity centres, inter-local activity centres. They also include strategic district roads and public roads in a district's secondary road network. Finally, village roads include all public roads that connect the region on the village as well as environment roads.

Figure 3-9: Distribution of the road types in Aceh

The Department of Public Work manages all four types of roads. However, more attention is paid to national roads (especially highways), and therefore more data and information are collected and available on highways, as these are the roads with the greatest traffic density. Given this data availability, this analysis focuses on the highways only. In total, 1,053 km of highways are analysed (figure 3.10). Annex 1 presents photos of road conditions and the surrounding environment from selected locations on the two main highways studied in the thesis.

Figure 3-10: Location of the two main highways of the province of Aceh

The east road (highway 1) runs from Banda Aceh to Medan, is 466 km long and passes through seven districts: Aceh Besar, Pidie, Bireun, Lhoksemawe, Aceh Utara, Aceh Timur and Aceh Tamiang. This road is a main artery connecting the northern and southern parts of the province on the east coast. The west road (highway 2) runs from Banda Aceh to Subulassalam, is 587 km long and passes through six districts: Aceh Besar, Aceh Jaya, Aceh Barat, Aceh Barat Daya, Aceh Selatan and Subulassalam. It joins the northern part of the province with the southern part on the west coast. Figure 3.11 represents a cross-sectional diagram of the East and West highways.

Figure 3-11: Cross-sectional diagram of the East and West highways (Source: Indonesian Ministry of Public Work, Aceh Province)

All the units on the diagram are in centimetres. The width of the shoulder is 1m, the width of the ditch is 1m. There are two lanes, and the width of each lane is 3.5m.

Table 3.3 presents the accident density per district. As can be seen, the districts in how much highway they enclose. All districts have more than 20 km of road, and four districts (Aceh Jaya, Aceh Barat Daya, Aceh Selatan and Aceh Timur) have over 100 km of highway each. Lhokseumawe is the district with the highest accident density and Pidie is the district with the lowest.

Table 3-3: Accident density per district

District name	Road section	Road length (km)	Number of accidents	Accident density (accidents per km)
Aceh Besar (West)	West	56.80	36	0.63
Aceh Jaya	West	134.10	92	0.68
Aceh Barat	West	58.70	47	0.80
Aceh Barat Daya	West	132.50	61	0.46
Aceh Selatan	West	165.50	144	0.87
Subulussalam	West	39.60	29	0.73
Aceh Besar (East)	East	77.10	202	2.62
Pidie	East	90.20	19	0.21
Bireun	East	69.90	366	5.23
Aceh Utara	East	63.00	470	7.46
Lhoksemawe	East	21.50	315	14.65
Aceh Timur	East	119.60	519	4.33
Aceh Tamiang	East	24.60	144	5.85
TOTAL		1,053.10	2,444	

Table 3.4 below shows the land use proportions in the districts. For this analysis, three types of land use are considered ('urban', 'rural', and 'urban-rural'). An 'urban' area is an area dominated by heterogeneous land use activities (for example, commercial areas, offices, or services). A 'rural' area is an area mostly dominated by homogenous land use activities, such as agricultural areas. Lastly, an 'urban-rural' area is defined as a rural area characterized by heterogeneous land use activities that are found in an urban area, and that is dominated by trade service functions.

The dominant land use in most districts is urban. Three districts [Aceh Besar (on both East and West side), Aceh Jaya and Subulussalam] have mostly rural land use. When accident densities from table 3.3 are compared with land use proportions, it seems that districts with a larger urban proper tend to have higher accident densities.

Table 3-4: Land use proportions along highways in each district

District name	Rural (km)	Urban (km)	Urban and rural (km)
Aceh Besar (West)	40.50	15.30	1.00
Aceh Jaya	75.10	44.00	15.00
Aceh Barat	19.00	35.70	4.00
Aceh Barat Daya	44.00	74.00	14.50
Aceh Selatan	67.30	80.20	18.00
Subulussalam	25.00	13.60	1.00
Aceh Besar (East)	37.06	28.06	11.99
Pidie	26.31	50.17	13.71
Bireun	10.01	56.88	3.00
Aceh Utara	6.00	54.00	3.00
Lhoksemawe	-	20.50	1.00
Aceh Timur	22.00	89.60	8.00
Aceh Tamiang	1.00	20.60	3.00

Figure 3.12 shows the distribution of the different types of land use on the two main highways. Land use 'urban' is predominant near important cities such as Lhoskemawe, Bireun, Sigli or Banda Aceh on the east road, and Meulaboh or Blang Pidie on the west road; but the 'urban' land use also extent near smaller towns such as Ulee Gle, Idirayeuk or Langsa on the east road or Calang and Takap Tuan on the west road.

Figure 3-12: Distribution of the 'urban', urban-rural' and 'rural' land uses on the two main highways

4. METHODOLOGY

4.1. Introduction

This chapter presents in detail the different models that are used to achieve the objectives of the thesis. The first section identifies the three models selected. The data collection and preparation phases are then explained in detailed. Finally, each of the three methodologies is further described, with background on the methodology and the associated tools that are used in the next chapter.

4.2. Selection of the models

The various methodologies presented in chapter 2 have helped broaden the global understanding of road crashes at different scales, from segment-level applications to more global network-level analyses. As more data are produced and made more readily available, the complexity of the methodologies being developed increases, producing more detailed results.

These methodologies can be classified into two broad categories: non-spatial and spatial methodologies. Non-spatial methodologies used to analyse road accidents are very diverse. In total, eleven methodologies were reviewed, and their advantages and drawbacks were discussed. The most commonly used approaches include Poisson regression models, negative binomial models and zero-inflated Poisson models. Some of these methods were reviewed and tested on the dataset for this thesis. Based on the results and on the available data, it was determined that the most suitable method for achieving the objectives of this research was the negative binomial approach, in part due to its ability to handle over-dispersion. Spatial methodologies, based on GIS, are now used extensively by many road traffic agencies around the world. These methods have improved the efficiency and effectiveness of transport operations by incorporating knowledge of crash characteristics that could not be observed with non-spatial

methodologies. Newer methodologies include more sophisticated, complex and resource-intensive spatio-statistical models. For example, NKDE has been widely applied as a tool to visualize accident analyses when quantitative statistical inference assessments are missing. Areas of high and low densities can also be represented simultaneously with NKDE. Bayesian spatial modelling considers the distribution of data and the unknown coefficient estimators. It also shows the considerable potential of Bayesian models in crash-rate analysis.

Therefore, based on the drawbacks and benefits of the different methodologies presented in chapter 2, based on the results obtained with some of the methods tested, based on the objectives of the thesis, and finally based on the available data, three methodologies were selected and are detailed in this chapter. These methodologies are as follows: (a) network kernel density estimation (NetKDE), (b) hot zones, and (c) the Bayesian spatial model.

NetKDE is used to identify concentrations of crashes on the two roads (hot spots), and the hot zones methodology is used to identify hazardous road locations that consist of two or more contiguous segments. Finally, the Bayesian spatial model is used to identify factors that contribute most to road crashes.

So far, the NetKDE and hot zone methodologies have mostly been used in macro-level analyses; but it is believed that they could also provide informative results when used at the micro level.

4.3. Data preparation

The data preparation step consists in cleaning and formatting the accidents and road data for subsequent analyses. This phase involves two intermediate steps: (1) the cleaning of the datasets and (2) the GIS formatting of the data.

4.3.1. Cleaning of the data

Large datasets are often not error-free. Therefore, a first key step after obtaining the relevant datasets was to quality check the information they contained.

The Excel file containing road accident records was cleaned, and information for each variable was reviewed and standardized. Road accidents for which no coordinates were available were removed from the file. After review and cleaning, 2,094 accident records for the east road accident and 409 for the west road were considered for the analysis.

Quality checks performed on the accident files included, for example, ensuring all point locations provided were located in Aceh province and ensuring information provided for each field was an acceptable value (or was not reported).

The spatial layer representing the road network was also reviewed and cleaned. The digitalization of the road was checked thoroughly and edited when necessary, using ArcGIS Open Street Map base map as a reference. The latter is a web map that references the live tiled map service from the Open Street Map project, which aims to create a free editable map of the world.

A large section of the road shape file provided was edited manually. In the shapefile provided, many road sections were digitized too simplistically, without representing curves and other important turns, and were therefore not representative of the network design on the ground. For example, in some cases, the line representing the road was drawn far from the road's actual location, and in other instances, the different segments representing the road were not connected together. A quick review and repair of the geometry highlighted vertices located at the extremity of each line.

4.3.2. GIS formatting of the data

Once the road crashes file was cleaned and the attribute information formatted, the file was imported into Arcmap. The accident records were georeferenced and mapped using an appropriate projection system, in this case UTM Zone 47N, which covers the northern part of the province.

The original road layer provided by the Department of Public Work included a limited amount of qualitative information in the spatial layer attribute table. A schema was created to compile into one file the different pieces of road-related information provided separately in the spatial layer. Available information included, for example, information on land use, AADT, alignment and speed limit. The new formatted layer, containing both the east and west roads, was then projected in the same projection as the road crash layer (UTM Zone 47N).

Then, the east road and the west road were divided into segments based on the vertical and horizontal alignment information, the AADT information and the geometric design of the road. This road layer was then used in the Bayesian model. The layer was also integrated in the hot zone and SANET tool, and further split into equal segments 200 meters long.

A spatial join between the accidents point layer and the road layer was created in Arcmap to combine into one layer the road accident variables and the road variables. Two new columns were added to this layer. The first field records information on the

number of intersections along a segment. The second captures the density of accidents in each segment.

4.4. Network kernel density analysis

4.4.1. Background

KDE methods have been very popular in the field of traffic safety over recent decades. They have notably been used to identify road crash hot spots. These methods use the Euclidean distance on a plane (planar spatial analysis). As spatial data become more abundant and accessible, and as GISs become more prevalent, the existing methodologies quickly revealed some limitations given the new questions they were used to answer. For example, the use of ordinary KDE creates patterns of accidents that frequently seem more clustered than they really are (Okabe et al., 2009).

To address some of the limitations associated with the ordinary KDE methodologies, a new series of KDE methods based on networks have been developed (also referred to as network-based KDE, or NKDE). These new methodologies were designed to estimate the density distribution of point events over a one-dimensional network space. Whereas ordinary KDE uses a planar area of influence, and every accident has a circle (or curve, contingent upon the type of kernel function) with a defined bandwidth, NKDE considers each point's influence along a system, and the network characterizes the shortest path distance.

Flahaut et al. (2003), for example, developed a method to estimate kernel density on a simple network defined as a line. However, this method presented some limitations when used to comprehend the complexities associated with a real road network (for example, the network topology is not considered). Borruo et al. (2006) developed another method to estimate kernel density on a network with a GIS. This methodology did not focus on the network only but also integrated the overall region in which the network was embedded, thereby looking at a two-dimensional space rather than being constrained to one dimension.

4.4.2. Network Kernel Density Estimation (NetKDE)

Okabe et al. (2006) developed another network-based KDE to overcome some of the issues found in other network-based methods. This method was embedded into a GIS tool named 'Kernel density estimation (NetKDE)'. This tool (and the method upon which it was developed) was created to address some of the drawbacks of using a planar kernel for network-based analyses. In their method, the researchers developed two unbiased kernel functions (discontinuous and continuous). These kernel functions assume that

events occur only in the given network and do not consider what happens in the broader region in which the network exists.

NetKDE is one among 16 tools created in the SANET toolbox. The latter is an ESRI toolbox developed by Atsuyuki Okabe and his team at the University of Tokyo (Okabe et al., 2006). The toolbox is freely available for non-commercial research and can be accessed at <http://sanet.csis.u-tokyo.ac.jp/>.

4.4.3. General framework of NetKDE

The principle behind the NetKDE tool is to estimate a nonparametric estimation based on known point locations. Let N be a roadway network with V ($v_1, v_2, v_3, \dots, v_n$) as the nodes on the network and L ($l_1, l_2, l_3, \dots, l_n$) as the roadway links on that network connecting the nodes. It creates a sub network L_q such that the shortest-path distance between q and any point on L_q is less than or equal to h , where q is an arbitrary point connected on any link L , and h is the chosen bandwidth. For a given point q and some arbitrary point p on any link L , the function $K_q(p)$ is defined as equation 4.1.

$$K_q(p) \begin{cases} \geq 0 & \text{for } q \in L_q \\ = 0 & \text{for } q \in \tilde{L} / L_q \end{cases}$$

$$\int K_q(p) dp = 1 \quad (4.1)$$

where L is the union of all links and $p \in L_q$ is the integration of $K_q(p)$ dp along the line segment of L_q . Here, $K_q(p)$ is the network kernel density function value at q , with q as the kernel centre. For the given sample points $p_1, p_2, \dots, p_n$ on L , a function $K(p)$ is given as equation 4.2.

$$K(p) = \frac{1}{n} \sum_{i=1}^n K_{p_i}(p) \quad (4.2)$$

where $K(p)$ is the network kernel density estimator for $f(p)$ at p .

This creates a kernel function for all the individual points on the network.

However, the calculation of NetKDE is changes if the point comes across as node where it has to create function on the two roadway links. In such situations, Okabe et al. (2006) proposed two unbiased methods: (a) equal-split discontinuous kernel density function, and (b) equal-split continuous kernel density function. The equal-split discontinuous function is defined as equation 4.3.

$$K_y(x) = \begin{cases} \frac{k(d(x, y))}{(n_1 - 1)(n_2 - 1) \cdots (n_s - 1)} & \text{for } 0 \leq d(y, x) \leq h, \\ 0, & \text{for } d(y, x) \geq h. \end{cases} \quad (4.3)$$

The equal-split continuous function is defined as equation 4.4.

$$K_y(x) = \begin{cases} k(x) & \text{for } -h \leq x \leq 2d - h, \\ k(x) - \frac{n-2}{n}k(2d-x) & \text{for } 2d - h \leq x \leq d, \\ \frac{2}{n}k(x) & \text{for } d \leq x \leq h. \end{cases} \quad (4.4)$$

In essence, these two kernel functions differ in the designation of values to l2, l3 and l4. In the continuous function, the values are adjusted to make the function continuous in the 'local area' around the vertex v (i.e. the area in which the distance from v is within h-d). The first method is preferred over the second, as the latter requires more computational time (which varies according to the complexity of the kernel support).

When using the equal split discontinuous function, two cases arise: (1) the kernel centre q does not coincide with a node in V and (2) the kernel centre q coincides with a node in V.

In the first case, a buffer network of q with a width of h is constructed with the shortest paths from q to the boundary points of the buffer network bi. If the boundary distance $0 \leq d_s(q, p) \leq d_s(q, v_{i1})$, it creates the function as $K_q = K(d_s(q, p))$, but when it reaches a node vi1, it splits the link which comes from the centre q into $n_{i1} - 1$, where n_{i1} is the degree of node vi1. It then distributes the value to every link that is attached to node vi1. This process continues until it reaches the boundary bi. The function $K_q(p)$ is defined as equation 4.5.

$$K_q(P) = \begin{cases} \frac{K(d_g(q, p))}{(n_{12}-1)(n_{12}-1) \cdots (n_{ik-1}-1)} & \text{for } d_s(q, v_{ik-1}) \leq d_s(q, p) < d_s(q, v_{ik}) \\ & \text{or for } d_s(q, p) \geq h \end{cases} \quad (4.5)$$

where $k = 1, \dots, m$; $v_{i0} = q$; and $v_{im} = b_i$.

In the second case, the value of function $K_q(v_{i1})$ is divided by n_{i1} and assigned to the roadway links. The function in this case is defined as equation 4.6.

$$K_q(P) = \left\{ \frac{2K(d_g(q,p))}{(n_{12}-1)(n_{12}-1)\dots(n_{ik-1}-1)} \right\} \begin{matrix} \text{for } d_s(q(V_{ik}-1) \leq d_s(q,p) < d_s(q,v_{ik})) \\ \text{or for } d_s(q,p) \geq h \end{matrix} \quad (4.6)$$

where $k = 2, \dots, m$.

The computation of the equal-split discontinuous method when there are more than two paths available is critical when the length is less than h from the centre q . For example, when the paths overlap on the network, N is carried out in three steps. In the first step, the method modifies the given network by creating dummy nodes for every link near to the nodes that have degree more than two, such that these dummy nodes divide the existing links further in order to manage the multiple values at every node with a degree more than two. In the second step, the method computes the split ratio at every node with the length of the bandwidth h starting from the kernel centre q . The third step sums all the densities obtained from the previous steps with respect to all kernel centres. As the choice of bandwidth effects the density estimation, studies suggest using a 100- to 300-meter bandwidth (Okabe and Sugihara, 2012), which can be taken as a rule of thumb when applying the methodology to urban areas. A 200-meter bandwidth was used for the network kernel estimation analysis in this study, which was determined based on trial and error (Okabe and Sugihara, 2012).

4.5. Hot zones identification

4.5.1. Background

Traditionally, the hot spot methodology has been popular in road safety analyses for identifying hazardous locations in a road network. In fact, it has been used extensively by road traffic agencies all over the world to detect areas with high concentrations of road crashes. The results of these analyses have been particularly useful to decision-makers, who with limited budgets could quickly identify the most dangerous sections of road that needed investments.

Some drawbacks associated with the hot spot methodology led to the development of the hot zone methodology. Whereas the hot spot methodology identifies point locations in a network, the hot zone methodology identifies hazardous road locations that consist of two or more contiguous segments. Indeed, some contiguous road segments may be more dangerous than a hot spot alone. Given that actual crash records determine the length of hot zones, they are not fixed but can vary in length. Secondly, unlike the hot spot methodology, the hot zone methodology does not consider the individual road segments as spatially independent units.

Many researchers have started to investigate the location of hot zones amongst hot spots to better understand the underlying hazardous road locations, and many have suggested using the hot zone methodology to complement the hot spot methodology (Loo, 2009; Flahaut et al., 2003; Moons et al., 2009).

4.5.2. Steps in hot zone identification

Loo (2009) developed a three-stage methodology to identify hot zones in a road network in a systematic and scientific manner. Her approach has been used by many other researchers. Recently, Grubenmann et al. (2019) developed an add-in tool for ArcGIS called the 'Hotzone Generation Addin' that identifies hazardous road locations and detects road crash hot zones. The Python-based add-in is used to automate and speed up the spatial process of hot zone identification, and it uses the three-step hot zone methodology, as described by Loo (2009).

The first step is to geo-validate road crashes by combining the crash database, the road network and the administrative database to validate crash locations. The second step is to cut the road network into segments, or BSUs, and to calculate crash statistics. The segments or BSUs should be long enough to allow for the identification of crash concentrations but short enough to reflect changes in the environment that could influence road safety. A few authors have used 100-meter segments (Black and Thomas, 1998; Flahaut et al., 2003; Flahaut, 2004). Finally, the last step involves detecting hot zones. The algorithm that was developed for this step is presented in figure 4.1 (Loo, 2009).

The algorithm starts with the first BSU and compares its actual crash rate (AR) with the threshold value. The crash rate describes the number of crashes in a given period as compared to the traffic volume (or exposure) to crashes. The crash rate is calculated by dividing the total number of crashes at a given road section over a specified period of time (commonly 3 to 5 years) by a measure of exposure. If the AR value is equal to or greater than the threshold crash rate, a new working table is created with an index number equal to one, and all contiguous BSUs are listed in this table. Then, each contiguous BSU is analysed, and whenever the AR value of the contiguous BSU is equal to or greater than the threshold crash rate, the index number increases by one. Once all the contiguous BSUs listed in the table have been checked, the index number is examined. If it is greater than one, a hot zone has been identified, and the variable indicating a hot zone is updated in the main table (Loo, 2009).

Figure 4-1: Flowchart for identifying hot zones (Loo, 2009)

4.6. Bayesian spatial model

4.6.1. Background

Crash frequency models, used at different scales, have been widely used by traffic agencies around the world to identify road crash locations and to analyse contributing factors in order to determine appropriate countermeasures to improve road safety. Among the benefits they offer, Bayesian models can predict risks and capture the variability of estimates, even in the presence of sparse data or rare events, such as in the study area of this thesis (province of Aceh). Bayesian models can also incorporate

prior knowledge without the restriction of distributional assumptions (Withers, 2002). They can produce comprehensive information through the conditional probabilities of interrelationships between the variables and identify the dependencies of the variables involved in the occurrence of road crashes.

Bayesian models using the MCMC framework have been commonly used because they are more flexible, they can account for spatial correlation, and they can handle complex models. Many authors have used these models to model crash risk at different scales and identify the relationships between, for example, road infrastructures, traffic-related factors and socio-economic and demographic factors (Song, 2006; Li et al., 2007; Agüero-Valverde and Jovani, 2008; Agüero 2010).

However, this approach has its drawbacks, notably in terms of convergence and computational time. For example, because the simulation time is dependent on the sample size and the model complexity, as models become complex, their computational time increases significantly. Despite all the benefits they provide, Bayesian models would still represent a barrier to developing more complex model forms (Lord and Mannering, 2010).

To overcome these limitations and as an alternative to the MCMC method, the INLA approach was recently developed with improved computational capabilities (Hue et al., 2009). In the field of traffic safety, this approach has been increasingly used over MCMC methods, notably because the INLA approach allows users to specify more accurate prior distributions and to consider the spatial and temporal structure of crashes in the inferential process which is used to derive the posterior distributions (Blangiardo et al., 2013; Wang et al., 2017).

4.6.2. Development of the models

The general model specification used in this analysis is as in equation 4.7.

$$Y_{it} \sim \text{Poisson}(\theta_{it}) \quad (4.7)$$

where

y_{it} is the number of crashes observed in segment i at time t (in years), and

θ_{it} is the expected Poisson rate (i.e. expected crash frequency) for segment i at time t .

The Poisson rate is modelled as a function of the covariates following the lognormal distribution stated by equation 4.8.

$$\text{Log}(\theta_{it}) = \beta_0 + \sum_k \beta_k x_{itk} + \varepsilon_{it} \quad (4.8)$$

where

β_0 is the intercept,

β_k is the coefficient for the k^{th} covariate,

x_{itk} is the value for the k^{th} covariate (or any suitable transformation of the covariate) for segment i at time t , and

ε_{it} is the error term for segment i at time t .

More precisely, a Poisson model with a Besag York Mollie (BYM) specification (spatial term) is assumed and applied in a Bayesian framework. This model is in fact, a union of the Besag model or the conditionally autoregressive term u and an independent identically distributed (iid) error term v . The benefit of this model is that it allows identification of the posterior marginal of the sum of the spatial and the iid model, as in equation 4.9.

$$\varepsilon_{it} = \begin{pmatrix} v_{it} + u_{it} \\ u_{it} \end{pmatrix} \quad (4.9)$$

with the length of x being $2n$ if the lengths of u and v are n . Then v is defined as equation 4.10.

$$V_i \sim N(0, \tau_v) \quad (4.10)$$

where

τ_v is the precision (inverse of the variance) and controls the Poisson extra-variation due to heterogeneity.

The random effects are pooled over time to improve the model estimation. The assumption of constant random effects over time is not restrictive, provided that the covariates explain most of this variation over time.

The spatially correlated effect is modelled using a Gaussian CAR prior as shown by equation 4.11 (Besag et al., 1991).

$$u_i \mid u_{-i} \sim N\left(\frac{\sum_{j \sim i} w_{ij} u_j}{w_{i+}}, \frac{1}{w_{i+} \tau_u}\right) \quad (4.11)$$

where u_{-i} refers to all the neighbours of i ,

τ_u controls the Poisson extra-variation due to clustering or spatial correlation,

$j \sim i$ denotes that segment j is a neighbour of segment i ,

w_{ij} is the weight of the j^{th} neighbour of the i^{th} segment, and

w_{i+} is the sum of the weights of the neighbours of segment i .

The hyper parameters are the precision τ_{v1} of the iid model (v) and the precision τ_u of the spatial model (u). The precision parameters are represented as equation 4.12.

$$\boldsymbol{\tau} = (\theta_{v1}, \theta_{u1}) = (\log \tau_v, \log \tau_u) \quad (4.12)$$

and the priors are defined on $\boldsymbol{\tau}$ such that $\tau_v, \tau_u \sim \text{loggamma}(1, 5 \times 10^{-4})$.

The model is a Poisson model with a BYM specification (spatial term) presented above and is then run using INLA and the R-INLA package. More precisely, the model developed $f()$ is the latent Gaussian field BYM for spatial effects. The hyper parameters are the precision of the iid model and the precision of the Besag models (Rue et al., 2009). The model formula is as in equation 4.13.

$$\text{Formula} = y - a + b + a:b + c * d + f(\text{idx1}, \text{model1}, \dots) + f(\text{idx2}, \text{model2}, \dots) \quad (4.13)$$

where

Formula describes the structured additive linear predictor,

y is the response variable,

the terms $a + b + a:b + c * d$ are responsible for the fixed effects specifications of the model and represent the smooth functions, and

idx1 and idx2 are latent building blocks that are combined together to form a joint latent Gaussian model of interest.

An offset term is sometimes necessary to covariate with the slope. This might be useful in situations when correcting the number of events with an estimate of population size.

4.6.2.1. Database creation using QGIS

To run the Bayesian model, the spatial road layer created in section 4.3 was used. Its attribute table was exported as a separate file (.csv). These two files were used in QGIS to produce the additional files necessary to run the model in R.

QGIS is a free and open source geographic information system (QGIS Development Team, 2019) which enables a user to create, visualise and analyse geospatial information. QSpatialite is a plugin that enables the use of Spatialite Databases in QGIS. Spatialites are SQLite database engines in which spatial functions have been added.

First, a spatial database in QGIS was created using Spatialites. This database was used to create several intermediary tables that were needed to produce the additional files to be used in R. The SQL language was used to create adjacency intermediary tables with selected information coming from the two input files (shapefile and csv). Two of these tables were exported and saved and were later used in R.

4.6.2.2. R-INLA software

MCMC methods have traditionally used WinBUGS. However this software cannot support complex model forms and is slow. An R-INLA package was recently produced. It enables the user to run a large class of statistical models using the INLA approach (Martins et al., 2013). In addition to INLA, three other packages were used (sp, MASS, and Matrix). The 'sp' package uses different methodologies that enable the user to integrate and handle spatial data in R. This is a powerful package and underpins the proper functioning of many GIS-related packages in R. The 'MASS' package provides functions and datasets to support a large range of statistical analyses. The 'Matrix' package was developed to allow manipulating dense and sparse matrices that extend the basic matrix data types.

4.6.3. Goodness of fit measures

Goodness-of-fit measures are used to compare and select Bayesian models that provide the best fit to the data. As explained in Spiegelhalter et al. (2002), model comparison is done by defining a measure of fit to the data and a measure of complexity in the model, the latter represented by the number of model parameters. An increase in model complexity is associated with a better fit to the data, and model comparison is usually done by trading between these two elements (Spiegelhalter et al., 2002; Pooley and Marion, 2018).

Different indicators have been developed to assess the performance of Bayesian models. The DIC developed by Spiegelhalter et al. (2002) is one of the most commonly used goodness-of-fit measures and considers measures of fit and of complexity as in equation 4.14.

$$DIC = \bar{D} + p_D \quad (4.14)$$

where $\bar{D}$ is the posterior deviance and measures the model fit and

p_D is the effective number of parameters and measures the model complexity as shown in equation 4.15.

The DIC measure can also be expressed as:

$$DIC = D(\bar{\theta}) + 2p_D \quad (4.15)$$

where $D(\bar{\theta})$ is the deviance at the posterior mean θ .

Models with smaller DIC values are preferred over models with larger DICs. Models with smaller DICs help to identify the best-fitted models. Once the best-fitted model is determined, other models with a DIC value three to seven points higher than the best model can be discarded. If the DIC value of a model is only one or two points higher than that of the best model, then this model deserves consideration.

Other measures include the Bayesian information criterion, the Akaike information criterion, the posterior deviance mean, and the Watanabe-Akaike information criterion (WAIC). The posterior deviance mean, for example, is a useful measure of how well the model fits the data. Large values of the mean deviance indicate that the data deviates substantially from the model's assumptions. The WAIC is used to estimate the out-of-sample expectation. It computes the log point-wise posterior predictive density and adds a correction for the effective number of parameters to adjust for over-fitting (Gelman et al., 2013). While this criterion has certain limitations (Gelman et al., 2013), it has the advantage of being invariant to reparameterization and can deal with mixture models (Spiegelhalter et al., 2014). WAIC is expressed as equation 4.16.

$$WAIC = -2 * (LPPD - p_D) \quad (4.16)$$

where p_D is an estimate of the effective number of parameters in the model and

$LPPD$ is the log posterior predictive density estimated as equation 4.17.

$$LPPD = \sum_{i=1}^n \log \int p(y_i|\theta) p_{post}(\theta) d\theta \quad (4.17)$$

The logarithm of the pseudo marginal likelihood can be expressed as equation 4.18.

$$LPML = \sum_{i=1}^n \log (CPO_i) \quad (4.18)$$

where CPO is the conditional predictive ordinate defined as equation 4.19.

$$CPO_i = f(y_i|\mathbf{y}_{-i}) \quad (4.19)$$

where y_i is the i -th observation for all segments.

Many studies have been dedicated to comparing the results of these various goodness-of-fit measures (Gelman et al., 2013; Pooley and Marion, 2018). The DIC remains a popular indicator for model selection due to its simplicity of use; its application to a wide range of models, including complex models; and its integration in the BUGS software (Spiegelhalter et al., 2003; Berg et al. 2004; Geedipally et al., 2014).

5. RESULTS AND DISCUSSION

5.1. Introduction

This chapter presents the results of running the three methodologies detailed in the previous chapter. For each methodology, the model parameters are described, and the results are presented and discussed. When applicable, the results of the different methodologies are compared.

Next, based on the national context (and on the broader South-East Asian context), and considering the findings obtained with the three methodologies and the objectives of the thesis, a two-step methodology is put forward. This methodology is based on two of the presented methodologies and is a combination thought to best respond to the objectives of the thesis. It can also be applied in a data-poor context and still provide key information for decision-makers.

5.2. NetKDE approach

This section describes the results of using the NetKDE approach with SANET presented in section 4.4 to estimate the density of accidents on the two main highways of Aceh. The two main layers used in the tool were the road accident layer and the road layer described in chapter 3.

5.2.1. Model parameters

The equal split continuous function was used with the bandwidth set to 200 m. The choice of the bandwidth is important because the estimated density will vary based on the bandwidth. While there are not established values, Porta et al. (2009) suggest a

bandwidth value between 100 and 300 m because these values are commonly used in urban planning. The cell width was set to 20 m as recommended by the SANET team (Okabe et al., 2006).

5.2.2. Results and discussion

Figure 5.1 presents the location of the hot spots identified using SANET. The colour scale on the map represents the concentration of hot spots, from green (no hot spot) to yellow (light concentration of hot spots) to red (important concentrations of hot spots). Sections of the road represented in red thus show hazardous segments characterized by high crash density.

Figure 5-1: Location of the hot spots identified on the east and west roads using SANET. The inset zooms in on the district of Lhokseumawe

More precisely, the concentration of hot spots is greater on the east road than on the west road. On the east road, hot spots were detected in three districts, Banda Aceh, Lhokseumawe and Aceh Timur. It is in the district of Lhokseumawe that the highest concentrations of hot spots can be seen, as represented by the continuous concentration of red dots on the maps. Additionally, in all three districts the hot spots were located in urban environments. Only a few hot spots were detected on the west road, located in the districts of Aceh Selatan, Aceh Besar and Aceh Jaya, and they are

more sparsely distributed than those on the east road. Overall, the map shows that the density of crashes is lower on the west road. Additional maps are presented in annex 2.

5.3. Hot zone approach

5.3.1. Model parameters

The two main layers used in the tool were the road accident layer and the road layers described in section 4.3.

After snapping the traffic crashes on the BSUs, each BSU recorded a certain number of crashes. Contiguous BSUs with an actual crash rate (AR) exceeding the critical crash rate (CR) were considered hot zones. The Python-based add-in developed by Grubenmann, Tsoi and Loo (2019a) and described in section 4.4.2 was used to automate and speed up the spatial process of hot zone identification. The add-in offers three BSU sizes (50 m, 100 m and 200 m) and four thresholds (3, 4, 5 or 6 crashes in a year in the same segment). These four threshold values are pre-defined in the tool; however, the user can also define different values, based on the local context. As a hot zone is defined by at least two contiguous segments, the minimum threshold value that can be entered in the tool is '2'. The pre-defined threshold values set up in the tool are used in this analysis, as areas of 'hot spots' in Indonesia are usually made of 6 or more crashes.

All 12 possible combinations were tested. Only three of the combinations identified any hot zones (combination 1: BSU length = 200 m, threshold = 3; combination 2: BSU length = 100 m, threshold = 3; combination 3: BSU length = 200 m, threshold = 4). The other combinations did not identify any hot zones. After review and analysis of all results of the three productive combinations, the first combination was selected because it provided the most interesting results for the local context.

5.3.2. Results and discussion

To identify the hot zones, a minimum threshold value of 3 crashes was used, and the BSU length used was 200 meters. In total, 99 hot zones were identified in all districts on the east road in the 2012–2015 period. Only the district of Pidie did not record any hot zone. On the west road, no hot zones were identified in any district. On the east road, the districts of Lhokseumawe and Aceh Utara recorded the highest numbers of hot zones (42 and 29, respectively). Districts with high numbers of hot zones are those with higher crash and population densities. Apart from Banda Aceh, the five districts with the highest numbers of hot zones are the most densely populated and have the highest crash densities. The longest hot zones (Max (n)) examined consisted of 8 contiguous BSUs, representing a 1.6-km segment. The maximum crash rate for an individual BSU,

for the two threshold values, was 11. Additional maps of the hot zones are presented in annex 3.

The distribution of the hot zones also shows that they were located primarily in urban environments (over 60%) (table 5.1). Figure 5.2 shows all the hot zones identified on the east road and zooms in on the district of Lhokseumawe to illustrate the distribution of hot zones in this district.

Table 5-1: Hot zones per district on the east and west roads

District	Road	Hot zones	Hot zones in urban (%)	Crash density	Pop density (2015)
Aceh Barat Daya	West	0	0	0.16	94.38
Aceh Besar	East	7	71.43	2.67	132.22
Aceh Besar	West	0	0	0.60	132.22
Aceh Jaya	West	0	0	0.68	22.65
Aceh Selatan	West	0	0	0.85	58.54
Aceh Tamiang	East	5	80.00	5.57	142.24
Aceh Timur	East	7	100.00	4.11	64.10
Aceh Utara	East	29	82.76	7.27	180.38
Banda Aceh	West	0	0	5.00	4,079.25
Bireun	East	6	66.67	5.18	228.96
Lhokseumawe	East	42	95.24	14.57	1,057.14
Pidie	East	0	0	0.21	135.69
Subulussalam	West	0	0	0.83	54.05

Figure 5-2: Location of the hot zones on the east road

A closer look at the 20 most dangerous hot zones in terms of both concentration of crashes and total fatalities shows that 12 out of the 20 hot zones were located in the district of Lhokseumawe (figure 5.3). This highlights the need to prioritize this district and set up appropriate countermeasures. In addition, all but two hot zones were located in urban environments. The two non-urban hot zones were located in an urban-rural environment in the district of Aceh Utara and in a rural environment in the district of Aceh Besar. To understand how the hot zones rank in terms of injury-severity level, each severity level was weighted (fatality = 1.15, major injury = 1.05 and minor injury = 0.95), and the total of weighted road safety burden was calculated. The hot zones that rank highest in terms of number of crashes also rank high in terms of injury-severity level. However, three hot zones in Lhokseumawe which rank 9, 12 and 13 in terms of number of crashes have higher weighted injury severities than other hot zones ranking higher in terms of number of crashes. This means that while these hot zones recorded fewer traffic crashes, those crashes tended to be more serious.

Figure 5-3: Location of the top 20 hot zones by number of fatalities

When looking at each hot zone individually and analysing environmental characteristics, some commonalities emerged (table 5.2). For example, most of the top 20 hot zones were sections of the road that contained many intersections. Some of these hot zones contained markets, which means that these sections were likely to have many pedestrians walking in these areas during market hours. Other hot zones included access roads to a railway station. These areas were likely to have much traffic and many pedestrians going to and from the train station. Some hot zones also encompassed shopping centres, implying intense traffic during the weekends, and schools, meaning substantial road traffic and motorcycles and pedestrians during term time. Some of these hot zones also included sharp curves, which can lead to road crashes. In the other 79 hot zones identified on the east road, commonalities across the hot zones included relatively high numbers of intersections in the hot zones and the presence of schools, universities, hospitals, shopping centres, markets and housing along the road sections.

Table 5-2: Top 20 hot zones by number of fatalities on the east road

Rank	Number of crashes	District	Land use	Total fatalities	Total major injuries	Total minor injuries	Total weighted	Environmental characteristics
1	11	Aceh Utara	Urban	7	0	13	20.4	Straight after sharp curve, 6 intersections, settlements, access road to railway station
2	6	Aceh Tamiang	Urban	6	1	3	10.8	Curve, rural housings
3	6	Aceh Timur	Urban	4	1	8	13.25	Curve, 5 intersections
4	8	Aceh Utara	Urban-rural	3	0	10	12.95	Straight road, 4 intersections, shopping places, settlements, access to railway station
5	5	Aceh Besar	Rural	2	2	6	10.1	2 intersections
6	7	Lhoksemawe	Urban	2	1	14	16.65	Slightly curvy, 2 intersections, settlements along the segment
7	6	Lhoksemawe	Urban	2	0	10	11.8	Meandering, 4 intersections, settlements along the segment, market
8	11	Lhoksemawe	Urban	1	2	24	26.05	Straight road after sharp curve, gas station, market, gas factory, settlements along the segment
9	8	Aceh Utara	Urban	1	2	21	23.2	Sharp curve, 3 intersections, settlements, access road to railway station, heat centre, police office
10	11	Aceh Utara	Urban	1	2	20	22.25	Straight road, 5 intersections, shopping places, settlements, access to railway station
11	10	Lhoksemawe	Urban	1	1	20	21.2	Straight road, 4 intersections, settlements along the segment
12	7	Lhoksemawe	Urban	1	1	16	17.4	Meandering, 3 intersections, settlements along the segment, market
13	7	Lhoksemawe	Urban	1	0	15	15.4	Straight road, 4 intersections, 1 fertilizer factory, settlements

14	7	Lhoksemawe	Urban	1	0	11	11.6	Straight road, 6 intersections, 1 fertilizer factory, settlements, access to railway station
15	7	Lhoksemawe	Urban	1	0	9	9.7	Curvy road, settlements, shopping place, market
16	5	Aceh Utara	Urban	0	0	13	12.35	Curve in the road, 3 intersections, 1 connecting road to road district
17	8	Lhoksemawe	Urban	0	0	12	11.4	Straight, 2 intersections, settlements along the segment, school
18	6	Lhoksemawe	Urban	0	0	10	9.5	Straight, 2 intersections, settlements along the segment
19	7	Lhoksemawe	Urban	0	0	10	9.5	Curvy road, settlements, shopping place, market
20	5	Lhoksemawe	Urban	0	0	9	8.55	Straight road, 3 intersections, settlements, access road to railway station

5.4. Bayesian spatial approach

5.4.1. Model parameters

SPFs are crash prediction models. They are mathematical equations that relate the number of crashes of different types to site characteristics. These models, most generally, include traffic volume (AADT) but may also include some site characteristics, including lane width, shoulder width, radius or degree of horizontal curves, presence of turn lanes, and traffic control at intersections.

In this analysis, the SPFs of the Bayesian spatial model are based on the formula described in section 4.6. The general formula of the crash is presented below. In total, five scenarios were run: total number of accidents, fatal accidents, serious injuries, minor injuries and property damage only (PDO).

$$N_{spf} = e^{\beta_0} \cdot L^{\beta_1} AADT^{\beta_2} \cdot \prod_{i=3}^n e^{\beta_i X_i} \quad (5.1)$$

where

- N_{spf} = average predicted crash frequency (crashes per year);
- L = length of the road segment (km);
- $AADT$ = annual average daily traffic volume, possibly disaggregated in multiplicand components for considering diverse traffic units (motor vehicles, cyclists and pedestrians) or the importance of

	the segments intersecting at road intersections (major vs. minor roads) and having separate coefficient estimates for each volume component (motor vehicles, cyclists or pedestrians per year);
X_i	= a set of predictor variables related to the road, traffic features, environmental variables and context variables; and
B_i	= a set of regression coefficients, which include β_0 , β_1 and β_2 .

The models also included another covariate frequently used in SPFs, which is the expected crash frequency for a segment given its length. As a result, the coefficients for AADT and length measure what deviation expected for this site, in absolute or relative terms, is used to rank sites for future safety improvements. The details of all the SPFs used in the different scenarios are presented in annex 4.

5.4.2. Spatial model

The results from the total number of crashes scenario model (table 5.3) identify several statistically significant variables. These are AADT, access density, land use 'urban', and land use 'urban and rural'. These variables are correlated with higher crash frequencies. The variables land use 'urban and rural' and alignment 'flat' were used as a base case in the models, and segment length was used as an offset.

The model indicated that the AADT coefficient was 1.1195 and the mean values of the land use variables 'urban and rural' and 'rural' were 0.2826 and -0.0752, respectively, for the spatial model of total number of crashes.

The variables horizontal alignment 'meandering' and 'windy' have coefficients of -0.2807 and -0.8841, respectively. Windy roads are characterised by sharper curves than meandering roads.

The mean value of the access density variable was 0.0469. In this scenario, the DIC value was 2620.55, and the WAIC and LPML values were 2564.866 and -4536.827, respectively. These results are consistent with previously published papers in which significant improvements were found for the goodness-of-fit measures of spatial models at the segment level (Aguero-Valverde and Jovanis, 2008; El-Basyouny and Sayed, 2009; Huang and Abdel-Aty, 2010; Guo et al., 2017; Huang et al., 2017).

The intercept value -12.0807 with a 95% credible interval equal to (-16.2511, -8.0372), and the coefficient of AADT is = 1.1195 with a 95% credible interval equal to (0.6945, 1.5560). This indicates that the AADT proportion has a positive effect on total crashes. The columns 0.025 and 0.975 quantiles contain the lower and upper limits of 95% credible intervals of accidents and provide measures of uncertainty.

Table 5.4 presents the estimates of the spatial model of the number of fatal crashes scenario. The AADT coefficient in this scenario was 0.6352, and the mean values of the land use variables 'urban and rural' and 'rural' were 0.3138 and -0.1542, respectively. The estimates of the horizontal alignment 'meandering' and 'windy' variables were -0.3948 and -0.8052, respectively, which means that they are negatively correlated to the crashes. The mean value of the access density variable was -0.0246, which means that in this scenario density is not associated with fatal crashes. These findings are also consistent with Aguero-Valverde et al. (2016) and Aguero-Valverde and Jovanis (2008), who found that by including spatial correlation, the bias associated with model misspecification can be reduced, as is illustrated by the reduction of the AADT coefficient. The intercept value -9.4081 with a 95% credible interval equal to (-14.424, -4.4037), and the coefficient of AADT is = 0.6352 with a 95% credible interval equal to (0.1038, 1.1662). This indicates that the AADT proportion has a positive effect on fatal crashes.

Table 5.5 presents the estimates of the major injury crash scenario resulting from the spatial models. The AADT coefficient is significant and has a positive value (0.5719), which means that it is positively correlated to crashes. The horizontal alignment variables are not significant in the models but are shown in the table for consistency with presentations of the results from other models. The land use variable 'rural' presents an important correlation to road crashes. This result is different from what was obtained with the total crash frequency model, in which the number of crashes decreased in rural areas. However, this finding is not unexpected, as vehicle speed is typically higher in rural areas. The alertness of drivers in these areas might often be lowered because of the monotony of driving in this environment compared with urban areas. This monotony could also contribute to an increase in major injury crash risk. The spatial model yields a more significant estimate for the horizontal alignment variable 'windy'. These findings are consistent with other studies that looked at the relationship between horizontal alignment and road crashes (Montella et al., 2008; Karlaftis and Golias, 2002; Ma et al., 2008). The land use variable 'urban and rural' also appears to be a significant parameter in the spatial model, with a coefficient of 0.3138 for the land use variable. These results indicate that in urban-rural areas the expected number of total crashes is higher than in urban areas (base case). These results are consistent with findings in the existing literature (Pulugurtha et al., 2013; Osama and Sayed, 2017). The intercept value -8.6655 with a 95% credible interval equal to (-13.2040, -3.8917), and the coefficient of AADT is = 0.5719 with a 95% credible interval equal to (0.0605, 1.0747). This indicates that the AADT proportion has a positive effect on major injury crashes.

Results of the spatial model for the minor injury crash scenario are presented in table 5.6. Some variables, including the horizontal alignment variable 'meandering', the alignment variable 'windy' and the land use variable 'urban', appear to be negatively significant in the spatial model. The coefficients for the horizontal alignment variables

'meandering', 'windy' and land use 'urban' are, respectively, -0.3631, -1.6153 and -0.1314. Furthermore, some other variables including AADT, access density and land use 'urban-rural' seem to be positively significant. The coefficients for AADT, access density and land use 'urban rural' were 0.8053, 0.0376 and 0.2635, respectively. The intercept value -10.6133 with a 95% credible interval equal to (-15.2417, -6.0421), and the coefficient of AADT is = 0.8053 with a 95% credible interval equal to (0.3203, 1.2917). This indicates that the AADT proportion has a positive effect on minor injury crashes.

Results of the spatial model for the PDO crashes scenario are presented in table 5.7. The AADT coefficient, land use 'urban and rural' and the access density variables are significant in the model. The values for AADT, access density and land use 'urban and rural' are 1.1580, 0.0732 and 0.5612, respectively. Conversely, the other variables, including alignment horizontal 'windy', 'meandering' and land use 'urban', are not significant in the model. The intercept value is -13.5074 with a 95% credible interval equal to (-17.6916, -9.6152), and the coefficient of AADT is = 1.1580 with a 95% credible interval equal to (0.7264, 1.5961). This indicates that the AADT proportion has a positive effect on PDO crashes.

Table 5-3: Results of the spatial model of the total number of crashes scenario for the case study

Variable	Spatial Model				
	Quantile				
	Mean	SD	0.025	0.5	0.975
(Intercept)	-12.0807	2.0896	-16.2511	-12.0572	-8.0372
log(AADT)	1.1195	0.2194	0.6945	1.1174	1.5560
Access density	0.0469	0.0490	-0.0469	0.0471	0.1426
Horizontal alignment (Meandering)	-0.2807	0.2071	-0.6878	-0.2806	0.1258
Horizontal alignment (Windy)	-0.8841	0.4095	-1.6895	-0.8838	-0.0814
Land use (Urban and rural)	0.2826	0.3070	-0.3231	0.2835	0.8828
Land use (Urban)	0.0752	0.2309	-0.3767	0.0746	0.5301
τ_v	0.3122				
τ_u	3.6490				
Expected number of effect parameters	460.61				
DIC	2,620.55				

Table 5-4: Results of the spatial model of the fatal crash frequency scenario for the case study

Variable	Spatial Model				
	Mean	SD	Quantile		
			0.025	0.5	0.975
(Intercept)	-9.4081	2.5527	-14.4242	-9.4068	-4.4037
log(AADT)	0.6352	0.2706	0.1038	0.6352	1.1662
Access density	-0.0246	0.0516	-0.1266	-0.0243	0.0760
Horizontal alignment (Meandering)	-0.1798	0.1857	-0.5439	-0.1800	0.1850
Horizontal alignment (Windy)	-0.3948	0.4304	-1.2444	-0.3933	0.4456
Land use (Urban and rural)	0.3138	0.3068	-0.2922	0.3150	0.9123
Land use (Urban)	-0.1542	0.2224	-0.5902	-0.1544	0.2827
τ_v	0.4211				
τ_u	45.3877				
Expected number of effect parameters	316.10				
DIC	1,599.89				

Table 5-5: Results of the spatial model of major injury crash frequency scenario for the case study

Variable	Spatial Model				
	Mean	SD	Quantile		
			0.025	0.5	0.975
(Intercept)	-8.6655	2.4224	-13.2040	-8.6734	-3.8917
log(AADT)	0.5719	0.2582	0.0605	0.5734	1.0747
Access density	0.0204	0.0506	-0.0798	0.0206	0.0192
Horizontal alignment (Meandering)	-0.4088	0.2008	-0.8856	-0.4880	-0.0963
Horizontal alignment (Windy)	-0.8052	0.2352	-1.6752	-0.8030	-0.7986
Land use (Urban and rural)	-0.1514	0.3271	-0.8006	-0.1493	0.4858
Land use (Urban)	-0.2480	0.2352	-0.7048	-0.2498	0.2187
τ_v	0.6786				
τ_u	1.4678				
Expected number of effect parameters	314.74				
DIC	1,552.13				

Table 5-6: Results of the spatial model of the minor crash frequency scenario for the case study

Spatial Model					
Variable	Mean	SD	Quantile		
			0.025	0.5	0.975
(Intercept)	-10.6133	2.3403	-15.2417	-10.6038	-6.0421
log(AADT)	0.8053	0.2472	0.3203	0.8049	1.2917
Access density	0.0376	0.0544	-0.0700	0.0379	0.1437
Horizontal alignment (Meandering)	-0.3631	0.2102	-0.7781	-0.3625	0.0479
Horizontal alignment (Windy)	-1.6153	0.4999	-2.6107	-1.6107	-0.6465
Land use (Urban and rural)	0.2635	0.3225	-0.3746	0.2640	0.8924
Land use (Urban)	-0.1314	0.2445	-0.6099	-0.1321	0.3503
τ_v	0.4334				
τ_u	1.9012				
Expected number of effect parameters	354.30				
DIC	18.0877				

Table 5-7: Results of the spatial model of the PDO of crash frequency scenario for the case study

Spatial Model					
Variable	Mean	SD	Quantile		
			0.025	0.5	0.975
(Intercept)	-13.5074	2.1054	-17.6916	-13.4914	-9.6152
log(AADT)	1.1580	0.2214	0.7264	1.1567	1.5961
Access density	0.0732	0.0489	-0.0234	0.0734	0.1688
Horizontal alignment (Meandering)	-0.2096	0.2026	-0.6079	-0.2095	0.1879
Horizontal alignment (Windy)	-0.7414	0.4176	-1.5644	-0.7404	0.0757
Land use (Urban and rural)	0.5612	0.3025	-0.0354	0.5620	1.1528
Land use (Urban)	-0.0711	0.2290	-0.5194	-0.0717	0.3799
τ_v	0.3873				
τ_u	2.9860				
Expected number of effect parameters	404.89				
DIC	21.2996				

The standard deviation (SD) parameter presents some important reductions. In the total number of crashes scenario, the SD value is lower than in the fatal crash scenario, going from 2.5527 to 2.0896. The SD of the models associated with the major injury crash scenario (2.4224) is higher than the SD value of the minor injury crash scenario (2.3403) and of the PDO scenario (2.1054).

To assess the goodness-of-fit, the DIC was calculated and is presented in the previous tables. To assess the accuracy of the DIC, three other goodness-of-fit measures are also calculated in the models: the deviance, the WAIC and LPML. Table 5.8 presents a summary of goodness-of-fit measures for the different scenarios.

The results for all goodness-of-fit measures are consistent. Compared to the other scenarios, the major injury crashes scenario fits the best model, as indicated by lower DIC, deviance and WAIC values and a higher LPML value for the spatial models. Amongst all scenarios, the major injury crashes scenario predicted crashes most accurately, with a difference in DIC values of more than 59 points among all the models. Similarly, the deviance and WAIC values are lower in the spatial model of major injury crashes (2620.11 and 2584.86, respectively). The higher LPML value for the major injury model (-4536.82), as compared with the other models, also indicates that the major injury model performed better in this scenario.

Table 5-8: Summary of goodness-of-fit measures for the spatial models for the case study

	DIC	Deviance	WAIC	LPML
Total crashes	2,620.118	1,021.47	2,564.866	-4,536.827
Fatal crashes	1,599.343	771.62	1,581.086	-4,168.561
Major injury crashes	1,551.610	744.18	1,513.704	-4,940.098
Minor injury crashes	1,808.264	791.85	1,759.678	4,972.685
PDO crashes	2,130.226	914.07	2,080.768	-4,803.846

5.4.3. Distribution maps

For each scenario, a distribution map was produced for the parameters that were found to be significant in the results of the spatial model. All the significant parameters are presented on individual maps. For clarity of representation, the road crashes are not represented on these maps. Detailed maps showing the distribution of each contributing

parameter, along with combined maps, are in annex 5, including maps both with and without road crashes to allow for better comparison.

In the total crash scenario, the land use 'urban', 'urban and rural', and access density variables were found to be the most significant parameters (figures 5.4 and 5.5). This means that accident hot spots are mainly concentrated in 'urban' and 'urban and rural' areas and in areas with many intersections.

In the fatal crash scenario, the variables land use 'urban', 'urban and rural', and AADT were the most significant parameters (figures 5.4 and 5.6). This indicates that fatal accident hot spots are mainly concentrated in 'urban' and 'urban and rural' areas and in areas with substantial traffic.

In the major injury crash scenario, the variables AADT and access density were the most significant parameters (figures 5.5 and 5.6). This suggests that concentrations of crashes with major injuries are most likely to happen in areas of high daily traffic and on road segments with many intersections.

In the minor injury crash scenario, the variables AADT, access density and land use 'urban and rural' were the most significant parameters (figures 5.5, 5.6 and 5.7). This signifies that concentrations of crashes with major injuries are most likely to happen in areas of high daily traffic, on road segments with many intersections, and in suburban areas.

In the PDO crash scenario, the variables AADT and access density were the most significant parameters (figures 5.5 and 5.6). This means that concentrations of crashes with major injuries are most likely to happen in areas of substantial daily traffic and on road segments with many intersections.

Figure 5-4: Map representing the distribution of the parameters land use 'urban' and land 'urban-rural', both found significant in two scenarios (total crash and fatal injury)

Figure 5-5: Map representing the distribution of the parameter access density, found significant in four scenarios (total crash, major injury, minor injury and PDO)

Figure 5-6: Map representing the distribution of the parameter AADT, found significant in four scenarios (fatal injury, major injury, minor injury and PDO)

Figure 5-7: Map representing the distribution of the parameter land use 'urban-rural', found significant in the minor injury scenario

5.5. Comparison of the methods

5.5.1. Comparison of results

NetKDE was used to identify hot spots on the two main highways of the province. The majority of hot spots were identified on the east road. In particular, three main areas were identified with high concentrations of hot spots. The first area is in Banda Aceh, where the highway starts. The second area is a long section that spreads from Bireun to Lhoksemawe and Aceh Utara. The district of Lhoksemawe in particular was identified with the greatest number of hot spots. The third area with a high concentration of hot spots was in the districts of Aceh Timur and Aceh Tamiang, where a small number of hot spots were identified in close proximity to each other. These districts also tended to have higher population densities. Few hot spots were identified on the west road. Districts on the west road record lower population densities than those on the east road.

The hot zones methodology was used to identify hazardous road locations that consisted of two or more contiguous segments. The location of hot zones was investigated because in some instances contiguous road segments may be more dangerous than a hot spot alone. The methodology did not identify any hot zones on the west road. On the east road, however, 99 hot spots were identified. The districts of Aceh Utara and Lhoksemawe together recorded approximately 70% of the total number of hot zones identified on the east road. Twelve out of the 20 top hot zones by number of fatalities were located in the district of Lhokseumawe.

Figure 5.8 presents a comparison of the locations of hot spots (represented in red) and hot zones (represented in orange). It can be seen that high concentrations of hot spots were located in hot zones, and only two hot zones did not capture any of the hot spots identified with NetKDE. All other hot zones included at least one hot spot. A concentration of dangerous hot spots was identified in the district Lhoksemawe. In this area, the longest hot zones were identified. All the top 20 ranks of hot zones identified included some hot spots.

Figure 5-8: Comparison of the locations of hot spots and hot zones identified with different methodologies

The Bayesian spatial approach was used to identify the main factors contributing to the accidents recorded in the study data. For this, five scenarios (total number of accidents, fatal accidents, major injuries, minor injuries and PDO) were developed and analysed. The model also identified the main variables that were considered to be significant in the different scenarios. For example, AADT was found to be significant in four scenarios (fatal accidents, major injuries, minor injuries and PDO). The access density variable was also found to be a significant variable in four scenarios (total accidents, major injuries, minor injuries and PDO). The variable 'urban and rural' land use was significant in three scenarios.

In a comparison of the locations of hot zones with the locations of hot spots and the associated contributing factors identified with the Bayesian model, some patterns emerged. In particular, the segments where AADT was highest overlapped with the hot zones. A similar pattern of hot zones with the locations of hot spots and the associated contributing factors identified with the Bayesian model was observed in the major injury scenario. Access density was also identified as a significant parameter in the major injury scenario. Segments where access density was greatest also overlapped with the hot zones located in Lhokseumawe district and to a lesser extent in Banda Aceh. Some segments recorded highest access density on the west road in Aceh Barat Daya district; however, no hot zone was identified in this area. In the minor injury scenario, AADT and access density, identified as contributing factors, overlapped with hot zones in the Lhoksemawe district (AADT and access density) and in Aceh Timur (access density). Both

variables also overlapped at the beginning of the hot zone located in the district of Banda Aceh.

5.5.2. Discussion

All three methodologies provided insightful results which contributed to a better understanding of the patterns of road crashes on the two main highways of the province. The NetKDE methodology identified locations of road accident hot spots with high and low crashes along the roadway. The hot zone methodology helped to move a step beyond hot spots by identifying hazardous stretches of road that extended over continuous segments. In a context where decision-makers often have a limited budget to invest in road safety, the hot zone methodology is useful for detecting road locations with concentrations of accidents and those considered most dangerous. While the hot spot methodology detected a substantial number of accident hot spots, the hot zone methodology enables decision-makers to select and focus on those areas considered most dangerous. Concentrations of hot spots identified with the NetKDE method were identified as hot zones in the hot zone methodology. This included, for example, the mountainous areas in Pidie. In these areas, the limited visibility caused by the abundant vegetation and by the road design (many sharp curves), the crossing of wildlife, the excessive speed encouraged by the light traffic and the lack of speed control are conditions that generate many serious road crashes. In addition to identifying hot zones, several variables were identified as significant in the hot zone methodology. These included, for example, 'urban' and 'urban and rural' land uses and horizontal alignment features such as curves, intersections or access roads.

Mapping where road crashes occur is important, but limiting the number of crashes on these roads will not be possible without understanding what factors contribute to these crashes. By weighting the different variables, the Bayesian model provides some valuable information for decision-makers trying to select the most appropriate countermeasures. While the hot zone methodology identified some variables as significant, the Bayesian model provides more precision and in-depth information on the factors contributing to road crashes. For example, the hot zones overlap with some hot spots identified by the Bayesian model (and by NetKDE). In those hot spots located within hot zones, AADT and access density were found to be significant variables in the fatal and major injury crash scenarios. Indeed, in areas of dense traffic, more accidents are likely to occur. For example, drivers often adopt risky behaviours when passing in front of other vehicles on very busy roads and might not control other lanes carefully enough or check for nearby access roads and vehicles entering the highway. In high-traffic situations, drivers might also attempt to pass a vehicle in front on either side of that vehicle, an action more likely to cause a major accident.

5.6. Proposed combined approach

As stated in the introduction, the main objective of this thesis is to gain a better understanding of the relationships between traffic accidents and road design on the primary roads of the province of Aceh. Although data are continuously collected by traffic agencies in the province, these are not systematically analysed, even less so spatially. As a result, there is a lack of information on where road crashes are occurring and what factors are contributing to these accidents.

The first methodology used in the previous chapter identified and mapped road crash hot spots on the two main highways of the province. The results provide information that officials can investigate further. However, hot spots are point locations and do not account for the dangerousness of the surrounding road segments. This is because the hot spot methodology considers road segments as spatially independent units. Furthermore, with this method, the factors contributing to road accidents are not identified.

The hot zone methodology, by contrast, was used to identify hot zone locations that consisted of two or more contiguous segments. The identification of hot zones as a complement to hot spots is important because some contiguous road segments may be more dangerous than a hot spot alone. In total, 99 hot zones were identified on the east road. The districts of Lhokseumawe and Aceh Utara recorded the highest number of hot zones. Not hot zone was identified on the west road.

The third methodology was used to identify factors contributing to road crashes. In particular, the model allowed the identification of factors based on severity level. Knowing where road crashes occur is important, but understanding which factors contribute most to these crashes is a requirement for developing the most appropriate, tailored and specific countermeasures and thereby improving traffic safety.

Based on the objectives of the thesis, based on the results of the three methodologies, and based on the data gaps in Indonesia as well as in other South-East Asian countries, a combination of the hot zones method and the Bayesian model is proposed. One of the main benefits of the Bayesian approach over traditional methodologies, is that it allows the investigation of various characteristics related to the estimation of unmeasured variables. In particular, it can deal with small population samples and/or a lack of data and still provide relevant and informative results. For example, the Bayesian approach allowed for identification of the variability of estimates in the present case study in the province of Aceh, which is characterized by little data on road crashes. The Bayesian approach can also produce reliable estimates of safety using only one years' worth of accident data, provided the population size is sufficiently large. This has the advantage that only 'fresh' data, untarnished by the effects of external influences, is used in

estimating safety at a location. This situation of poor data availability is very common in South-East Asian countries. Working with a methodology that is able to provide new knowledge about the distribution of crashes, even in contexts of lower data quality and availability, is essential for taking appropriate actions to alter crash patterns. A second important aspect of this methodology is that, unlike traditional methodologies which focus on identifying a single optimum estimate, sample-based methods of Bayesian estimation estimate the entire density of a parameter.

This analysis therefore identified the Bayesian approach as suitable for investigating local factors contributing to road crashes. The advantage of the hot zone approach, as previously mentioned, is that it considers the spatial dependency of road segments. Thus, it enables the identification of hazardous road locations that encompass contiguous segments. This is a main benefit over the hot spots approach in that contiguous road segments may be more dangerous than a hot spot alone. This approach, however, does not identify the factors contributing to these hazardous road sections. A manual identification of factors that could explain the dangerousness of the top 20 hot zones in the case study returned the same factors identified with the Bayesian approach.

Combining these two approaches, which have different but complementary purposes, will not only identify the most hazardous sections of a road but will also identify the factors contributing to road crashes in these locations and weight them by importance so that decision-makers can quickly identify such factors for a given road. The approach is important in that it can be run at different scales, from small road sections to bigger networks, and can handle small data samples and lower-quality datasets, both of which can be common in Indonesia and in the broader South-East Asian region. Thus, this approach is suggested for researchers wanting to better understand road crashes and their patterns and characteristics, at a road level or across a district, and who are working with limited data and information. More precisely, this approach will enable a researcher to map accident locations in a studied area and analyse the factors contributing to these accidents. The results will provide important knowledge on where crashes occur and what factors contribute most to these crashes. With this information in hand, decision-makers can take action to put in place the most relevant countermeasures to reduce the number of accidents on a road or district network.

A practitioner's step-by-step guide to undertake an accident study using the combined approach is presented in figure 5.9. Crashes and road information are first collected, and a digital map, appropriate for the scale of the analysis, is chosen. The data collected are then cleaned, quality-checked and formatted into a spatial database (GIS).

Figure 5-9: A practitioner's step-by-step guide to undertake an accident study using the combined approach

Software such as ESRI's Arcmap or the open source software QGIS can be used to format the data. An investigation of the data is then undertaken to learn about the patterns, distributions, variables and characteristics of the variables. This information is mapped and compared to other relevant local information. This includes, for example, data on land use, population density or distribution over time. The aim is to gain better insight into the data.

To identify hazardous contiguous segments, the hot zone methodology is run on the dataset. The hot zone add-in developed at the University of Hong Kong is used to identify hot zones in the dataset. The full methodology and the tool are explained in Loo (2009). The GIS-based algorithm is detailed in the above flowchart. The first segment (or BSU) is analysed to see if its crash rate is greater than or equal to the threshold crash rate. If it is, a working table is created with an index number equal to one. All the contiguous BSUs are listed in the working table and analysed. The index number increases by one when the actual crash rate of a contiguous BSU is greater than or equal to the threshold crash rate. When all the BSUs have been checked, the index number is analysed. If it is greater than one, a hot zone is identified, and the hot zone variable of the BSU in the main table is updated to one. All the hot zones can then be mapped, and an analysis of the spatial pattern of hazardous contiguous road locations undertaken.

Once the hot zones have been identified and mapped, the use of the Bayesian model on the dataset enables the researcher to identify and weight factors contributing to road accidents. For this, QGIS is used to create a series of temporary tables that serve to create two main tables that will be used in the model. The R-INLA package is then used to run the models, under different scenarios (for example, according to different severity types), and will identify factors contributing most to road accidents.

So far, many methodologies only identify hot spots or hot zones but do not investigate closely or identify the explanatory factors behind these events. More importantly, by providing decision-makers the knowledge they need to develop countermeasures, the proposed methodology can be run at small or large scales and does not require a large amount of highly detailed information. This latter factor is especially important in Indonesia, as well as in other countries in the region that might have data collection, management and sharing issues.

5.7. Accident countermeasures

The two main highways in the province of Aceh represent an important part of the road network, and because these roads are the only way to drive from the south of the province to the north, they absorb a large part of the daily traffic. Overall, these roads account for an important fraction of all road crashes in the province, at all severity levels.

In this section, recommendations are made to develop a series of countermeasures to reduce the number of accidents on these highways. The countermeasures put forward are based on the results obtained from the three methodologies, as well as on countermeasures implemented in other countries which may be relevant in an Indonesian context. In the field of traffic safety, three main factors should be considered when developing appropriate countermeasures. Indeed, it is important to identify the road crash types occurring and the factors causing them and to understand the road situation (figure 5.10). Countermeasures involving reconstruction are considered too expensive measures and are excluded from the proposed measures.

Figure 5-10: Flow chart for the development and planning of road traffic accident countermeasures

The Bayesian spatial approach identified factors contributing significantly to road accidents of four severity types (fatalities, major injuries, minor injuries and PDO). The hot zone approach identified a series of hazardous road locations. Hot spots were located in areas identified as hot zones. These areas were characterised as curvy or meandering areas with many intersections and access roads. These results were consistent with those obtained with the Bayesian spatial model, which found that access density was a significant factor in most scenarios. Therefore, in urban environments, suggested countermeasures proposed on the highways of Aceh province include enhancing road users' awareness of upcoming intersections, sharp curves, or stops by adding additional signs before these events, and installing raised pavement markers at connection points to increase visibility and drivers' awareness. It is also suggested to reduce speed limits and add a higher number of speed cameras, especially in areas with many connection points to encourage drivers to reduce their speed and be more attentive.

Other measures, in rural environments, would include adding sufficient lights before, at and after intersections and sharp curves, as roads in rural environments tend to be poorly lighted. It is also suggested to add more sign ahead of these events to increase drivers' alertness. Finally, ensuring stronger enforcement of speed regulations is also recommended to decrease driver speed ahead of sharp turns or intersections. This could

also be done by adding fixed and mobile speed cameras. Finally, in both rural and urban areas, there is a need to increase road users' awareness of the use of seatbelt, which will contribute to reducing the severity of accidents in these areas.

In places with many pedestrians, motorcyclists or cyclists, it is suggested to segregate lanes for each road user type and to install multiple bumps over short distances to force reduced speeds, which will increase safety for all users. To increase drivers' awareness and alertness, it is strongly suggested to add additional road signs to inform drivers they are approaching these events. It is also recommended to lower speed limits, add speed controls, and improve enforcement to ensure that speed-related tickets are paid. Another suggestion is to develop public campaigns on road safety for road users and pedestrians. This could be done via various means such as highway signs, radio and schools. On busy urban roads crossing the city centre, it is encouraged to add secondary road traffic-density-related signs to encourage drivers to use secondary roads, which may be less busy for some sections of the travel route.

In rural areas with dense traffic – for example, in Aceh Besar – the installation of centreline rumble strips on undivided sections of highways is suggested to help prevent crashes in these areas. In addition, enhancing the presence of medical and emergency services in most rural areas with high crash rates is encouraged. Indeed, medical services tend to be less accessible in these areas, and reinforcing their presence and easing their access to areas where high crash rates have been identified could help save lives. Wildlife crossing in rural areas is also common. To reduce crashes related to wildlife crossing, it is suggested to install more frequent bridges for wildlife to use. Fencing in areas frequently crossed by wildlife and designing wildlife corridors to direct them towards wildlife crossing bridges could also reduce crashes related to wildlife crossing, which tend to be quite severe due to the higher speed of vehicles in rural areas.

Table 5.9 provides some examples of countermeasures that could be put in place to address factors identified as contributing most to road accidents.

Table 5-9: Examples of countermeasures to address road accident hot zones and mitigate contributing factors

Contributing factor targeted	Example of countermeasures	Type
AADT	<ul style="list-style-type: none"> • Install centreline rumble strips • Widen lanes or shoulders in dangerous areas of the road • Reduce speed limit on sections of the road with difficult roadway geometry over a substantial length • Enhance monitoring of seatbelt use and more strictly enforce related fines 	<p>Engineering</p> <p>Engineering</p> <p>Engineering</p> <p>Enforcement</p>
Access density	<ul style="list-style-type: none"> • Enhance road user awareness by installing more intersection warning signs • Install additional lighting before, in and after intersections 	<p>Education</p> <p>Engineering</p>
Horizontal alignment: Windy	<ul style="list-style-type: none"> • Verify super elevation and cross slope • Enhance road user awareness by installing more warning signs ahead of sharp curves • Enhance road user awareness by installing more warning signs ahead of stops • Install guardrails on longer sections of the road and as medians on highways • Enhance road user awareness by installing more warning signs ahead of horizontal curves 	<p>Engineering</p> <p>Engineering</p> <p>Engineering</p> <p>Engineering</p> <p>Education</p>
Land use: Urban-rural	<ul style="list-style-type: none"> • Remove unnecessary fixed objects to reduce the number of fixed-object crashes • Paint the bottom of trees on the side of the roads in white so that road users can see them more easily at night • Install guardrails on longer sections of the road and in the middle of highways 	<p>Engineering</p> <p>Engineering</p> <p>Engineering</p> <p>Engineering</p>

	<ul style="list-style-type: none"> • Mow and clear vegetation as required along relevant sections of the road • Ensure sufficient sight distance on dangerous sections of the road • Enhance road user awareness and lower speed limits ahead of road sections frequented by animals • Install more barrier reflectors, at smaller intervals, over dangerous road sections • Raise awareness of the main causes of road crashes (at schools and through communication campaigns) 	Engineering Engineering Engineering Education
Land use: Urban	<ul style="list-style-type: none"> • Widen lanes or shoulders along dangerous areas of the road • Create more clear zones and improve existing ones • Install more raised pavement markers on longer sections of the road • Enhance monitoring of helmet use by motorcycle riders and enforcement related fines more strictly • Increase public awareness of road accidents by undertaking campaigns advertising good practices on roads (signs on highways such as on the dangers of using cell phones while driving, etc.) • Require driving retesting and vision testing every few years for senior citizens 	Engineering Engineering Engineering Enforcement Education Enforcement

6. CONCLUSIONS AND FUTURE RESEARCH

6.1. Conclusions

The main objective of this thesis was to gain a better understanding of the relationships between traffic accidents and road design on the primary roads of the province of Aceh. With three persons dying on Indonesian roads every hour, road fatalities represent a major concern in the country. Yet, the road data and information collected by the national government, sub-national authorities and road traffic agencies, are not used in any systematic analyses to analyse road accident patterns and identify contributing factors. Nor are they used to inform better decision-making. The current gap in knowledge of the locations of crashes and the key factors contributing to them prevents authorities from developing effective countermeasures to, in the long term, reduce the number of road crashes and associated fatalities. In this thesis, some statistical techniques were applied to a provincial road safety database, characterised by some gaps in data availability and quality. Despite these limitations, the results of the quantitative spatial analyses enabled the generation of some new knowledge about patterns and characteristics of traffic accidents in Aceh.

At the provincial level, the findings of this thesis can help decision-makers in the province to develop appropriate measures to reduce accidents on the two main highways of the province. The thesis also offers a methodology that can be used by the main traffic agency in the province to monitor traffic crashes annually and adapt existing measures or develop new measures. The methodology proposed in this thesis can be applied to 'simple' datasets and still generates insightful results. As information about new parameters is collected, such as the socio-economic profile of road users, sight distance, or pedestrian crossings, the models can be expanded to include more factors.

At the national and broader regional levels, it is hoped that the findings of this thesis will encourage traffic agencies in other parts of Indonesia as well as in neighbouring countries, to reproduce these types of analyses to better understand road accident patterns and characteristics in other places.

The main conclusions are as follows:

1. While the Indonesian government collects and monitors some basic information primarily for global reporting purposes, there is still a lack of high quality and precise information collected on road crashes. There is a need to collect information that is more detailed – for example, socio-economic information about drivers. Importantly, there is a need to collect, in a systematic manner, this information on all road types in the country and not only on national roads. Today, there is a gap in knowledge as to the contributions of small provincial or village unpaved roads to road crashes. These roads are even more important in that they absorb the daily traffic of motorcyclists and cyclists, which in South-East Asia and Indonesia represent two of the categories most involved in fatal crashes.
2. Collecting and managing spatial information is costly, but in an era in which GIS and spatial methodologies are used daily to understand and monitor crash trends and patterns, there is a need in Indonesia, as well as in the province of Aceh, to establish GIS units within traffic departments and to move to spatial databases. Indeed, spatial information is essential to understanding how and why crashes occur and to monitoring the long-term effects of implemented countermeasures.
3. To improve data collection practices, establish GIS units and manage spatial databases, there is a need to develop open data and data sharing policies. Indeed, although road crash information is collected, it is difficult for researchers to access it. Researchers are key to analysing data that are collected and turning these into information and knowledge. Only with this information will the global community be able to make the most effective decisions to reduce the number of crashes and associated fatalities.
4. The results obtained with the NetKDE allowed the identification of hot spots on both highways. Results showed a higher concentration of hot spots on the east road and in urban areas. Fewer hot spots were identified on the west road, which is located mainly in rural areas. The results of the Bayesian model enabled the identification of the most important factors contributing to road accidents. Indeed, five scenarios, based on accident severity level, were analysed. Factors contributing to accidents were identified and weighted for each scenario. This revealed how significant a part they play in road accidents. The results of the hot zone methodology identified hot zones. These areas aligned with important concentrations of hot spots identified with NetKDE on the east road. Overall,

these three methodologies allowed the identification and mapping of traffic hot spots on the west road and the east road.

5. The results showed that there was a higher concentration of crashes in urban areas on the east road. This is because the east road is predominantly located in urban environments. In urban areas, traffic is increased, and speed limits are lower. There are also more pedestrians and motorcyclists, as well as more buildings such as supermarkets and schools. It was also discovered that the concentration of accidents was greater on road segments with many intersections and access roads. This likely resulted from poor signalisations and visibility ahead of intersections. The locations of hot zones also revealed that more accidents occur on road sections with major curves and that there is a need to improve visibility and signalisation ahead of these dangerous areas.
6. According to the Bayesian spatial models, AADT was found to be a significant variable in most scenarios. This means that the likelihood of crashes of the most severe types increases significantly in areas of high daily traffic. Secondly, the model identified that the risk of fatal accidents increased in winding and meandering urban-rural and rural environments. Overall, the Bayesian spatial model highlighted the importance of including spatial relationships in road crash models.
7. The hot zone method enabled the mapping of road hot zones, but it did not identify and map factors contributing to accidents in these identified areas. Understanding where hot zones are located and mapping them is not sufficient information for developing the most appropriate measures to lower the numbers of accidents in these areas. Identifying contributing factors, and the extent of their contribution to road accidents, is essential in order to implement measures to address these factors and improve traffic safety. This information is what the Bayesian spatial approach provides. Therefore, the combination of mapping road accident hot zones and identifying factors contributing most to accidents in these areas is an important step, in a country like Indonesia, to reducing the number of accidents. While the hot spots methodology appeared useful for identifying concentrations of accidents and was also well-suited for segment-level analyses, it does not consider the spatial connection between the segments, and it identifies point locations. In the context of limited budgets, identifying hot zones produces results useful enough to inform decision-making that can quickly put in place appropriate measures to reduce the level of danger in some of these identified areas.
8. While the hot zone methodology has thus far mostly been developed and run on larger networks (for example, at the neighbourhood level), the present analysis shows that it can also produce informative results on small road sections and can quickly identify priority areas.

6.2. Future research

Based on the findings of this thesis, future research could focus on different aspects.

These analyses could be extended to all roads in the province (including secondary and unpaved roads) to improve understanding of what variables contribute most significantly to accidents on these roads. Indeed, there is currently a complete lack of knowledge as to where and why road crashes occur on these other roads.

The present analyses could be complemented by adding more social variables focused on profiles of road users. Understanding how variables such as age, profession, education and gender affect the distribution of road crashes will allow the development of more appropriate countermeasures. The present thesis did not include variables related to the socio-economic profiles of road users, because these are currently not collected by local traffic agencies. Collecting this information would constitute a first step to expanding these analyses with more social variables.

It is also suggested that the effects of countermeasures should be monitored and evaluated on a mid- and long-term basis, to assess how they have helped to reduce the number of crashes. For example, the characteristics of road safety should receive administrative attention. This would include strengthening the research into and implementation of counteractive measures in terms of road geometry or pavement surface or developing proper road safety training. Future implementation efforts should not skim the surface by providing redundant greening and landscapes in developed areas but should increase the investment in road safety research in disaster-prone environments.

Future work could include focusing on traffic crashes at intersections under different settings, including, for example, the characteristics of crashes occurring on curves at the end of straight roads. This will help better inform road infrastructure design to address road safety problems before they occur. Secondly, investigating the long-term effects of both hot spot and hot zone countermeasures and understanding their individual and joint impact on road safety (when both measures are taken on the same segment versus taken on different segments) could advance understanding of how to best develop certain types of measures in given contexts. Third, future work could investigate the potential for developing a single-approach methodology to understand how important multivariate models are, compared with spatial correlation. This would involve observing and analysing more than one statistical outcome variable at a time. Fourth, investigating the use of random parameters would be helpful to facilitate and enhance understanding on how crash factors affect crash frequency and severity along a highway segment. Lastly, it could be useful to investigate the potential of using joint Bayesian to model crash prediction, including both road segments and intersections located on a

road network. This would give insights into the spatial correlations between heterogeneous types of entities.

7. REFERENCES

Abdel-Aty, M.A., & Radwan, A.E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5), 633-642.

Aguero-Valverde, J., & Jovanis, P. (2008). Analysis of road crash frequency with spatial models. *Transportation Research Record: Journal of the Transportation Research Board*, (2061), 55-63.

Aguero-Valverde, J., & Jovanis, P. P. (2010). Spatial correlation in multilevel crash frequency models: Effects of different neighboring structures. *Transportation Research Record*, 2165(1), 21-32

Aguero-Valverde, J., Wu, KK., and Donnell, ET. (2016) A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention* 87: 8-16.

Anastasopoulos, P.C., Tarko, A.P., & Mannering, F.L. (2008). Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis & Prevention*, 40(2), 768-775.

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., & Mannering, F.L. (2012a). A study of factors affecting highway accident rates using the random-parameters tobit model. *Accident Analysis & Prevention*, 45, 628-633.

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., & Mannering, F.L. (2012b). A multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis & Prevention*, 45, 110-119.

Anderson, T.K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3), 359-364.

Aghajani, M. A., Dezfoulan, R. S., Arjroody, A. R., & Rezaei, M. (2017). Applying GIS to identify the spatial and temporal patterns of road accidents using spatial statistics (case study: Ilam Province, Iran). *Transportation research procedia*, 25, 2126-2138.

Banos, A., & Huguenin-Richard, F. (2000). Spatial distribution of road accidents in the vicinity of point sources: application to child pedestrian accidents. *Geography and Medicine*, 8, 54-64.

Barua, S., El-Basyouny, K., & Islam, M. T. (2014). A full Bayesian multivariate count data model of collision severity with spatial correlation. *Analytic Methods in Accident Research*, 3, 28-43.

Barua, S., El-Basyouny, K., & Islam, M. T. (2016). Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic methods in accident research*, 9, 1-15.

Benedek, J., Ciobanu, S. M., & Man, T. C. (2016). Hotspots and social background of urban traffic crashes: A case study in Cluj-Napoca (Romania). *Accident Analysis & Prevention*, 87, 117-126.

Berg, A., Meyer, R., & Yu, J. (2004). DIC as a model comparison criterion for stochastic volatility models. *Journal of Business and Economic Statistics*, 22, 107-120.

Besag J, York J and Mollié A (1991) A Bayesian image restoration with two applications in spatial statistics *Ann Inst Statist Math* 43: 1–59. Find this article online.

Black, W. R., & Thomas, I. (1998). Accidents on Belgium's motorways: a network autocorrelation analysis. *Journal of Transport Geography*, 6(1), 23-31.

Blangiardo M, Cameletti M, Baio G and Rue H (2013) Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology* 4: 33-49.

Blazquez, C.A., & Celis, M.S. (2013). A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile. *Accident Analysis & Prevention*, 50, 304-311.

Blazquez, C. A., Picarte, B., Calderón, J. F., & Losada, F. (2018). Spatial autocorrelation analysis of cargo trucks on highway crashes in Chile. *Accident Analysis & Prevention*, 120, 195-210.

Borruso, G. (2005). Network density estimation: analysis of point patterns over a network. In *International Conference on Computational Science and Its Applications* (pp. 126-132). Springer, Berlin, Heidelberg.

- Bruyninckx, H. (2002). Bayesian probability. Belgium: K.U. Leuven, Department of Mechanical Engineering.
- Cai, Q., Abdel-Aty, M., Lee, J., & Eluru, N. (2017). Comparative analysis of zonal systems for macro-level crash modeling. *Journal of safety research*, 61, 157-166.
- Caliendo, C., & Guida, M. (2014). A new bivariate regression model for the simultaneous analysis of total and severe crashes occurrence. *Journal of Transportation Safety & Security*, 6(1), 78-92.
- Caliendo, C., de Guglielmo, M.L., & Guida, M. (2015). Comparison and analysis of road tunnel traffic accident frequencies and rates using random-parameter models. *Journal of Transportation Safety and Security*, 8(2), 00-00.
- Caliendo, C., De Guglielmo, M. L., & Guida, M. (2016). Comparison and analysis of road tunnel traffic accident frequencies and rates using random-parameter models. *Journal of Transportation Safety & Security*, 8(2), 177-195.
- Carson, J., & Mannering, F. (2001). The effect of ice warning signs on ice-accident frequencies and severities. *Accident Analysis & Prevention*, 33(1), 99-109.
- Çela, L., Shiode, S., & Lipovac, K. (2013). Integrating GIS and spatial analytical techniques in an analysis of road traffic accidents in Serbia. *International Journal for Traffic and Transport Engineering*, 3(1), 1-15.
- Chang, S., & Dixit, V. V. (2018). Application of Fractal theory for crash rate prediction: Insights from random parameters and latent class tobit models. *Accident Analysis & Prevention*, 112, 30-38.
- Chen, J. R., Zhang, M. G., Yu, S. J., & Wang, J. (2018). A Bayesian Network for the Transportation Accidents of Hazardous Materials Handling Time Assessment. *Procedia Engineering*, 211, 63-69.
- Cheng, E. and A. Tarko (2014). Modeling Safety of Highway Work Zones with Random Parameters and Random Effects Models. *Analytic Methods in Accident Research*. 1(1), 86-95.
- Chin, H. C., & Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention*, 35(2), 253-259.
- Conway, R.W., & Maxwell, W.L. (1962). Network dispatching by the shortest-operation discipline. *Operations research*, 10(1), 51-73.

Cressie, N. A. C., (1993). *Statistics for Spatial Data*, revised edition. John Wiley & Sons, New York, p.359

Da Cruz Figueira, A., Pitombo, C.S., & Larocca, A.P.C. (2017). Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil. *Case Studies on Transport Policy*, 5(2), 200-207.

Dadashova, B., Arenas, B., Mira, J., & Aparicio, F. (2014). Bayesian model selection of structural explanatory models: Application to road accident data. *Procedia-Social and Behavioral Sciences*, 160, 55-63.

Debnath, A.K., Blackman, R., & Haworth, N. (2014). A Tobit model for analysing speed limit compliance in work zones. *Safety Science*, 70, 367-377.

Dechter, R., & Pearl, J. (1988). Network-based heuristics for constraint-satisfaction problems. In *Search in Artificial Intelligence* (pp. 370-425). Springer, New York, NY.

Dong, C., Clarke, D. B., Nambisan, S. S., & Huang, B. (2016). Analyzing injury crashes using random-parameter bivariate regression models. *Transportmetrica A: Transport Science*, 12(9), 794-810.

Druzdel, M. J., & Van Der Gaag, L. C. (2000). Building probabilistic networks: "Where do the numbers come from?". *IEEE Transactions on knowledge and data engineering*, 12(4), 481-486.

Eckley, D. and Curtin, K. (2013). Evaluating the spatiotemporal clustering of traffic accidents. *Computers, Environment and Urban Systems*. Volume 37: pp70-81.

Effati, M., Rajabi, M., Samadzadegan, F., & Shabani, S. (2014). A geospatial neuro-fuzzy approach for identification of hazardous zones in regional transportation corridors. *International Journal of Civil Engineering*, 12(3), 289-303.

El-Basyouny, K., & Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis & Prevention*, 41(4), 820-828.

El-Basyouny, K., & Sayed, T. (2013). Depth-based hotspot identification and multivariate ranking using the full Bayes approach. *Accident Analysis & Prevention*, 50, 1082-1089.

Elvik, R. (2008). The predictive validity of empirical Bayes estimates of road safety. *Accident Analysis & Prevention*, 40(6), 1964-1969.

- Elvik, R., Ulstein, H., Wifstad, K., Syrstad, R. S., Seeberg, A. R., Gulbrandsen, M. U., & Welde, M. (2017). An Empirical Bayes before-after evaluation of road safety effects of a new motorway in Norway. *Accident Analysis & Prevention*, 108, 285-296.
- Erdogan, S., Yilmaz, I., Baybura, T., & Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. *Accident Analysis & Prevention*, 40(1), 174-181.
- Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research*, 40(5), 341-351.
- Erdogan, S., Ilci, V., Soysal, O.M., Kormaz, A. (2015). A model suggestion for the determination of the traffic accident hotspots on the Turkish highway road network: a pilot study. *Boletim de Ciencias Geodésicas*, 21(1).
- ESCAP, U. (2010). Energy Security and Sustainable Development in Asia and the Pacific, Policy Options for Energy Security and Sustainable development.
- Flahaut, B., Mouchart, M., San Martin, E., & Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. *Accident Analysis & Prevention*, 35(6), 991-1004.
- Flahaut, B. (2004). Impact of infrastructure and local environment on road unsafety: Logistic modeling with spatial autocorrelation. *Accident Analysis & Prevention*, 36(6), 1055-1066.
- Famili, A., Sarasua, W. A., Ogle, J. H., Shams, A., & Asadi, N. (2018). GIS Based Spatial Analysis of Pedestrian Crashes: A Case Study of South Carolina. In *International Conference on Transportation and Development* (p. 368).
- Farah, H., Bekhor, S., & Polus, A. (2009). Risk evaluation by modeling of passing behavior on two-lane rural highways. *Accident Analysis & Prevention*, 41(4), 887-894.
- Fountas, G., & Anastasopoulos, P. C. (2017). A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. *Analytic methods in accident research*, 15, 1-16.
- Fountas, G., & Anastasopoulos, P. C. (2018). Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances. *Analytic methods in accident research*, 20, 30-45.

Fu, Y., Li, J., Weng, Q., Zheng, Q., Li, L., Dai, S., & Guo, B. (2019). Characterizing the spatial pattern of annual urban growth by using time series Landsat imagery. *Science of The Total Environment*, 666, 274-284.

Garber, N. J., & Ehrhart, A. A. (2000). Effect of speed, flow, and geometric characteristics on crash frequency for two-lane highways. *Transportation Research Record*, 1717(1), 76-83.

Gaber, M., Wahaballa, A. M., Othman, A. M., & Diab, A. (2017). Traffic accidents prediction model using fuzzy logic: Aswan desert road case study. *J. Eng. Sci. Assiut Univ*, 45, 2844.

Geedipally, S. R., Lord, D., & Dhavala, S. S. (2014). A caution about using deviance information criterion while modeling traffic crashes. *Safety science*, 62, 495-498.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Getis, A., & Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3), 189-206.

Goldin, P. R., & Gross, J. J. (2010). Effects of mindfulness-based stress reduction (MBSR) on emotion regulation in social anxiety disorder. *Emotion*, 10(1), 83.

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention*, 35(2), 273-285.

Grossman, D., Domingos, P., & Domingos, P. (2004, July). Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proceedings of the twenty-first international conference on Machine learning* (p. 46). ACM.

Grubenmann, T., Tsoi, K.H. and Loo, B.P.Y. (2019a) Hotzone Generation Add-in for ArcGIS. Available at <http://arcg.is/1T0bKv>

Guadamuz-Flores, R., & Aguero-Valverde, J. (2017). Bayesian Spatial Models of Crash Frequency at Highway-Railway Crossings. *Transportation Research Record Journal of the Transportation Research Board*, 2608, 27-35.

Gundogdu, I.B. (2010). Applying linear analysis methods to GIS-supported procedures for preventing traffic accidents: Case study of Konya. *Safety Science*, 48(6), 763-769.

- Gundogdu, I.B. (2014). Geostatistical Analyst using the Junction Risk Factor to Analyse and Prevent Urban Traffic Accidents. *Latest Trends in Energy, Environment and Development*, 42-49.
- Guo Q., Xu P., Pei X., Wong SC. and Yao D. (2017) The effect of road network patterns on pedestrian safety: a zone-based Bayesian spatial modelling approach. *Accident Analysis & Prevention* 99, 114-124.
- Harirforoush, H., & Bellalite, L. (2016). A new integrated GIS-based analysis to detect hotspots: a case study of the city of Sherbrooke. *Accident Analysis & Prevention* 130, 62-74.
- Hashimoto, S., Yoshiki, S., Saeki, R., Mimura, Y., Ando, R., & Nanba, S. (2016). Development and application of traffic accident density estimation models using kernel density estimation. *Journal of traffic and transportation engineering (English edition)*, 3(3), 262-270.
- Hauer, E. 1997. *Observational before-after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety*. Oxford: Pergamon/Elsevier Science Ltd.
- Heckerman, D., & Breese, J. S. (1996). Causal independence for probability assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 26(6), 826-831.
- Hezaveh, A. M., & Cherry, C. R. (2018). Neighborhood-level factors affecting seat belt use. *Accident Analysis & Prevention*, 122, 153-161.
- Hosseinpour, M., Sahebi, S., Zamzuri, Z. H., Yahaya, A. S., & Ismail, N. (2018). Predicting crash frequency for multi-vehicle collision types using multivariate Poisson-lognormal spatial model: A comparative analysis. *Accident Analysis & Prevention*, 118, 277-288.
- Huang, H., H. C. Chin, and M. M. Haque. (2009). "Empirical Evaluation of Alternative Approaches in Identifying Crash Hot Spots." *Transportation Research Record: Journal of the Transportation Research Board* 2103: 32–41. doi: 10.3141/2103-05.
- Huang H. and Abdel-Aty M. (2010) Multilevel data and bayesian analysis in traffic safety *Accident Analysis & Prevention*, 42, 1556-1565.
- Huang, H., Abdel-Aty, M., and Darwiche, A., (2010). County-level Crash Risk Analysis in Florida: Bayesian Spatial Modeling. *Transportation Research Record*, No. 2148, 27-37, *Transportation Research Board of the National Academics*, Washington, D.C.

- Huang H., Zhou H., Wang J., Chang F. and Ma M. (2017) A multivariate spatial model of crash frequency by transportation modes for urban intersections *Analytic Methods in Accident Research*, 14, 10-21
- Imprialou, M. I. M., Quddus, M., Pitfield, D. E., & Li, L. (2014). High-Accuracy Generic Method for Automatic Crash Mapping Using Fuzzy Logic (No. 14-3868).
- IFT, UN, Eurostat (2020). Glossary for Transport Statistics 2019 5th edition, OECD Publishing Paris/European Union, Brussels. <https://doi.org/10.1787/505b670b-en>.
- Ivan, I., & Tesla, J. (2015). Road and intersection accidents: localization of black spots in Ostrava. *Geografický časopis*, 67(4), 323-340.
- Jensen, F.V., 2001. Bayesian Networks and Decision Graphs. Springer, New York. NETICA. Bayesian network development software. Norsys Software Corp. www.norsys.com
- ARTICLE IN PRESS Fig. 7. Histogram of the sensitivity coefficient s_1 (absolute value of derivative).
- Karlaftis MG and Golias I (2002) Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident Analysis & Prevention* 34(3): 357-365.
- Kaygisiz, Ö., Düzgün, Ş., Yildiz, A., & Senbil, M. (2015). Spatio-temporal accident analysis for accident prevention in relation to behavioral factors in driving: The case of South Anatolian Motorway. *Transportation research part F: traffic psychology and behaviour*, 33, 128-140.
- Khalid, S., Shoaib, F., Qian, T., Rui, Y., Bari, A. I., Sajjad, M., ... & Wang, J. (2018). Network Constrained Spatio-Temporal Hotspot Mapping of Crimes in Faisalabad. *Applied Spatial Analysis and Policy*, 11(3), 599-622.
- Khan, G., Qin, X., & Noyce, D.A. (2008). Spatial analysis of weather crash patterns. *Journal of Transportation Engineering*, 134(5), 191-202.
- Kim, K., & Nitz, L. (1995). Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns. *Accidents Analysis & Prevention*, 27(5), 663-674.
- Kingham, S., Sabel, C.E., & Bartie, P. (2011). The impact of the 'school run' on road traffic accidents: A spatio-temporal analysis. *Journal of Transport Geography*, 19(4), 705-711.
- Kjaerulff, U. B., & Madsen, A. L. (2008). Bayesian networks and influence diagrams. Springer Science+ Business Media, 200, 114.

- Kumara, S., & Chin, H.C. (2003). Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. *Traffic Injury Prevention*, 4(1), 53-57.
- Lee, J., & Mannering, F. (2002). Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis & Prevention*, 34(2), 149-161.
- Li, L., Zhu, L., & Sui, D.Z. (2007). A GIS-based Bayesian approach for analysing spatial-temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography*, 15(4), 274-285.
- Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic information systems and science*. John Wiley & Sons.
- Loo, B. P. (2009). The identification of hazardous road locations: a comparison of the blacksite and hot zone methodologies in Hong Kong. *International Journal of Sustainable Transportation*, 3(3), 187-202.
- Loo, B.P.Y., Cheung, W., and Yao, S. (2011). The Rural-Urban Divide in Road Safety: The Case of China. *Open Transportation Journal*, 5(1), 9-20
- Loo B.P.Y and Yao S. (2013). The identification of traffic crash hot zones under the link-attribute and event-based approaches in a network-constrained environment. *Computers, Environment and Urban Systems*, 41, 249-261
- Lord, D., & Persaud, B. N. (2004). Estimating the safety performance of urban road transportation networks. *Accident Analysis & Prevention*, 36(4), 609-620.
- Lord, D., Washington, S.P., & Ivan, J.N. (2005a). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1), 35-46.
- Lord, D., Washington, S.P., & Ivan, J.N. (2005b). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1), 35-46.
- Lord, D., Washington, S., & Ivan, J.N. (2007a). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, 39(1), 53-57.
- Lord, D., Washington, S., & Ivan, J.N. (2007b). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, 39(1), 53-57.

Lord, D., & Miranda-Moreno, L.F. (2008). Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. *Safety Science*, 46(5), 751-770.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 291-305.

Lu, P., & Tolliver, D. (2016). Accident prediction model for public highway-rail grade crossings. *Accident Analysis & Prevention*, 90, 73-81.

Ma J, Kockelman KM and Damien P (2008) A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention* 40(3), 964-975.

Ma, X., Chen, F., Chen, S., (2015a). Modeling crash rates for a mountainous highway by using refined-scale panel data. *Transportation Research Record* 2515, 10-16

MacNab, Y. C. (2004). Bayesian spatial and ecological models for small-area accident and injury analysis. *Accident Analysis & Prevention*, 36(6), 1019-1028.

Manepalli, U., & Bham, G. (2013). Identification of Crash-Contributing Factors: Effects of Spatial Autocorrelation and Sample Data Size. *Transportation Research Record: Journal of the Transportation Research Board*, 2386, 179-188.

Manepalli, U., & Bham, G. (2016). An evaluation of performance measures for hotspot identification. *Journal of Transportation Safety & Security*, 8(4), 1-19.

Mannering, F.L., & Bhat, C.R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic methods in accident research*, 1 (2014), 1-22.

Martins, T., Simpson D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA : new features. *Computational Statistics & Data Analysis*, 67: 68-83. <http://www.r-inla.org/download>.

MathWorks, I. (1999). Optimization Toolbox, User's Guide. The MathWorks.

Meng, X.-H., Zheng, L. & Qin, G.-M. (2009). Traffic accidents prediction and prominent influencing factors analysis based on fuzzy logic. *Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/ Journal of Transportation Systems Engineering and Information Technology*. 9. 87-92.

- Mensah, A. and Hauer, E. (1998). Two Problems of Averaging Arising in the Estimation of the Relationship Between Accidents and Traffic Flow. *Transportation Research Record*. 1635. 37-43. 10.3141/1635-05.
- Miaou, S., & Lord D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes Versus Empirical Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*. 1840(1), 31-40.
- Miaou, S., Song, J.J., & Mallick, B.K. (2003). Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics*, 6, 33-58.
- Miranda-Moreno, L. F., Labbe, A., & Fu, L. (2007). Bayesian multiple testing procedures for hotspot identification. *Accident Analysis & Prevention*, 39(6), 1192-1201.
- Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian-vehicle collision occurrence at signalized intersections. *Accident Analysis & Prevention*, 43(5), 1624-1634.
- Mitra, S., & Washington, S. (2007). On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis & Prevention*, 39(3), 459-468.
- Mitra, S. (2009). Spatial autocorrelation and Bayesian spatial statistical method for analyzing intersections prone to injury crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2136, 92-100.
- Mohammadi, M. A., Samaranayake, V. A., & Bham, G. H. (2014). Crash frequency modeling using negative binomial models: An application of generalized estimating equation to longitudinal data. *Analytic Methods in Accident Research*, 2, 52-69.
- Mohamadzadeh Moghaddam, A. (2016, March). GIS based spatial analysis of pedestrian, Case study of Mashhad city. In *پانزدهمین کنفرانس بین المللی مهندسی حمل و نقل و ترافیک*.
- Montella, A., Colantuoni, L., and Lamberti, R. (2008). Crash prediction models for rural motorways. *Transportation Research Record: Journal of the Transportation Research Board* (2083), 180-189.
- Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis & Prevention*, 49, 58-72.
- Moons, E., Brijs, T., & Wets, G. (2009). Improving Moran's index to identify hot spots in traffic safety. In: *Geocomputation and urban planning*. Springer, 176, 117-132.

Moran, P.A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2), 243-251.

Mothafer, G. I., Yamamoto, T., & Shankar, V. N. (2017). A negative binomial crash sum model for time invariant heterogeneity in panel crash data: some insights. *Analytic methods in accident research*, 14, 1-9.

Murè, S., & Demichela, M. (2009). Fuzzy Application Procedure (FAP) for the risk assessment of occupational accidents. *Journal of loss prevention in the process industries*, 22(5), 593-599.

Myllymaki, P. (2005). Advantages of Bayesian networks in data mining and knowledge discovery. <http://www.bayesit.com/docs/advantages.html>.

Nair, B. M., & Cai, J. (2007, June). A fuzzy logic controller for isolated signalized intersection with traffic abnormality considered. In *2007 IEEE intelligent vehicles symposium* (pp. 1229-1233). IEEE.

Neapolitan, R. E. (1990). Probabilistic reasoning in expert systems. A Wiley-Interscience Publication.

Norsys (2000). Netica. <http://www.norsys.com>.

O'Sullivan, D., & Unwin, D. (2010). *Geographic information systems*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Oh, J., Washington, S. P., & Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention*, 38(2), 346-356.

Okabe, A., Satoh, T., & Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science*, 23(1), 7-32.

Okabe, A. and Sugihara, K. (2012). Spatial analysis along networks: statistical and computational methods. 10.1002/9781119967101.

Osama A and Sayed T (2017) Evaluating the Impact of Socioeconomics, Land Use, Built Environment, and Road Facility on Cyclist Safety. *Transportation Research Record: Journal of the Transportation Research Board* 2659, 33-42.

Ouni, F., & Belloumi, M. (2018). Spatio-temporal pattern of vulnerable road user's collisions hot spots and related risk factors for injury severity in Tunisia. *Transportation Research Part F: Traffic Psychology and Behaviour*, 56, 477-495.

- Ouni, F., and Belloumi, M. (2019). Pattern of road traffic crash hot zones versus probable hot zones in Tunisia: A geospatial analysis. *Accident Analysis and Prevention*, 128, 185-196.
- Pearl, J. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450), 428-431.
- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3, 96-146.
- Persaud, B. and Lyon C. (2007). Empirical Bayes before-after studies: lessons learned from two decades of experience and future directions. *Accident Analysis and Prevention*. 39(3), 546-55.
- Pirdavani, A., Bellemans, T., Brijs, T., & Wets, G. (2014). Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. *Journal of Transportation Engineering*, 140(8), 04014032.
- Plug, C., Xia, J.C., & Caulfield, C. (2011). Spatial and temporal visualisation techniques for crash analysis. *Accident Analysis & Prevention*, 43(6), 1937-1946.
- Pooley CM, Marion G. (2018). Bayesian model evidence as a practical alternative to deviance information criterion. *R. Soc. open sci.*5: 171519.
- Prasannakumar, V., Vijith, H., Charutha, R., & Geetha, N. (2011). Spatio-temporal clustering of road accidents: GIS based analysis and assessment. *Procedia-Social and Behavioral Sciences*, 21, 317-325.
- Pulugurtha, S.S., Krishnakumar, V.K., & Nambisan, S.S. (2007). New methods to identify and rank high pedestrian crash zones: An illustration. *Accident Analysis & Prevention*, 39(4), 800-811.
- Pulugurtha SS, Duddu VR and Kotagiri Y (2013) Traffic analysis zone level crash estimation models based on land use characteristics. *Accident Analysis & Prevention* 50: 678-687.
- QGIS Development Team (2019). QGIS Geographic Information System. Open Source Geospatial Foundation Project. <http://qgis.osgeo.org>
- 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.

Raihan, M. A., Alluri, P., Wu, W., & Gan, A. (2019). Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. *Accident Analysis & Prevention*, 123, 303-313.

Ramp, D., Caldwell, J., Edwards, K.A., Warton, D., Croft, D.B. (2005). Modelling of wildlife fatality hotspots along the Snowy Mountain Highway in New South Wales, Australia. *Biological Conservation* 126, 474-490.

Rankavat, S., & Tiwari, G. (2013). Pedestrian Accident Analysis in Delhi using GIS. *Journal of the Eastern Asia Society for Transportation Studies*, 10, 1446-1457.

Rue H, Martino S and Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)* 71(2), 319-392.

Sacchi, E., Sayed, T. & El-Basyouny, K. (2016). A full Bayes before-after study accounting for temporal and spatial effects: Evaluating the safety impact of new signal installations. *Accident Analysis & Prevention*. 94. 52-58. 10.1016/j.aap.2016.05.015.

Saha, D., Alluri, P., Gan, A., & Wu, W. (2018). Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. *Accident Analysis & Prevention*, 118, 166-177.

Shaheed, M. S., Gkritza, K., Carriquiry, A. L., & Hallmark, S. L. (2016). Analysis of occupant injury severity in winter weather crashes: A fully Bayesian multivariate approach. *Analytic methods in accident research*, 11, 33-47.

Samadzadegan, F., Azizi, A., Hahn, M., & Lucas, C. (2005). Automatic 3D object recognition and reconstruction based on neuro-fuzzy modelling. *ISPRS journal of photogrammetry and remote sensing*, 59(5), 255-277.

Saravanan, S., Sabari, A., & Geetha, M. (2014). Fuzzy-based approach to predict accident risk on road network. *International Journal of Emerging Technology and Advanced Engineering*, 4(5), 536-540.

Schneider, R.J., Ryznar, R.M., & Khattak, A.J. (2004). An accident waiting to happen: a spatial approach to proactive pedestrian planning. *Accident Analysis & Prevention*, 36(2), 193-211.

Selby, B. and Kockelman, K. (2013). Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, 29, 24-32.

- Shafabakhsh, G., Famili, A., Sadegh Bahadori, M. (2017). GIS-based spatial analysis of urban traffic accidents: case study in Mashad Iran. *Journal of Traffic and Transportation Engineering*, 4(3), 290-299.
- Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention*, 27(3), 371-389.
- Shankar, V., Ulfarsson, G.F., Pendyala, R.M., & Nebergall, M.B. (2003). Modelling crashes involving pedestrians and motorized traffic. *Safety Science*, 41(7), 627-640.
- Shuttleworth, M' (2009). Bayesian Probability. Retrieved Nov 25, 2019 from Explorable.com: <https://explorable.com/bayesian-probability>
- Sireli, Y., Ozan, E., Kauffmann, P., Gupta, S., & Kachroo, P. (2002). A QFD-Based Decision Model for Application of ITS Technologies to Cockpit Weather Information Systems for General Aviation. In 9th World Congress on Intelligent Transport SystemsITS America, ITS Japan, ERTICO (Intelligent Transport Systems and Services-Europe).
- Song JJ, Ghosh M, Miaou S and Mallick B (2006) Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, 97(1), 246-273.
- Spiegelhalter D, Best N, Carlin B, and Van Der Linde A (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583-639.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS user manual.
- Spiegelhalter D, Best N, Carlin B, and Van Der Linde A (2014) The Deviance Information Criteria: 12 years on. *J.R.Stat.Soc. B76*, 485-493
- Steenberghen, T., Dufays, T., Thomas, I., & Flahaut, B. (2004). Intra-urban location and clustering of road accidents using GIS: a Belgian example. *International Journal of Geographical Information Science*, 18(2), 169-181.
- Tayfur, G., Ozdemir, S., & Singh, V. P. (2003). Fuzzy logic algorithm for runoff-induced sediment transport from bare soil surfaces. *Advances in water resources*, 26(12), 1249-1256.
- Thakali, L., Kwon, T.J., & Fu, L. (2015). Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation*, 23(2), 93-106.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24-36.

Tortum, A., & Atalay, A. (2015). Spatial analysis of road mortality rates in Turkey. In: *Proceedings of the Institution of Civil Engineers-Transport*. 168 (4), 532-542.

Truong, L.T., & Somenahalli, S.V. (2011). Using GIS to identify pedestrian-vehicle crash hot spots and unsafe bus stops. *Journal of Public Transportation*, 14(1), 6.

Vandenbulcke, G., Thomas, I. & Int Panis, L. (2013). Predicting cycling accident risk in Brussels: A spatial case-control approach. *Accident; analysis and prevention*. 62. 10.1016/j.aap.2013.07.001.

Vangala, P., Lord, D., & Reddy Geedipally, S. (2015). Exploring the application of the Negative Binomial-Generalized Exponential model for analysing traffic crash data with excess zeros. *Analytic Methods in Accident Research*, 7 (2015), 29-36.

Wackernagel, H. (2003). Ordinary Kriging. In: *Multivariate Geostatistics*. Springer Berlin Heidelberg, pp79-88.

Wang, X., Fan, T., Chen, M., Deng, B., Wu, B., & Tremont, P. (2015). Safety modeling of urban arterials in Shanghai, China. *Accident Analysis & Prevention*, 83, 57-66.

Wang, K., Ivan, J., Ravishanker, N. and Jackson, E. (2017). Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. *Accident Analysis and Prevention*. Volume 99, Part A. P 6-19.

Wang, T., Wang, Y., Zhao, X., & Fu, X. (2018). Spatial distribution pattern of the customer count and satisfaction of commercial facilities based on social network review data in Beijing, China. *Computers, Environment and Urban Systems*, 71, 88-97.

Withers, S. D. (2002). Quantitative methods: Bayesian inference, Bayesian thinking. *Progress in Human geography*, 26(4), 553-566.

World Health Organization. (2019). Global action plan on physical activity 2018-2030: more active people for a healthier world. World Health Organization.

Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32(5), 396-406.

Xie, Z., & Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. *Journal of Transport Geography*, 31, 64-71.

- Xie, K., Wang, X., Ozbay, K. & Yang, H. (2014a). Crash frequency modeling for signalized intersections in a high-density urban road network. *Analytic Methods in Accident Research*. 2. 39–51. 10.1016/j.amar.2014.06.001.
- Xie, K., Ozbay, K. & Yang, H. (2014b). Spatial analysis of highway incident durations in the context of hurricane sandy. *Accident Analysis & Prevention*. 74. Doi: 10.1016/j.aap.2014.10.015.
- Xu, P., Huang, H., Dong, N. & Wong, S.C. (2017). Revisiting crash spatial heterogeneity: A Bayesian spatially varying coefficients approach. *Accident Analysis & Prevention*. 98. 330-337. 10.1016/j.aap.2016.10.015.
- Yang, B. Z. & Loo, B. (2016). Land use and traffic collisions: A link-attribute analysis using Empirical Bayes method. *Accident Analysis & Prevention*. 95. 236-249. 10.1016/j.aap.2016.07.002.
- Yanmaz-Tuzel, O., & Ozbay, K. (2010). A comparative Full Bayesian before-and-after analysis and application to urban road safety countermeasures in New Jersey. *Accident Analysis & Prevention*, 42(6), 2099-2107.
- Yin, H., Wong, S., Xu, J., & Wong, C. K. (2002). Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2), 85-98.
- Young, J. & Park, P. (2013). Benefits of small municipalities using jurisdiction-specific safety performance functions rather than the Highway Safety Manual's calibrated or uncalibrated safety performance functions. *Canadian Journal of Civil Engineering*. 40. 10.1139/cjce-2012-0501.
- Yu, R. and Abdel-Aty, M. (2013). Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. *Accident Analysis and Prevention*. 56:51-8. Doi: 10.1016/j.aap.2013.03.023.
- Yu, H., Liu, P., Chen, J., & Wang, H. (2014). Comparative analysis of the spatial analysis methods for hotspot identification. *Accident Analysis & Prevention*, 66, 80-88.
- Yu, R., Xiong, Y., & Abdel-Aty, M. (2015). A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. *Transportation Research Part C Emerging Technologies*, 50 (2015), 68-77.
- Zadeh, L.A. (1965). "Fuzzy sets". *Information and control*, 8, 338-353.
- Zadeh, L.A. (1989). *Knowledge representation in fuzzy logic*. IEEE transactions on knowledge and data engineering. Volume 1 (1).

Zeng, Q. and Huang, H. (2014). Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis & Prevention*, 67, 105-112.

Zeng, Q., Wen, H., Huang, H., Pei, X. & Wang, Shuai. (2017a). A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity. *Accident Analysis & Prevention*. 99. 184-191. 10.1016/j.aap.2016.11.018.

Zeng, Q., Wen, H., Huang, H., & Abdel-Aty, M. (2017b). A Bayesian spatial random parameters Tobit model for analyzing crash rates on roadway segments. *Accident Analysis & Prevention*, 100, 37-43.

8. Annex A: PICTURES OF SELECTED HIGHWAYS LOCATIONS

Figure 8-1: View of a section of the East road in a mountainous area, Aceh Besar district

Figure 8-2: View of a section of the East road in a mountainous area

Figure 8-3: View of a section of the East road in a rural environment (district of Bireun)

Figure 8-4: View of a road section of the East road in a rural environment

Figure 8-5: View of a section of the West road in a rural environment

Figure 8-6: View of a section of the West road in an urban environment

Figure 8-7: View of a section of the West road in the province of Aceh

9. Annex B: ADDITIONAL MAPS OF THE NetKDE METHOD

Figure 9-1: Hot spot distribution with NetKDE, with zoom in on Lhokseumawe (East road)

Figure 9-2: Hazardous sections identified with NetKDE, with zoom in on Lhokseumawe (East road)

Figure 9-3: Hot spot distribution with NetKDE, with zoom in on Manggeng (West road)

Figure 9-4: Hazardous sections identified with NetKDE, with zoom in on Manggeng (West road)

10. Annex C: ADDITIONAL MAPS OF THE HOT ZONE METHODOLOGY

Figure 10-1: Localization of the hot zones identified on the East Road

Figure 10-2: Localization of the hot zones identified in Lhokseumawe (East Road)

Figure 10-3: Comparison of the distribution of the hot zones (Hot zone method) and hot spots (NetKDE) in Lhokseumawe (East Road)

11. Annex D: SAFETY PERFORMANCE FUNCTIONS (BAYESIAN SPATIAL APPROACH)

In total, five scenarios were run to investigate. The safety performance functions used for each scenario are detailed below.

- **Total accident scenario:**

```
>route1.adj<-inla.geobugs2inla(adj$X1,cnt$X1,graph.file="route1.adj")

>formula <- tot_choq ~ offset(Length_km) + log(AADT) + as.factor(Land_use1) +
as.factor(Ali_hor) + f(Seg,model="bym", graph = route1.adj, hyper = list(prec.unstruct =
list(prior = "loggamma", param = c(1,0.01),prec.spatial = list(prior="loggamma",param =
c(1,0.01))))))

> model1.sp<-inla(formula,family="poisson",data = Segments, control.compute =
list(dic=TRUE))

> summary(model1.sp)
```

- **Fatal accidents scenario:**

```
>formula2 <- no_p_died ~ offset(Length_km) + log(AADT) + as.factor(Land_use1) +
as.factor(Ali_hor) + f(Seg,model="bym", graph = route1.adj, hyper = list(prec.unstruct =
list(prior = "loggamma", param = c(1,0.01),prec.spatial = list(prior="loggamma",param =
c(1,0.01))))))

>model2.sp<-inla(formula = formula2,family="poisson",data = Segments, control.compute =
list(dic=TRUE))

> summary(model2.sp)
```

- **Major injuries scenario:**

```
>formula3 <- Heavy_inj ~ offset(Length_km) + log(AADT) + as.factor(Land_use1) +
as.factor(Ali_hor) + f(Seg,model="bym", graph = route1.adj, hyper = list(prec.unstruct =
list(prior = "loggamma", param = c(1,0.01),prec.spatial = list(prior="loggamma",param =
c(1,0.01))))))

>model3.sp<-inla(formula = formula3,family="poisson",data = Segments, control.compute
= list(dic=TRUE))

> summary(model3.sp)
```

- **Minor injuries scenario:**

```
>formula4 <- Injury_Crs ~ offset(Length_km) + log(AADT) + as.factor(Land_use1) +
as.factor(Ali_hor) + f(Seg,model="bym", graph = route1.adj, hyper = list(prec.unstruct =
```

```
list(prior = "loggamma", param = c(1,0.01),prec.spatial = list(prior= "loggamma",param =  
c(1,0.01))))))  
  
>model4.sp<-inla(formula = formula4,family="poisson",data = Segments, control.compute  
= list(dic=TRUE))  
  
> summary(model4.sp)
```

- **PDO scenario:**

```
>formula5 <- PDO ~ offset(Length_km) + log(AADT) + as.factor(Land_use1) +  
as.factor(Ali_hor) + f(Seg,model ="bym", graph = route1.adj, hyper = list(prec.unstruct =  
list(prior = "loggamma", param = c(1,0.01),prec.spatial = list(prior= "loggamma",param =  
c(1,0.01))))))  
  
>model5.sp<-inla(formula = formula5,family="poisson",data = Segments, control.compute  
= list(dic=TRUE))  
  
> summary(model5.sp)
```

12. Annex E: ADDITIONAL MAPS OF THE BAYESIAN SPATIAL METHOD

- Total crash scenario: significant parameters land use (urban, and urban-rural) and access density.

Figure 12-1: Total crash scenario: representation of the significant parameter 'land use' (urban, and urban-rural)

Figure 12-2: Total crash scenario: representation of the significant parameter 'land use' (urban, and urban-rural), with the distribution of road accidents

Figure 12-3: Total crash scenario: representation of the significant parameter 'access density'

Figure 12-4: Total crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents

Figure 12-5: Total crash scenario: representation of all significant parameters, without the distribution of road accidents

Fatal crash scenario: significant parameters land use (urban-rural) and AADT.

Figure 12-6: Fatal crash scenario: representation of the significant parameter 'land use' (urban-rural)

Figure 12-7: Fatal crash scenario: representation of the significant parameter 'land use' (urban-rural), with the distribution of road accidents

Figure 12-8: Fatal crash scenario: representation of the significant parameter 'AADT'

Figure 12-9: Fatal crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents

Figure 12-10: Fatal crash scenario: representation of all significant parameters, without the distribution of road accidents

Major injury crash scenario: significant parameters access density and AADT.

Figure 12-11: Major injury crash scenario: representation of the significant parameter 'access density'

Figure 12-12: Major injury crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents

Figure 12-13: Major injury crash scenario: representation of the significant parameter 'AADT'

Figure 12-14: Major injury crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents

Minor injury crash scenario: significant parameters land use (urban-rural), access density and AADT.

Figure 12-15: Minor injury crash scenario: representation of the significant parameter 'land use' (urban-rural)

Figure 12-16: Minor injury crash scenario: representation of the significant parameter 'land use' (urban-rural), with the distribution of road accidents

Figure 12-17: Minor injury crash scenario: representation of the significant parameter 'access density'

Figure 12-18: Minor injury crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents

Figure 12-19: Minor injury crash scenario: representation of the significant parameter 'AADT'

Figure 12-20: Minor injury crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents

PDO crash scenario: significant parameters density access and AADT.

Figure 12-21: PDO crash scenario: representation of the significant parameter 'access density'

Figure 12-22: PDO crash scenario: representation of the significant parameter 'access density', with the distribution of road accidents

Figure 12-23: PDO crash scenario: representation of the significant parameter 'AADT'

Figure 12-24: PDO crash scenario: representation of the significant parameter 'AADT', with the distribution of road accidents