Masters Program in Geospatial Technologies

Spatio-temporal Modeling of Traffic Risk Mapping on Urban Road Networks

Somnath Chaudhuri

Dissertation submitted in partial fulfilment of the requirements for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

Spatio-temporal Modeling of Traffic Risk Mapping on Urban Road Networks

Dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Geospatial Technologies

February 21, 2020

Somnath Chaudhuri
som.rtc@gmail.com

Supervised by:
Pablo Juan Verdoy
Department of Mathematics
Universidade Jaume I

Co-supervised by:
Jorge Mateu
Department of Mathematics
Universidade Jaume I

Co-supervised by:
Ana Cristina Marinho da Costa
Information Management School
Universidade Nova de Lisboa

UNIVERSITAT JAUME I

NOVA IMS Information Management School

ifgi Institute for Geoinformatics University of Münster

Declaration of Academic Integrity

I hereby confirm that this thesis on *Spatio-temporal Modeling of Traffic Risk Mapping on Urban Road Networks* is solely my own work and that I have used no sources or aids other than the ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

February 21, 2020

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

February 21, 2020

Acknowledgements

At the outset, I intend to extend my heartful thanks and indebtedness to my supervisor Pablo Juan Verdoy, for his guidance and advice throughout the thesis work. I am extremely lucky to have a supervisor like him who responded to my every query so promptly and with patience. His inimitable energetic style of inspiration makes me enjoy the work and complete it on scheduled time. I would like to express my sincere gratitude to my co-supervisor Jorge Mateu for his continuous support, motivation, and immense knowledge. His valuable suggestions and encouragement kept me on the right track. My sincere thanks also go to my co-supervisor Ana Cristina Costa for her thoughtful comments and encouragement. I am also thankful to all the respected teachers for sharing their knowledge during the course, who through their teachings made me equipped with the necessary background needed to undertake such a work.

I must express my gratitude to my family, especially my parents for their continued support and encouragement throughout my studies and life. I owe special thanks to a very special person, my life-partner, Pam, for her continued support and motivation.

I would like to thank my global classmates for their help and suggestions. Special thanks go to my friends Ignacio Ponsoda Llorens, Mutaz Wajeh Abdlmajid Qafisheh and Alexys Herleym Rodríguez Avellaneda for their support and encouragement during the times of difficulty.

I acknowledge the UK Government Open Data Project for providing the traffic accident data of the city of London, UK. Finally, I would like to thank the r-spatial and R-INLA community for providing open-source tools and support.

Contents

List of Tables

List of Figures

List of Acronyms

AIC	Akaike Information Criterion
AUC	Area Under the Curve
CPO	Conditional Predictive Ordinate
DIC	Deviance Information Criterion
GF	Gaussian Function
GIS	Geographical Information System
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GMRF	Gaussian Markov Random Fields
GWR	Geographically Weighted Regression
INLA	Integrated Nested Laplace Approximation
LSOA	Lower Layer Super Output Area
MCMC	Markov Chain Monte Carlo
OSM	Open Street Map
RMSE	Root Mean Square Error
RSS	Residual Sum of Squares
SDF	Spatial Points Data Frame
SPDE	Stochastic Partial Differential Equations
WAIC	Watanabe-Akaike Information Criterion

Abstract

Over the past few years, traffic collisions have been one of the serious issues all over the world. Global status report on road safety, reveals an increasing number of fatalities due to traffic accidents, especially on urban roads. The present research work is conducted on five years of accident data in an urban environment to explore and analyze spatial and temporal variation in the incidence of road traffic accidents and casualties.

The current study proposes a spatio-temporal model that can make predictions regarding the number of road casualties likely on any given road segments and can generate a risk map of the entire road network. Bayesian methodology using Integrated Nested Laplace Approximation (INLA) with Stochastic Partial Differential Equations (SPDE) has been applied in the modeling process. The novelty of the proposed model is to introduce "SPDE network triangulation" precisely on linear networks to estimate the spatial autocorrelation of discrete events. The result risk maps can provide geospatial baseline to identify safe routes between source and destination points. The maps can also have implications for accident prevention and multi-disciplinary road safety measures through an enhanced understanding of the accident patterns and factors. *Reproducibility self-assessment : 3, 1, 1, 3, 2 (input data, preprocessing, methods, computational environment, results).*

Keywords: network triangulation, spatio-temporal modeling, traffic risk mapping

Chapter 1

Introduction

1.1 Background

Road traffic crashes is one of the serious issues around the globe. According to the global status report on road safety (2018) by the World Health Organisation, approximately 1.35 million people die each year as a result of traffic collisions (WHO, 2019). Rate of occurrence and severity of traffic crashes are the principal indicators of urban road safety measures (WHO, 2019). Literature suggests factors like road infrastructure, types of roads such as highways, double or, single carriage tracks play a vital role in road safety measure (Demasi et al., 2018). Uncontrolled vehicle speed, street junctions without traffic signals (Briz-Redón et al., 2019) incur accident risk. Temporal factors like, time of the day or, week-end nights also act as decisive factors in the count and impact of accidents (Farmer, 2005; Liu & Sharma, 2017). Identifying significant components has been a central focus of research in the domain of road safety.

Available map applications offered by larger corporations, such as Google Maps or, collaborative geospatial projects like OpenStreetMap (OSM) can provide information about the fastest (shortest) route from source to a destination point as depicted in Figure 1.1. The existing applications will suggest the shortest route without considering probable risk factors. Multi-disciplinary predictor aspects are not implemented in most of these applications.

Relevant spatio-temporal factors (Prasannakumar et al., 2011) play a significant role in identifying safe roads. According to (Williamson & Feyer, 1995), a particular road can be safe during mid-day, but the same road might not be safe during office hours. Traffic factors like street light, type of roads, speed limits act as significant factors in determining safe routes (Cantillo et al., 2016; Mohanty & Gupta, 2015). Thus, a multi-disciplinary approach is essential to explore spatio-temporal effects in road collisions. Identifying significant components (Deublein et al., 2013; Salifu, 2004) and spatio-temporal modeling of traffic accidents (Khulbe & Sourav, 2019; Zhong-xiang et al., 2014) have been an increasing trend in the domain of road safety management. But introducing Bayesian methodology on road networks using Integrated Nested Laplace Approximation (INLA) with Stochastic Partial Differential Equations (SPDE) in the domain of spatio-temporal predictive modeling is under explored in literature.

The current study is conducted on five years data of road traffic accidents from the city of London, UK. Spatial point process modeling on road networks

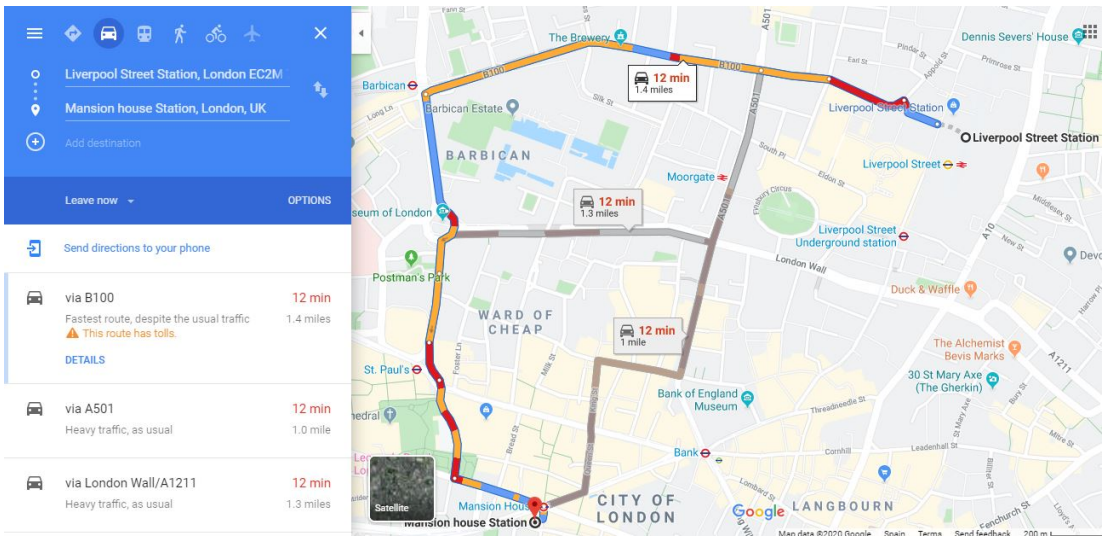


Figure 1.1: Google map application depicting the shortest route and potential alternatives between source-destination pair (GoogleMaps, 2019)

has been implemented using Bayesian methodology with INLA-SPDE. The model acts as a comprehensive scoring system that can predict risk index of individual road segment and can generate a categorized risk map of the entire road network. The methodology can be adapted and implemented to other locations of the UK and globally.

1.2 Related Work

Traffic accident fatalities have been on an increasing trend for the past few decades. According to the global status report on road safety (2018), road traffic injuries rank 9th among the leading causes of death globally (WHO, 2019). The impact of traffic collision fatalities is a social and public health challenge (Anderson, 2009). Therefore, to control the occurrence of traffic accidents and the impact of it on road traffic, it is essential to explore and analyze the factors that influence the occurrence of traffic accidents and to propose and implement corresponding accident analyzing models. Initial research works on road safety performance indicators (SPIs) (Gitelman et al., 2014) depict safe operational conditions of the road traffic system. Research works by (Ashraf et al., 2019; Azuike, 2018; Mohanty & Gupta, 2015) made notable contributions in identifying significant factors influencing traffic collisions. (Bhawkar, 2018) in his research work explored and analyzed the leading factors causing road accidents in the UK. (Shahid et al., 2015) mentioned that the causes of traffic collisions can be broadly classified into spatial and temporal components. Series of studies (Aghajani et al., 2017; Farmer, 2005; Jegede, 1988; Shafabakhsh et al., 2017) analyze historic data to identify risk factors and assess likelihoods of crash related events to categorize spatio-temporal factors affecting traffic accidents. These factors are considered as significant explanatory variables in statistical analysis and prediction modeling. Several statistical techniques, starting from traditional models like Poisson model's variations (Castro et al., 2012; Lord & Persaud, 2000; Miaou, 1993; Oh et al., 2006) or a negative binomial error structure (Pulugurtha & Sambhara, 2011)

to the logistic (Karacasu et al., 2013) and linear regressions (Abdel-Salam et al., 2008) have been applied to analyze spatial variability of traffic accidents. In this regard, (Sawalha & Sayed, 2003) highlights on statistical issues while modeling traffic accidents using Poisson and negative binomial regression. Similarly, (W. Wang et al., 2019) in their work on factors influencing traffic accident frequencies on urban roads mentioned that traditional traffic accident models assume accidents occurring at different locations are not related. In many cases, spatial autocorrelation of the traffic accidents has been ignored. But literature shows that, spatial methods are capable of incorporating geographical correlation in the model fitting process and in most of the cases spatial methods outperform the non-spatial models (Guo et al., 2018; Xu & Huang, 2015). Spatial heterogeneity in traffic accident modeling was implemented by using Geographically weighted regression (GWR). "*All accidents are not equal*" was established by (Zheng et al., 2011) using GWR to assess and forecast accident impacts. (Pirdavani et al., 2014) used GWR for spatial Analysis of fatal and injury crashes in Flanders, Belgium. Similarly, (Hezaveh et al., 2019) used GWR to estimate the cost of traffic crashes at a zonal level. To date publications regarding modeling of road traffic accidents using *descriptive analysis*, *linear regression* and *geographically weighted regression (GWR)* are summarized in Table 1.1 and Table 1.2

Table 1.1: Publications on modeling using descriptive analysis, linear regression and GWR

Year	**Reference**	**Model type**	**Case study**
2005	(Farmer, 2005)	Descriptive Analysis	US
2011	(Zheng et al., 2011)	GWR	Virginia, US
2012	(Castro et al., 2012)	Poisson model's variations	Texas, US
2013	(Karacasu et al., 2013)	Logistic regression and discriminant analysis	Turkey
2014	(Pirdavani et al., 2014)	GWR	Belgium
2015	(Ashraf et al., 2019)	Descriptive analysis, multivariate regression	South Korea
2015	(Shahid et al., 2015)	Non-parametric Mann–Kendall trend test	Malaysia
2015	(Xu & Huang, 2015)	GWR	Florida, US

Table 1.2: Publications on Modeling using descriptive analysis, linear regression and GWR (*continued*)

Year	**Reference**	**Model type**	**Case study**
2017	(Aghajani et al., 2017)	GIS-based Spatio-temporal analysis	Iran
2017	(Shafabakhsh et al., 2017)	GIS-based spatial analysis	Iran
2018	(Bhawkar, 2018)	Descriptive analysis	UK
2018	(Azuike, 2018)	Descriptive cross sectional analysis	Nigeria
2019	(W. Wang et al., 2019)	Spatio-temporal correlation	China
2019	(Lovelace et al., 2019)	GIS-based Spatio-temporal analysis	UK

Successive research works by (Juan et al., 2012; Karaganis & Mimis, 2006; Khulbe & Sourav, 2019; Loo et al., 2011) suggest that spatial point process modeling is one of the most appealing analytical tools to analyze the spatial and spatio-temporal distribution of traffic collisions. (Karaganis & Mimis, 2006) used spatial point process method to evaluate the probability of traffic accident occurrence on the national roads of Greece. Recently, a number of models on road safety have been proposed following Bayesian methodology. Cantilo et al. (2016) used a combined GIS-Empirical Bayesian approach in modeling traffic accidents in the urban roads of Columbia (Cantillo et al., 2016). A similar research work on urban road network of Florida by (Zeng & Huang, 2014) explored Bayesian spatial joint modeling of traffic crashes. A space–time multivariate Bayesian model was designed by (Boulieri et al., 2016) to analyze road traffic accidents by severity in different cities of UK. Traditionally, Bayesian approach with Markov chain Monte Carlo (MCMC) simulation methods are used to fit generalized linear mixed model (GLMM) (Wikle et al., 1998). (Song et al., 2006) depicted Bayesian multivariate spatial models using MCMC for mapping traffic crashes in Texas. Prior research suggests that with higher number of geo-locations the performance of MCMC models drop significantly (Rue et al., 2009). To balance speed and accuracy the possibility of studying spatial point processes by using integrated nested Laplace approximation (INLA) was suggested in literature (Bakka et al., 2018). Recently (Galgamuwa et al., 2019) used Bayesian spatial modeling with INLA in predicting road traffic accidents based on unmeasured information at road segment levels.

The above outlined studies focus on spatio-temporal modeling of traffic accidents based on diverse statistical methods. To date, the proposed models of road traffic accidents are summarized in Table 1.3 and Table 1.4 and described below.

Table 1.3: Publications on spatial point process modeling and INLA

Year	**Reference**	**Model type**	**Case study**
2006	(Karaganis & Mimis, 2006)	Inhomogeneous Poisson distribution using SUR method	Greece
2006	(Song et al., 2006)	Spatial multivariate Bayesian model	Texas, US
2007	(Rue & Martino, 2007)	INLA	-
2011	(Loo et al., 2011)	GIS-based network-constrained kernel density model	Sanghai, China
2012	(Juan et al., 2012)	Spatio-temporal point pattern analysis	Catalonia, Spain
2014	(Zeng & Huang, 2014)	Bayesian spatial joint model	Florida, US
2015	(Manley, 2015)	MCMC	UK
2016	(Cantillo et al., 2016)	GIS-Empirical Bayesian model	Colombia
2016	(Boulieri et al., 2016)	Spatio-temporal multivariate Bayesian model	UK
2017	(Rue et al., 2017)	INLA	-
2017	(Huang et al., 2017)	INLA-SPDE in environmental mapping	-
2018	(Bakka et al., 2018)	INLA-SPDE	-

Considering traffic accident events as discrete spatial points, spatial point process models can have higher predictive precision. But previous studies have emphasized little on implementing Bayesian methodology with INLA-SPDE in traffic accident modeling. Application of INLA-SPDE in spatial point process modeling in environmental mapping (Huang et al., 2017) and modeling environmental hazards like forest fire (Verdoy, 2019) are noteworthy. Another very recent trend in this domain is the spatial and spatio-temporal point pattern analysis on linear networks (Moradi, 2018).

Table 1.4: Publications on spatial point process modeling and INLA (*continued*)

Year	**Reference**	**Model type**	**Case study**
2018	(Moradi, 2018)	Spatio-temporal point pattern on linear networks	Australia, China, Colombia, Spain, UK and US
2019	(Verdoy, 2019)	INLA-SPDE in environmental hazards	Valencia, Spain
2019	(Galgamuwa et al., 2019)	INLA on road segment	Kansas, US
2019	(Moradi & Mateu, 2019)	Spatio-temporal point processes on linear networks	Colombia, UK, US

The present study suggests a spatio-temporal risk modeling of traffic accidents where Bayesian methodology has been implemented using INLA with SPDE. But the novel approach in this research work is the use of INLA-SPDE precisely on linear networks for spatial point process modeling of traffic accidents.

1.3 Aim and Objectives

The principal aim of the research work is to explore and analyze the spatial and temporal variation in the incidence of road traffic accidents and casualties. The current study seeks to propose the novel concept of multi-disciplinary road-safety analysis by introducing spatio-temporal risk modeling of traffic accidents using Bayesian methodology with INLA-SPDE precisely on road networks.

The objectives of the current study are:

- To assess the spatial and temporal road-safety components.
- To implement SPDE modeling of spatial point processes with INLA.
- To find predictions regarding the number of road casualties likely at any given road segment.
- To generate predicted traffic risk map on urban road networks.

Chapter 2

Tools and Applications

2.1 Tools

R programming language (version R 3.6.1) for statistical computing and graphical analysis and QGIS (version QGIS-3.8) geographic information system application have been extensively used throughout the current research analysis. RStudio (version RStudio 1.2.1335) integrated development environment has been used to implement R. Both QGIS (QGIS Development Team, 2009) and RStudio (RStudio Team, 2015) are open source and cross-platform desktop applications that is the advantage of selecting them as the principal processing tools in the study. Moreover, the online support for both R and QGIS from various open forums and communities are strong and reliable.

All the simulations during the current study were conducted on a quad-core Intel i7-4790 (3.60 GHz) processor with 16 GB (DDR3-1333/1600) RAM.

2.2 R Packages

The list of R packages used in the present research work is mentioned in Table 2.1, Table 2.2 and Table 2.3. For each package a brief description of its purpose along with the reference is reported.

Table 2.1: R packages for OSM network and data access

Package	**Purpose**	**Reference**
osmdata	Access and import OSM data as *sf* or *sp* objects	(Padgham et al., 2017)
stats19	Access UK official road traffic accident database (accidents, vehicles and casualties)	(Lovelace et al., 2019)

Table 2.2: R packages for data manipulation and representation

Package	**Purpose**	**Reference**
maptools	Manipulating spatial data	(R. Bivand & Lewin-Koh, 2019)
mapview	Interactive visualisations of spatial data in a map platform	(Appelhans, 2015)
rgdal	Bindings for geospatial data abstraction library	(R. Bivand et al., 2019)
sf	Support for simple features and standardize control to encode spatial vector data	(Pebesma, 2018)
sp	Classes and methods for spatial data	(R. S. Bivand et al., 2013)
statfunc	Clean redundant data and analyze basic metrics of the data set	(Chaudhuri, 2020) *unpublished working package*
stplanr	Provides spatial transport planning tools	(Robin Lovelace & Richard Ellison, 2018)
tidyverse	Collection of R packages like *dplyr, tidyr, ggplot2* for data exploration	(Wickham et al., 2019)

Table 2.3: R packages for data analysis and modeling

Package	**Purpose**	**Reference**
GWmodel	Geographically weighted (GW) model design	(Gollini et al., 2015)
INLA	Bayesian analysis using integrated nested Laplace approximation	(Rue et al., 2017)
spatstat	Spatial point pattern analysis	(Baddeley & Turner, 2005)
spgwr	compute geographically weighted regression (GWR)	(R. Bivand, 2017)

Chapter 3

Data and Experimental Design

This chapter illustrates an overview of the complete methodology followed in the current research work and provides detailed description of the data source along with data retrieval and preprocessing. It is structured as follows. The first section describes the complete workflow of the current study. Information about the study area and data source is reported in section two. The third section illustrates data preprocessing phases.

3.1 Design Overview

The complete workflow is depicted in Figure 3.1. Temporal and non-temporal (physical) variables along with Open Street Map (OSM) spatial data has been used as input to design the model. The final output is a traffic risk map for the entire road network of the study region.

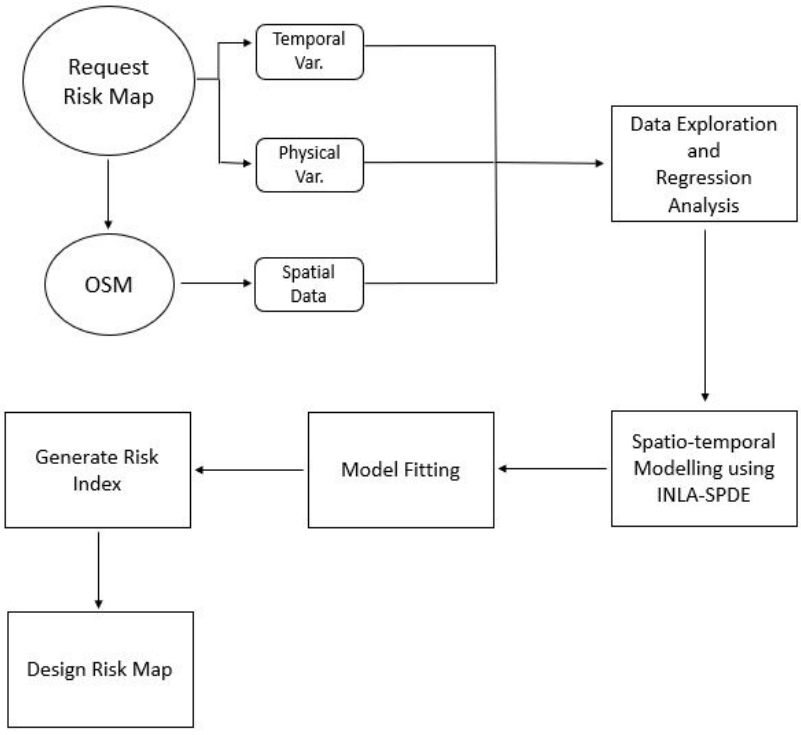


Figure 3.1: Complete workflow diagram

The workflow has been divided into two broad phases namely:

- Data exploration and regression analysis
- Spatio-temporal modeling

Data exploration and regression analysis phase consists of data preprocessing and exploratory data analysis followed by regression analysis of the processed data set. Data exploration refined the selection of explanatory variables that will be used in model fitting process. Regression analysis is performed to further investigate the relationship between the selected explanatory variables with the response variable. Detail workflow is illustrated in Figure 3.2.

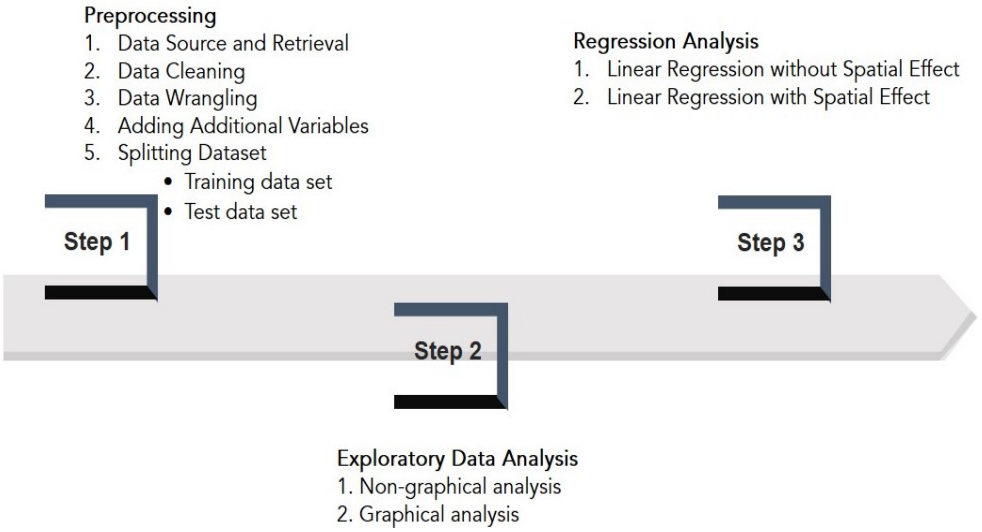


Figure 3.2: Workflow: Data exploration and regression analysis

Spatio-temporal modeling phase plays the vital role of model design and model fitting to achieve the objectives of the current study. Detail workflow of the modeling phase is illustrated in Figure 3.3. Agile methodology has been applied while executing both the phases.

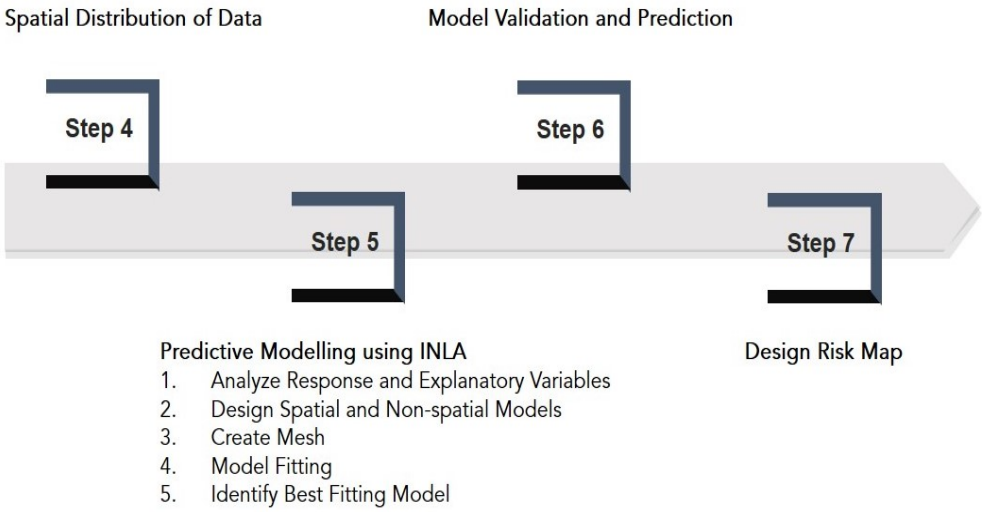


Figure 3.3: Workflow: Spatio-temporal modeling

3.2 Study Area and Data Source

The data set used in the current study contains detailed information of traffic accidents for thirteen years (2005-2017) that have occurred in the city of London, United Kingdom illustrated in Figure 3.4. After initial data cleaning and data exploration, a subset (2013-2017) of the initial data set was selected for the modeling process.

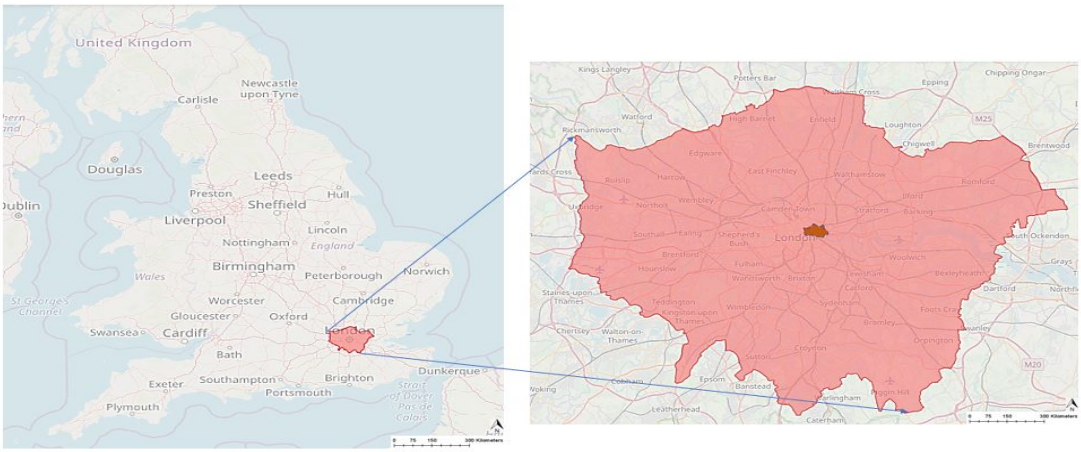


Figure 3.4: Study region: London, UK

City of London has an area of 2.90 square km. and comprises of six Lower Layer Super Output Area (LSOA) as depicted in Table 3.1. The study area is an important local government district that contains the historic center and the primary central business district (CBD) of London ('Wards', 2018). Figure 3.5 highlights the study area for the current research work.

Table 3.1: Study Area: Selected LSOA in city of London, UK

LSOA Code	**LSOA Name**
E01000001	City of London 001A
E01000002	City of London 001B
E01000003	City of London 001C
E01000005	City of London 001E
E01032739	City of London 001F
E01032740	City of London 001G

LSOA boundary and road network data is accessed from Open Street Map repository as *sf* or *sp* objects using "osmdata" (Lovelace et al., 2019) R package. OpenStreeMap plugin of QGIS application (QGIS Development Team, 2009) is used to facilitate download and conversion of network to shape files.

The Department for Transport, Government of UK publish road casualty statistics twice each year. Detailed data about the circumstances of road accidents on public roads reported to the police and consequential casualties are recorded

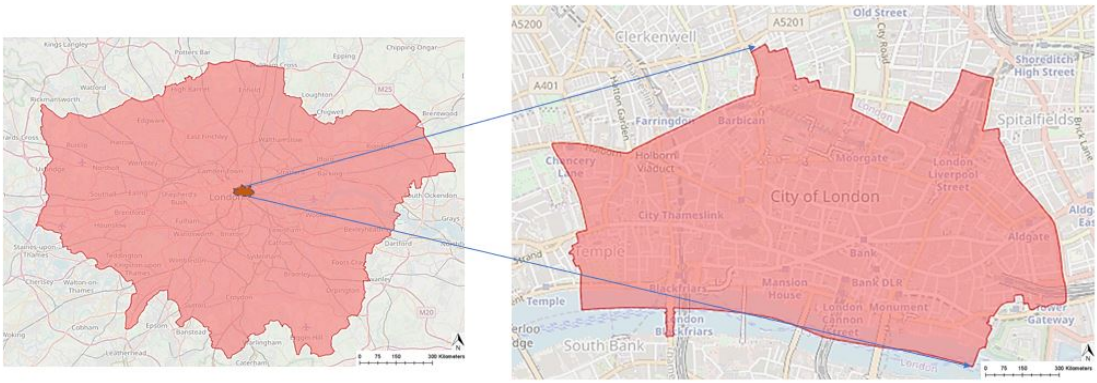


Figure 3.5: Study region: Selected LSOA regions of city of London, UK

using the *STATS19* accident reporting form. The complete data set since 1979 is available in the UK government open data repository (Transport for London, 2019). The data is free and available under the *Open Government Licence v3.0.* for public sector information, government of UK. It is also available from a number of online interactive geospatial portals. Figure 3.6 depicts an example of one such web portals of UK traffic accident data.

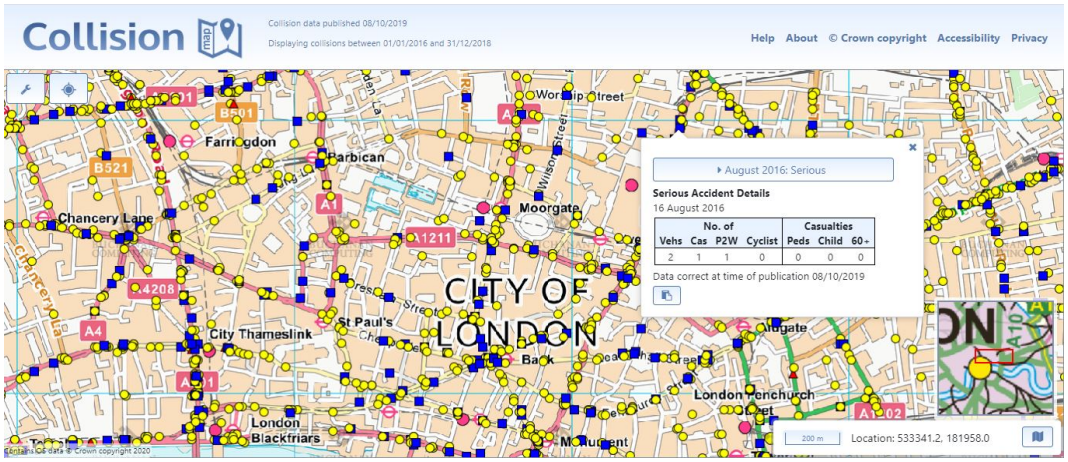


Figure 3.6: Example of UK traffic accident interactive data portal ('Buchanan Computing Collision Map', 2019)

R package "stats19" (Lovelace et al., 2019) provides UK road traffic accidents data in simple CSV downloadable format. In the current study, this package has been used to access and process UK official road traffic casualty database directly in the R platform. The data set accessed contains records from January 2005 to December 2017 for the selected LSOA regions mentioned in Table 3.1.

The reasons behind selecting the current study region are as follows. The data portal of the UK government maintains organized primary and secondary data published by the central government, local authorities and public bodies. Department for Transport (Transport for London, 2019), Government of UK provides open access to the data set. Government data is standardized and likely to be more accurate than other sources. This is one of the main reasons for selecting the collisions data set for the city of London, UK. Besides, previous

road safety literature conducted in the same region supports easy availability of reliable data. Additionally, the accident data set from "stats19" (Lovelace et al., 2019) provides wide range of explanatory variables (both temporal and spatial) for spatio-temporal modeling of road traffic accidents in an organized R package.

3.3 Data Preprocessing

An R package "statfunc" (Chaudhuri, 2020) is developed to perform initial data cleaning and preprocessing on the data set. The purpose of the package is to identify significant variables and data types, clean redundant data and analyze basic metrics of the data set. The package can be considered as a preprocessing tool having principal focus on identifying missing values and exploring the data set graphically. Initial preprocessing of data resolves redundant and missing data problems. It is ensured that the data set is consistent, free from error and convenient for exploratory data analysis.

3.3.1 Data Retrieval

Annual road collision data (from 2005 to 2017) can be accessed as a comma separated value (CSV) file using the R package "stats19" (Lovelace et al., 2019). Figure 3.7 depicts a glimpse of all the 32 variables available in the downloaded CSV files.

```
Variables: 32
$ Accident_Index                              <chr> "200501BS00001", "200501BS00002", "200501BS00003", "200501BS00004", "2005…
$ Location_Easting_OSGR                       <dbl> 525680, 524170, 524520, 526900, 528060, 524770, 524220, 525890, 527350, 5…
$ Location_Northing_OSGR                      <dbl> 178240, 181650, 182240, 177530, 179040, 181160, 180830, 179710, 177650, 1…
$ Longitude                                   <dbl> -0.191170, -0.211708, -0.206458, -0.173862, -0.156618, -0.203238, -0.2112…
$ Latitude                                    <dbl> 51.48910, 51.52007, 51.52530, 51.48244, 51.49575, 51.51554, 51.51270, 51.…
$ Police_Force                                <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ Accident_Severity                           <dbl> 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 2, 3, 3, 3, 3, 3…
$ Number_of_Vehicles                          <dbl> 1, 1, 2, 1, 1, 2, 2, 1, 2, 2, 1, 2, 1, 2, 1, 1, 2, 2, 2, 1, 1, 2, 2, 2, 2…
$ Number_of_Casualties                        <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 2, 5, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ Date                                        <chr> "04/01/2005", "05/01/2005", "06/01/2005", "07/01/2005", "10/01/2005", "11…
$ Day_of_Week                                 <dbl> 3, 4, 5, 6, 2, 3, 5, 6, 7, 7, 1, 3, 3, 3, 3, 3, 5, 6, 6, 7, 2, 2, 2, 3, 7…
$ Time                                        <time> 17:42:00, 17:36:00, 00:15:00, 10:35:00, 21:13:00, 12:40:00, 20:40:00, 17…
$ `Local_Authority_(District)`                <dbl> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1…
$ `Local_Authority_(Highway)`                 <chr> "E09000020", "E09000020", "E09000020", "E09000020", "E09000020", "E090000…
$ `1st_Road_Class`                            <dbl> 3, 4, 5, 3, 6, 6, 5, 3, 3, 4, 3, 3, 6, 3, 3, 3, 6, 3, 4, 3, 6, 4, 3, 5, 3…
$ `1st_Road_Number`                           <dbl> 3218, 450, 0, 3220, 0, 0, 0, 315, 3212, 450, 4, 3220, 0, 3217, 4, 3217, 0…
$ Road_Type                                   <dbl> 6, 3, 6, 6, 6, 6, 6, 3, 6, 6, 6, 6, 2, 2, 3, 6, 6, 6, 6, 6, 6, 6, 3, 6, 3…
$ Speed_limit                                 <dbl> 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 3…
$ Junction_Detail                             <dbl> 0, 6, 0, 0, 0, 0, 3, 0, 6, 3, 6, 6, 3, 3, 0, 3, 3, 3, 0, 3, 0, 3, 6, 6, 3…
$ Junction_Control                            <dbl> -1, 2, -1, -1, -1, -1, 4, -1, 2, 4, 2, 2, 4, 4, -1, 4, 4, 4, -1, 4, -1, 4…
$ `2nd_Road_Class`                            <dbl> -1, 5, -1, -1, -1, -1, 6, -1, 4, 5, 4, 3, 3, 3, -1, 6, 6, 3, -1, 6, -1, 5…
$ `2nd_Road_Number`                           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 304, 0, 325, 308, 3220, 3216, 0, 0, 0, 4, 0, 0, 0…
$ `Pedestrian_Crossing-Human_Control`         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ `Pedestrian_Crossing-Physical_Facilities`   <dbl> 1, 5, 0, 0, 0, 0, 0, 0, 5, 8, 5, 5, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 5…
$ Light_Conditions                            <dbl> 1, 4, 4, 1, 7, 1, 4, 1, 4, 1, 4, 4, 1, 4, 1, 1, 4, 1, 4, 4, 4, 4, 4, 4, 1…
$ Weather_Conditions                          <dbl> 2, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ Road_Surface_Conditions                     <dbl> 2, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 4, 1, 2…
$ Special_Conditions_at_Site                  <dbl> 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Carriageway_Hazards                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Urban_or_Rural_Area                         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ Did_Police_Officer_Attend_Scene_of_Accident <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 2, 1…
$ LSOA_of_Accident_Location                   <chr> "E01002849", "E01002909", "E01002857", "E01002840", "E01002863", "E010028…
```

Figure 3.7: Available variables in data set accessed using stats19 R package

Department for Transport, United Kingdom maintains detailed scope and connotation of each field used in the data set. The document entitled *Instructions for the Completion of Road Accident Reports* is available online as STATS 20 (Transport for London, 2019). For every reported road accidents, the above mentioned 32 distinct variables are required to be collected by the London police department. Before using the data, analytical user's manual from STATS 20 was referred. It is noteworthy to mention that the traffic accident data from the

Department for Transport, United Kingdom and R package "stats19" contain the records of *only fatal accidents*. Traffic accidents having at least one causality count have been used in the current study. Therefore, the sampled data set does not contain information about accidents where there is no casualty.

3.3.2 Data Cleaning

The pre-processed data set required to be cleaned to ensure consistency and reliability. Initially, redundant variables having meta data information and duplicate observations are discarded. R package "statfunc" (Chaudhuri, 2020) has been used extensively during this phase. Next, data cleaning is executed as follows:

- *Identify observations having missing or, null values*: Observations having one or more empty fields are identified.
- *Discard identified observations*: Delete the identified complete tuple (observation) if one or, more fields values are missing in it.

3.3.3 Data Wrangling

Good variable names are crucial for better readability and ease to use in dynamically typed languages. Thus, all the variables are converted to lower case and assigned relevant and self-explanatory names.

Literature states, in a combined model which can reflect the influence of multiple factors on traffic accidents and improve prediction accuracy, the number of death tolls for road traffic accidents play a vital role (Zhong-xiang et al., 2014). In context to the research objectives (mentioned in Section 1.3), *number_of_casualties* has been considered as the response variable in the current study. As discussed in Section 3.3.1, the traffic accident data set contains the records of only fatal accidents, thus it can be stated that the minimum value of response variable is 1. At the same time, it was observed that the number of casualties less than equal to 4 is comprising 98.97% of the data set. It indicates that the frequency of accident casualties greater than 4 is very low compared to the other range. Thus for computational ease, the field values are re-coded as factors mentioned below.

$$\text{number of casualties} = 1 \implies 1$$

$$\text{number of casualties} = 2 \implies 2$$

$$\text{number of casualties} = 3 \implies 3$$

$$\text{number of casualties} = 4 \implies 4$$

$$\text{number of casualties} > 4 \implies 5$$

Other categorical variables used in the study are also re-coded as factors.

3.3.4 Adding Additional Variables

A number of derived explanatory variables are added with the original data set. These variables will play vital role in the exploratory data analysis and in designing the final spatio-temporal model.

- From the field *date*, derived variables like *year*, *month*, *day* and *week* are added as individual fields.

- Random day variable (name: *rw_ date*, type: *integer*): The complete time period from January 2005 to December 2017 is assigned a continuous day variable using Julian Counts function (Chalabi et al., 2011). The first day of 2005 i.e. 01.01.2005 has been assigned value 1 and so on in sequential order. The last day of the data set i.e. 31.12.2017 has been assigned the maximum sequential value generated as the total number of days in the data set. This variable will be used to fit the temporal effect in the SPDE model.

- Week-end night time-slots (name: *week_ end_ night*, type: *binomial*): This variable is used as an important explanatory variable in regression and INLA models. Standard week-end nights are generated using the following time periods: *Friday 22:00 to 23:59 Hrs.*, *Saturday 00:00 to 6:00 Hrs.* and *22:00 to 23:59 Hrs.* and *Sunday 00:00 to 6:00 Hrs.*. If any accident occurred during the above-mentioned time period then the field value will be assigned 1 else it will have value 0.

- Logistic regression response variable (name: *logi_ reg_ var*, type: *binomial*): Logistic regression analysis requires a binary response variable. The traditional method of categorizing response variables in case of logistic regression is dichotomous in nature (Hoffman, 2019). The response value is generally assigned 0 when there is no occurrence and all other frequencies of occurrence are categorized as 1. Section 3.3.1 states that the data set contains the records of fatal accidents only. Thus, the traditional categorization method based on "events" or, "non-events" could not be applied in this study. But research works on quality prediction and control using logistic regression (Jin et al., 2007; Sampson et al., 2016) have adopted quality measure as the response variable. If the predicted quality is lower or equal to a threshold value then it is assigned label 0 otherwise response variable will have value 1. In contrast to the current study, impact of traffic crashes is one of the principal indicators of urban road safety. Number of casualties in a traffic collision is an important measure of the impact (WHO, 2019). Section 3.3.3 highlights the fact that 98.97% of the sampled traffic accidents have number of casualties less than or equal to 4. This implies, accidents having casualties more than four have considerably low occurrence compared to other groups. Thus, casualty value 4 has been set as a threshold value to convert response variable into binary form. The value of response variable has been assigned label 0 when the number of casualties is greater than 4. All other casualty values ranging from 1 to 4 are assigned label 1.

3.3.5 Splitting Dataset

The pre-processed data set is divided into two categories:

- Training data set (2005-2014): to develop and train the predictive model.

- Test data set (2015-2017): to assess the performance of the proposed model.

Chapter 4

Pre-modeling Data Analysis

This chapter presents and discusses the results of exploratory data analysis. The first section analyzes the data set to summarize main characteristics and identify outliers with the help of relevant graphical methods. The selected explanatory and response variables are also reported in this section. In the next section, spatial and non-spatial generalised linear regression models are explored to identify relationships between explanatory and response variables.

4.1 Data Exploration

Exploratory data analysis is the initial investigations on data to spot outliers, to identify patterns and to test hypothesis by performing summary statistics and graphically representing the data. Current data exploration initiated with a glimpse of the complete data set as depicted in Figure 4.1.

```
Observations: 957
Variables: 28
$ accident_index            <fct> 200501CP00009, 200501CP00036, 200501CP00039, 200501CP00060, 200501CP00063, 200501CP0…
$ longitude                 <dbl> -0.096983, -0.097308, -0.097149, -0.078865, -0.075025, -0.076890, -0.097293, -0.0742…
$ latitude                  <dbl> 51.51629, 51.51540, 51.51575, 51.51528, 51.51405, 51.51425, 51.51576, 51.51439, 51.5…
$ accident_severity         <fct> Serious, Serious, Slight, Slight, Slight, Slight, Fatal, Slight, Slight, Slight, Sli…
$ number_of_vehicles        <int> 2, 2, 3, 3, 1, 2, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 1, 2, 1, 1, 2, …
$ number_of_casualties      <int> 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, …
$ date                      <fct> 2005-01-10, 2005-03-03, 2005-03-23, 2005-04-11, 2005-04-16, 2005-05-12, 2005-05-18, …
$ day_of_week               <int> 2, 5, 4, 2, 7, 5, 4, 7, 7, 2, 3, 2, 6, 6, 4, 3, 5, 3, 2, 4, 1, 2, 5, 6, 4, 4, 6, 2, …
$ time                      <fct> 1899-12-31 15:00:00, 1899-12-31 13:31:00, 1899-12-31 10:05:00, 1899-12-31 10:59:00, …
$ road_type                 <fct> One way street, One way street, One way street, One way street, Single carriageway, …
$ speed_limit               <int> 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, …
$ junction_detail           <int> 3, 0, 3, 3, 0, 1, 3, 1, 0, 0, 3, 3, 0, 6, 3, 9, 1, 3, 0, 1, 1, 0, 6, 3, 7, 9, 3, 3, …
$ junction_control          <int> 4, -1, 4, 4, -1, 2, 4, 4, -1, -1, 4, 4, -1, 2, 4, 4, 4, 4, -1, 4, 2, -1, 2, 4, 2, 4,…
$ light_conditions          <int> 1, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 1, 4, 1, 1, 4, 1, 4, 4, 4, 4, 1, 4, 1, 1, 1, 1, 1, …
$ weather                   <fct> Raining no high winds, Fine no high winds, Fine no high winds, Fine no high winds, F…
$ road_surface              <int> 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, …
$ special_conditions_at_site <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, …
$ lsoa                      <fct> E01000001, E01000001, E01000001, E01000005, E01000005, E01000005, E01000001, E010000…
$ day_of_week_name          <fct> Monday, Thursday, Wednesday, Monday, Saturday, Thursday, Wednesday, Saturday, Saturd…
$ week_end_night            <fct> No, No, No, No, No, No, No, No, Yes, No, No, No, No, No, No, No, No, No, No, No, Yes…
$ newDate                   <fct> 2005-01-10, 2005-03-03, 2005-03-23, 2005-04-11, 2005-04-16, 2005-05-12, 2005-05-18, …
$ year                      <int> 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, …
$ month                     <int> 1, 3, 3, 4, 4, 5, 5, 6, 6, 6, 8, 8, 9, 9, 10, 10, 10, 11, 11, 12, 10, 10, 10, 12,…
$ week                      <int> 2, 9, 12, 15, 16, 19, 20, 23, 23, 26, 25, 33, 35, 39, 39, 43, 43, 40, 46, 47, 50, 42…
$ day                       <int> 10, 62, 82, 101, 106, 132, 138, 155, 155, 178, 172, 227, 238, 266, 271, 298, 300, 27…
$ time_slot                 <int> 15, 13, 10, 10, 16, 8, 18, 16, 0, 13, 8, 6, 20, 16, 17, 19, 11, 22, 17, 18, 2, 10, 5…
$ rw_date                   <int> 10, 62, 82, 101, 106, 132, 138, 155, 155, 178, 172, 227, 238, 266, 271, 298, 300, 27…
$ logi_reg_var              <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
```

Figure 4.1: Glimpse of pre-processed data set

Pre-processed data set comprises of 957 observations and 28 variables. Out of which *number_ of_ causalities* is selected as the response variable and from the rest 27, significant explanatory variables are identified through exploratory data

analysis. Most of the variables in the data set are categorical variables. Thus, performing summary statistics or, frequency distribution and scatter plot will not be beneficial. Result of basic metrics and missing value analysis are depicted in Figure 4.2. It confirms that, after initial data cleaning phase, no column and observation have null or, missing values.

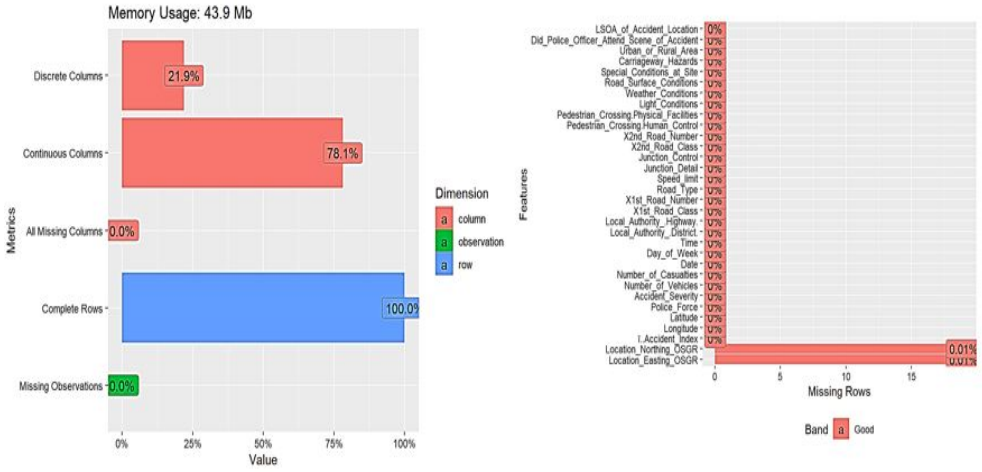


Figure 4.2: Summary of basic metrices and missing values

Graphical representation helps in identifying special effects in data set, indicate outliers, detect patterns, diagnose models and generally search for novel and perhaps unexpected phenomena (Everitt, 2006). According to (Chambers et al., 1983) "*there is no statistical tool that is as powerful as a well-chosen graph.*" Figure 4.3 depicts the total annual accident count and mean monthly accident count during the study period (2005 to 2014). It is observed that, the total number of accidents from 2005 to 2012 are comparatively low. But there is a sharp increase in the number from 2012 to 2013 and 2014.

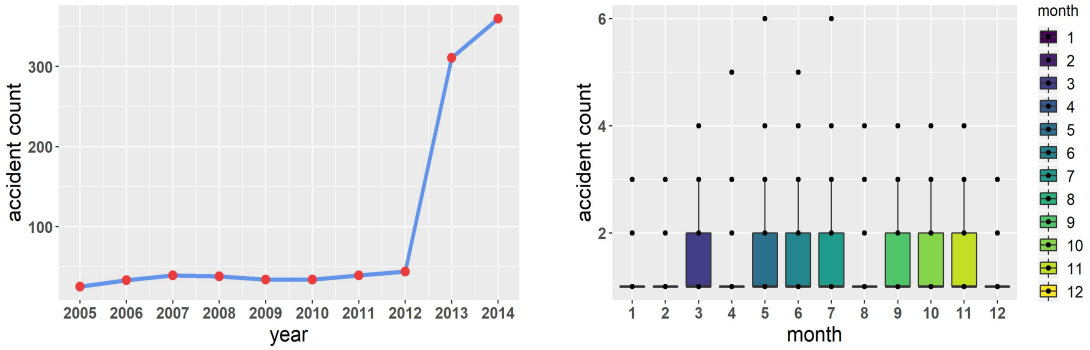


Figure 4.3: Total annual and mean monthly accident counts (2005-2014)

Figure 4.4 illustrates distribution of daily and hourly accident count. It can be noted that Sunday (marked as 1 in Figure 4.4 has the minimum occurrence of accidents compared to other week-days. On the other hand, there is a trend of relatively high accident count during the office hours (7:00-9:00 Hrs. and 17:00-18:00 Hrs.).

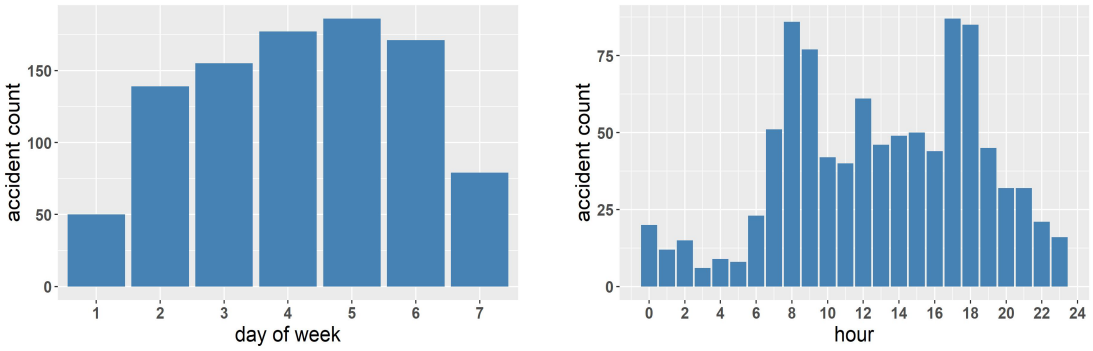


Figure 4.4: Daily and hourly accident counts (2005-2014)

To identify the distribution of traffic accidents over all the months and days of the week, a combined graph depicting mean number of daily accidents by day of the week and month is created as shown in Figure 4.5. It can be identified that the months of July and November are having relatively higher accident counts compared to other times of the year. In context to days of the week, Monday and Thursday are found to have higher records of traffic accidents.

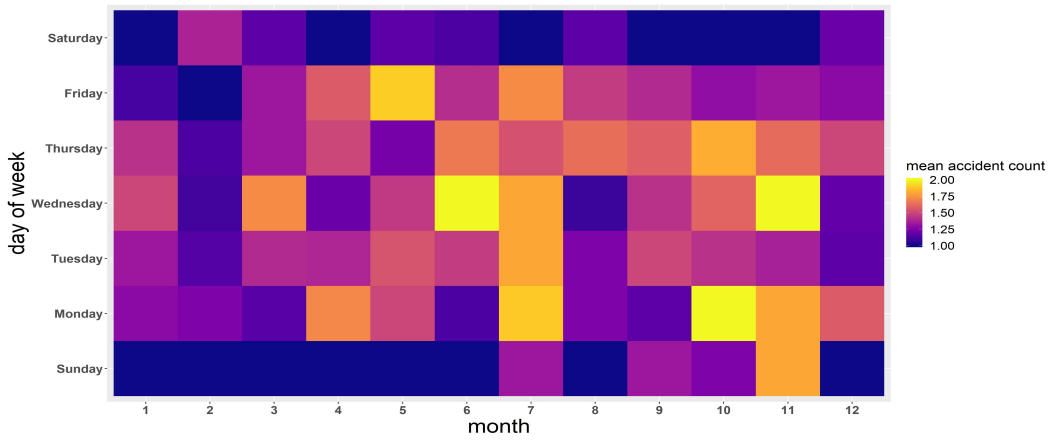


Figure 4.5: Mean daily accident count grouped by month and day of the week

Figure 4.6 illustrates that accident type "slight" has comparatively higher occurrence than serious or fatal accidents over any months of the year.

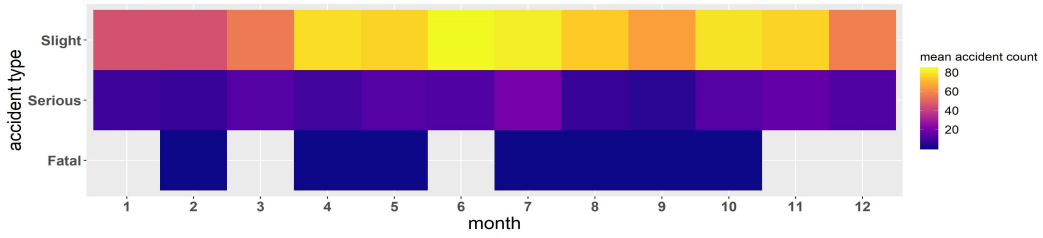


Figure 4.6: Mean monthly accident count grouped by severity of accident

Literature shows there is significant difference in traffic accident counts between week-days and week-ends (Scott et al., 2016). In the current study, traffic accidents during the week-end nights (mentioned in Section 3.3.4) and during whole

week-days are explored. With reference to Figure 4.7, week-end nights do not seem to influence the toll of traffic accidents for the current data set.

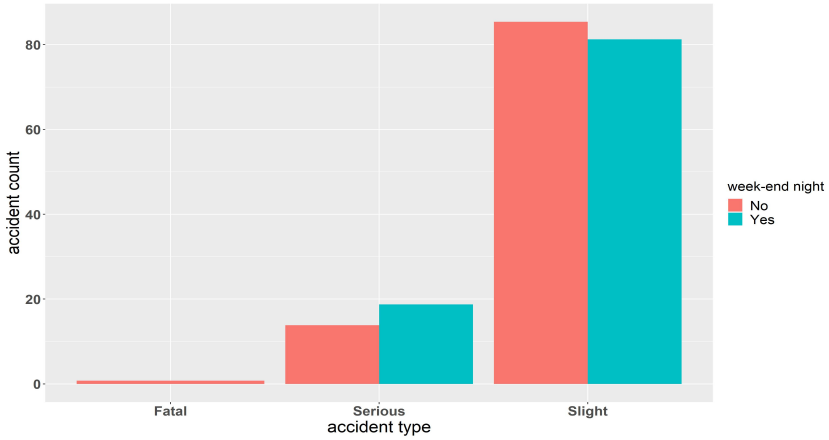


Figure 4.7: Count of types of accidents grouped by week-end nights and whole week-days

Adverse weather has an impact on vehicle crash rates on roads and highways. It usually increases during precipitation (Qiu & Nixon, 2008). Figure 4.8 depicts that, in the current data set it seems there is no significant influence of weather conditions in all three types of accidents. Most of the accidents have been observed during fine weather condition.

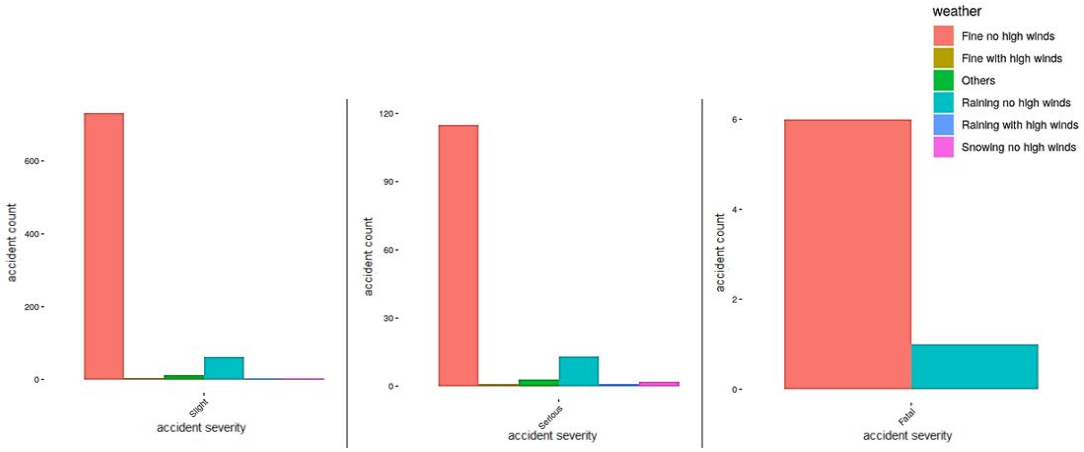


Figure 4.8: Count of accidents grouped by accident severity and weather conditions

Literature (Pikūnas et al., 2004) states speed limit and type of road has strong influence on the accident rate. Figure 4.9 establishes the fact that in the current data set for the city of London, most of the accidents have occurred in single carriage way roads. On the other hand, speed limit does not seem to influence the rate of accidents.

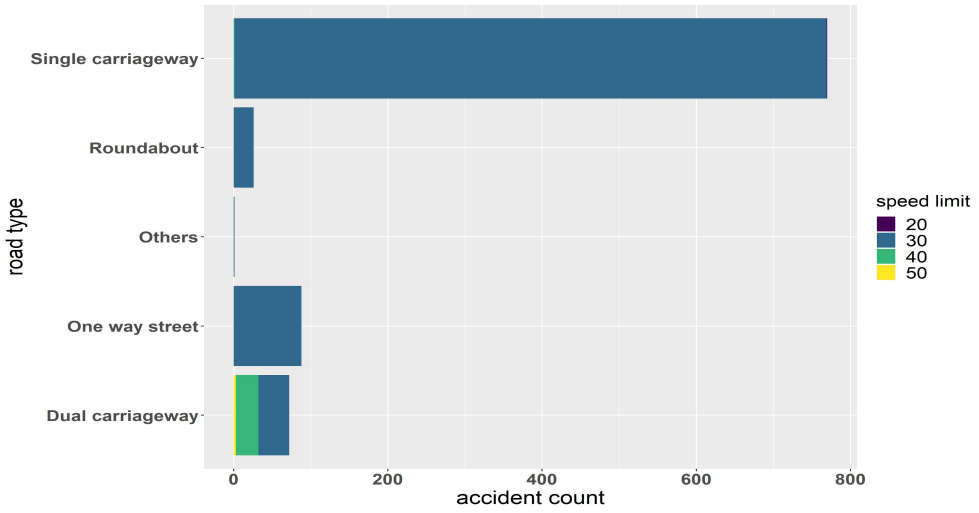


Figure 4.9: Count of accidents grouped by speed-limit and road type

Distribution of traffic accidents in study region over the Open Street Map (OSM) road network is visualized in the next phase of data exploration. Figure 4.10 depicts the road network accessed using R package "osmdata" (Padgham et al., 2017) from the OSM repository and visualized in an interactive environment using R package "mapview" (Appelhans, 2015). It is noteworthy to mention that, in the current study OSM highways of type *unclassified, bus_guideway, raceway, bridleway, path* are not included in the list of OSM highway categories.

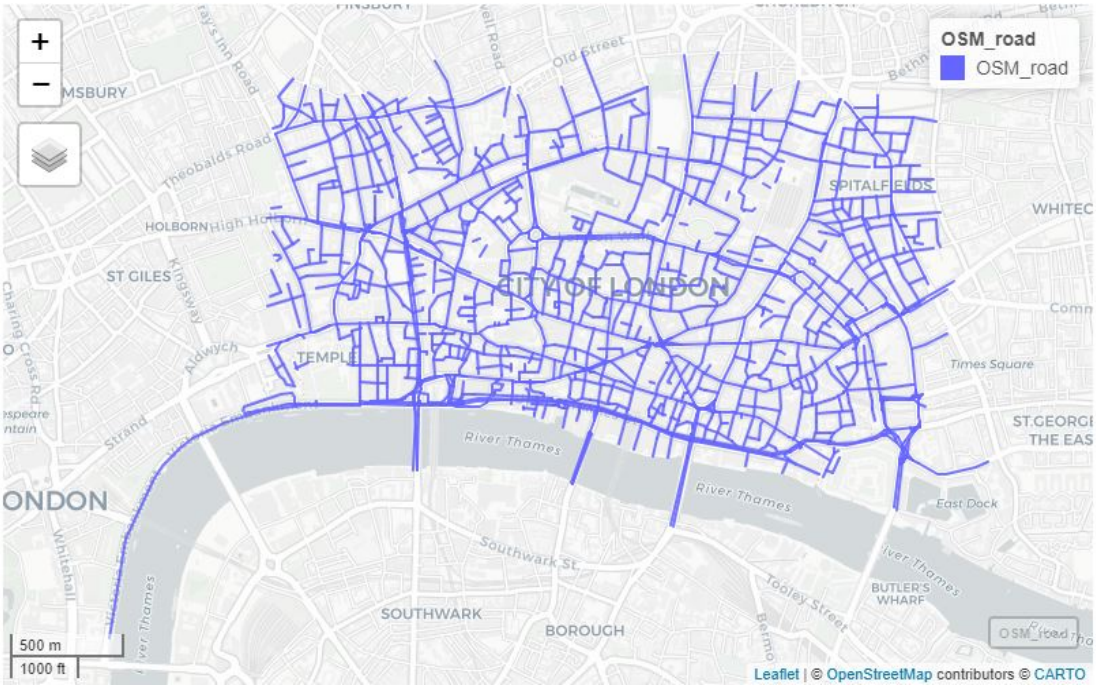


Figure 4.10: OSM road network

Figure 4.11 represents OSM road network with the spatial distribution of traffic accidents during the study period of 2005 to 2014; visualized using QGIS application (QGIS Development Team, 2009). Three distinct types of accident based on severity measure (slight, mild and severe) are depicted in the map.

From the illustrations, it can be stated that, the traffic accidents are basically concentrated in the central city location and as expected are observed mostly on the primary or, secondary road networks. The city outskirts and residential or, tertiary road networks are found to have relatively low accident occurrence.

Figure 4.11: OSM road network with spatial distribution of traffic accidents

Thus, by performing exploratory data analysis, significant variables required to develop the proposed spatio-temporal model are identified. It helps to discard redundant explanatory variables and make sure that the results they produce are valid, duly interpreted, and applicable to the desired objectives. In the current study, out of twenty seven independent variables, finally eleven are identified to be used for further analysis and model design. The identified explanatory variables used in the modeling process are mentioned in Table 4.1. It is noteworthy to mention that in case of INLA-SPDE modeling temporal effect has been implemented using *rw_ date* explanatory variable as discussed in Section 3.3.4.

Table 4.1: Selected explanatory variables

Physical variables	**Temporal variables**
Road type	Time slot
Road surface	Week-end night
Junction location	Day of week
Junction detail	Month
Speed limit	
Light condition	
Weather	

In case of logistic regression analysis, binary variable *logi_ reg_ var* (mentioned

in Section 3.3.4) has been used as the response variable. On the other hand, variable *number_of_casualties* in the original discrete form has been used as response variable in Poisson regression and GWR analysis. But the same variable *number_of_casualties* has been converted into categorical form (mentioned in Table 4.3) to be implemented as the predicted variable in INLA-SPDE modeling processes. Response variables along with their respective category labels used in different statistical methods of the current study are reported in Table 4.2 and Table 4.3.

Table 4.2: Response variables used in statistical models

Model type	**Variable name**	**Variable type**
Poisson regression	*number_of_casualties*	discrete
Logistic regression	*logi_reg_var*	binary
GWR	*number_of_casualties*	discrete
INLA-SPDE	*number_of_casualties*	categorical

Table 4.3: Response variable values

Response Variable *number_of_casualties (categorical)*		**Response Variable** *logi_reg_var*	
Factor levels	**Factor labels**	**Variable value**	**Binary value**
Casualty = 1	1	Casualty > 4	0
Casualty = 2	2	Casualty ≤ 4	1
Casualty = 3	3		
Casualty = 4	4		
Casualty >4	5		

4.2 Regression Analysis: Generalized Linear Model (GLM)

Data exploration refined the selection of explanatory variables that will be used in model fitting process. In the next phase, regression analysis is performed to further investigate the relationship between the selected explanatory variables with the response variable.

Regression analysis is conducted in two different approaches as depicted in Figure 4.12. The first process is multiple linear regression method without having any spatial influence. This method is executed using two Generalized Linear Model (GLM) regression techniques namely, Poisson and logistic regression methods. The second process is spatial regression method by implementing spatially varying fields. In this case, Geographically Weighted Regression (GWR) method is performed which extends the traditional regression framework by incorporating the estimation of local rather than global variables.

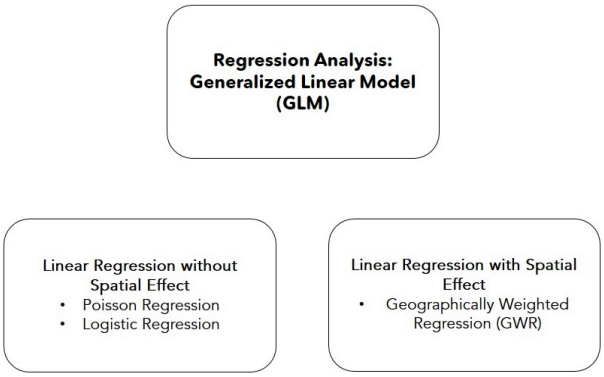


Figure 4.12: Implemented Generalized Linear Model (GLM)

4.2.1 Linear Regression without Spatial Effect

Ordinary linear regression predicts the expected response variable as a linear combination of one or more predictors. This is applicable when the response variable has a normal distribution. On the other hand, a generalized linear model is made up of a linear predictor and allow the magnitude of the variance of each measurement to be a function of its predicted value (Poul Thyregod, 2010). Mathematically it is represented as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i \tag{4.1}$$

where the response variable y_i, $i = (1, ..., n)$ is modeled by a linear function of explanatory variables x_j, $j = (1, ..., p)$ plus an error term ε.

Results obtained from Section 4.1 indicate that in the current data set the response variable *number_of_casualties* and *logi_reg_var* are not normally distributed. Thus, in this case generalized linear model is implemented which allows arbitrary distributions of response variable as well as an arbitrary function of the response variable.

Poisson Regression:

"*A Poisson Regression model is a Generalized Linear Model (GLM) that is used to model count data and contingency tables. The output (count) is a value that follows the Poisson distribution*" (Zamani & Ismail, 2013). A link function is used to transform non-linear relationship to linear form. In case of Poisson regression, it is log function. Mathematically Poisson Regression model can be represented as:

$$log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{4.2}$$

where, response variable y and x_i are the response and explanatory variables respectively. β are numeric coefficients.

Logistic Regression:

Logistic regression is the statistical modeling process of a binomial response variable with one or more explanatory variables. It measures the relationship by estimating probabilities using a logistic function (Hoffman, 2019). Mathematically represented as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{4.3}$$

where, p represents the probability of the response variable. x_i and β are the explanatory variables and numeric coefficients respectively.
In case of Poisson regression, discrete variable *number_of_casualties* and in case of logistic regression, binomial variable *logi_reg_var* has been considered as the response variable (mentioned in Table 4.2). In both cases, explanatory variables are the selected eleven variables reported in Table 4.1. The R code for generalized linear models for Poisson (Bruin, 2011) and logistic regression (Zhang, 2016) are adopted from previous studies and modified for the current analysis. Residual plot for Poisson and logistic regression methods are depicted in Figure 4.13.

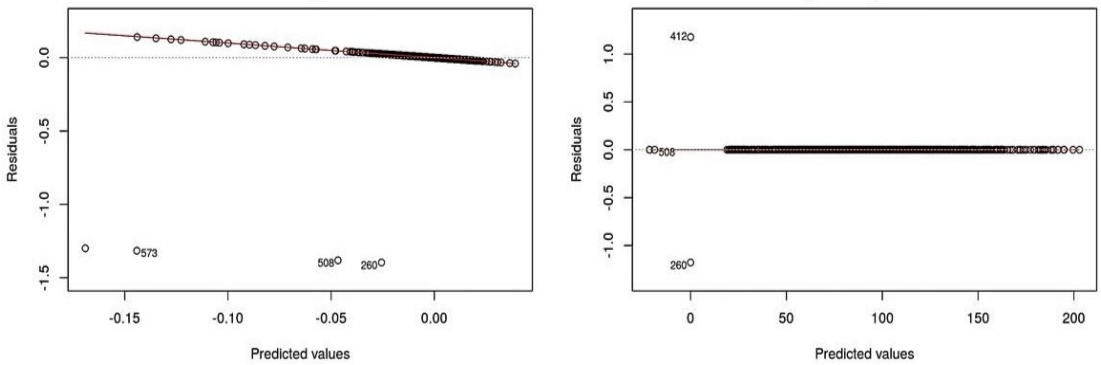


Figure 4.13: Residual vs. fitted plot for a) Poisson and b) logistic regression

The residual diagnostics in Figure 4.13 represents the residual values in both models are close to zero. Table 4.4 depicts the results of Poisson and logistic regression analysis.

Detailed analysis of deviance tables for Poisson and logistic regression models are reported in Appendix A.1. From the result it can be stated that, not a single

Table 4.4: Result: Poisson and logistic regression

Reg.	Deviance Residual Values					AIC	AUC
Family	Min	1Q	Median	3Q	Max		
Poisson	-1.39624	-0.00684	0.00119	0.01113	0.10463	2051.6	0.9773085
Logistic	-1.177	0.00	0.00	0.00	1.177	140.77	0.9998688

**AIC: Akaike information criterion*

**AUC: Area under the curve*

variable can be considered as a significant fit for the regression models. At the same time, from the residual plot in Figure 4.13 it is not possible to identify in a geographical pattern where the model over or under-predicts. Thus, to explore spatial heterogeneity the next sub-section is focused on local (non-stationary) statistical models having spatially varying relationships between response and explanatory variables.

4.2.2 Linear Regression with Spatial Effect

In case of linear regression analysis, influence of the indicators remain constant over the response variable throughout time and space. But the weight of an observation might not be constant in the calibration, in fact it varies with distance. As a result, observations closer to a point of event can have a stronger influence on the prediction.

Geographically Weighted Regression (GWR)

The basic concept behind GWR is to analyze how the relationship between a response variable and one or more explanatory variables might vary geographically (Nakaya et al., 2005). GWR is a modified extension of classical regression modeling (Brunsdon et al., 1998). It follows the same framework of multivariate linear regression as mentioned in Equation 4.1. But GWR adds the relationships between the response and explanatory variables to vary by locality. In fact, GWR constructs a separate linear regression equation for individual location within the bandwidth of each target location. The fundamental mathematical framework of GWR for continuous response variables can be represented as:

$$y_i(u) = \beta_{0i}(u) + \beta_{1i}(u)x_{1i} + \beta_{2i}(u)x_{2i} + ... + \beta_{pi}(u)x_{pi} \tag{4.4}$$

where y is the response variable and $\beta(u)$ is the vector of the location-specific parameter estimate. $u(u_i, v_i)$, represents the geographic coordinates of location i in space. $x_k, k = 1...p$ is a set of p explanatory variables. The parameters in GWR are estimated by weighted least squares calculated on the basis of their proximity to any event in location i. The weighing matrix w_{ij} is calculated using kernel density function. A common Gaussian weighing function can be mathematically expressed as:

$$w_{ij} = \exp(-\frac{d_{ij}^2}{h^2}) \tag{4.5}$$

where h is the bandwidth which controls the smoothness of the estimates and d_{ij} is the distance between location i and j.

Generally two types of weighing techniques are used namely: fixed-kernel and adaptive kernel methods. In the current study, GWR is initially tested with fixed bandwidth using *gwr.basic* function from R package "GWmodel" (Gollini et al., 2015). In case of fixed kernel method, initially, the traffic accident locations are converted to spatial objects and plotted. Figure 4.14 represents the same. This depicts the visual distribution of the sample data set.

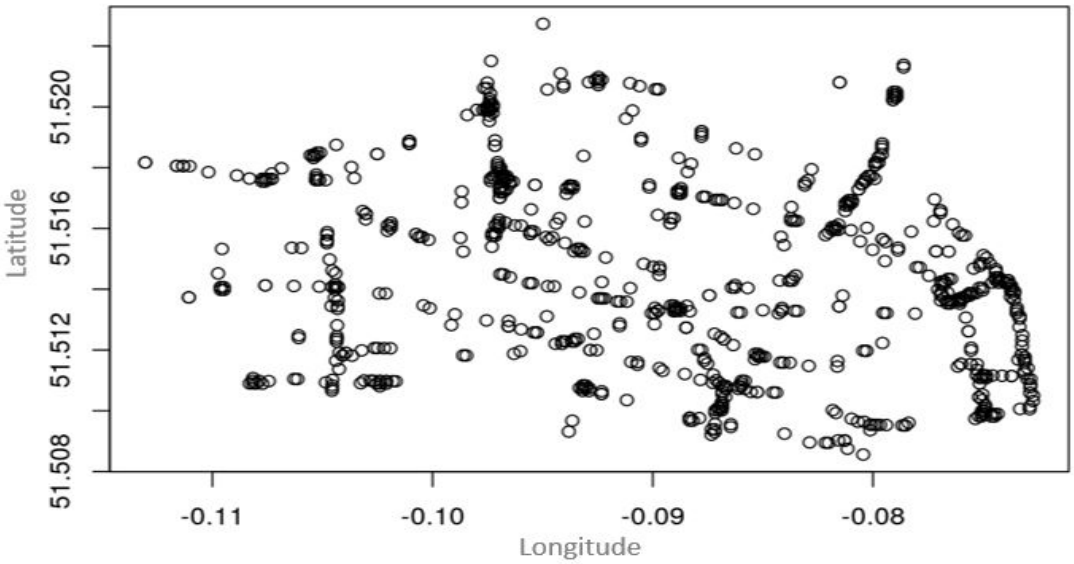


Figure 4.14: Spatial distribution of sampled traffic accidents in the study area

Using the spatial data points a grid of fixed cell size has been created covering the study region. Figure 4.15 depicts the grid covering the entire study area. In the next stage, the grid points are used as the regression points to execute GWR using fixed kernel method.

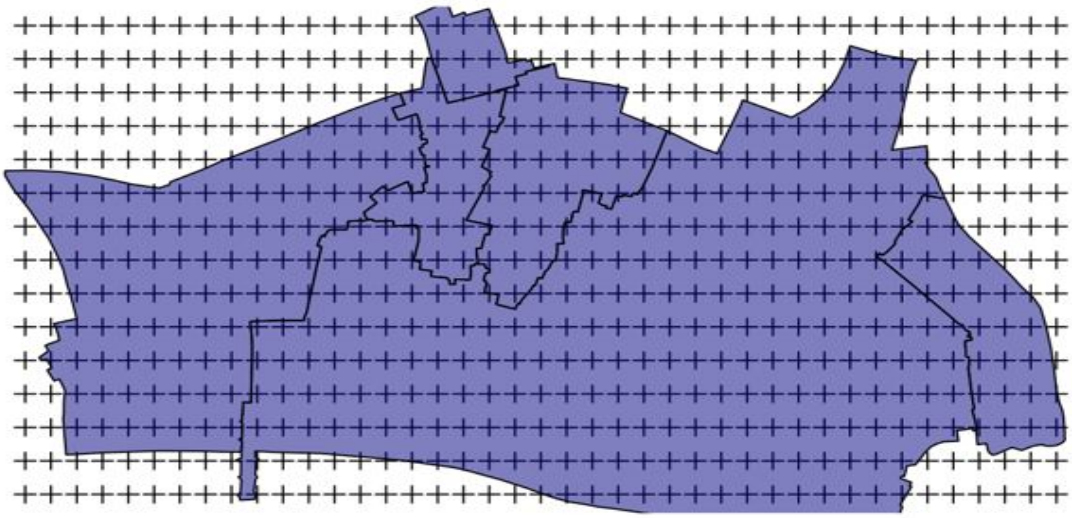


Figure 4.15: Spatial grid over study area

On the other hand, in case of adaptive kernel method, the bandwidth of the space kernel is calculated by using function *bw.gwr* with (adaptive=TRUE, approach=AICc) from R package "GWmodel" (Gollini et al., 2015). The function automatically selects optimal bandwidth that can calibrate a basic GWR model. The calculated bandwidth value is used during GWR model fitting process using *gwr.basic* function from the same R package. Explanatory variables mentioned in Table 4.1 and original discrete values of *number_of_casualties* (as response variable) are fitted in the model with Gaussian kernel function. The model can generate results for both global regression (without considering spatial effects) and GWR together. Table 4.5 illustrates the comparison result of the global regression and GWR model with both fixed and adaptive kernel methods.

Table 4.5: Result: Global regression and GWR

Family	**AIC**	**RSS**
Global	893.2707	132.1712
GWR (fixed kernel)	1143.181	178.3772
GWR (adaptive kernel)	851.5098	131.2478

**AIC: Akaike information criterion*
**RSS: Residual sum of squares*

Lower AIC value in Table 4.5 suggests GWR model using adaptive kernel method to be a better fit than ordinary global linear regression as well as GWR using fixed kernel method. This indicates the existence of spatial heterogeneity in the sample data set.

Coefficients of individual variables can be plotted to observe how the relationship between the response and each explanatory variable varies across space. Figure 4.16 is an example plot of basic GW regression coefficients estimates for one variable *speed_limit* used in the GWR model.

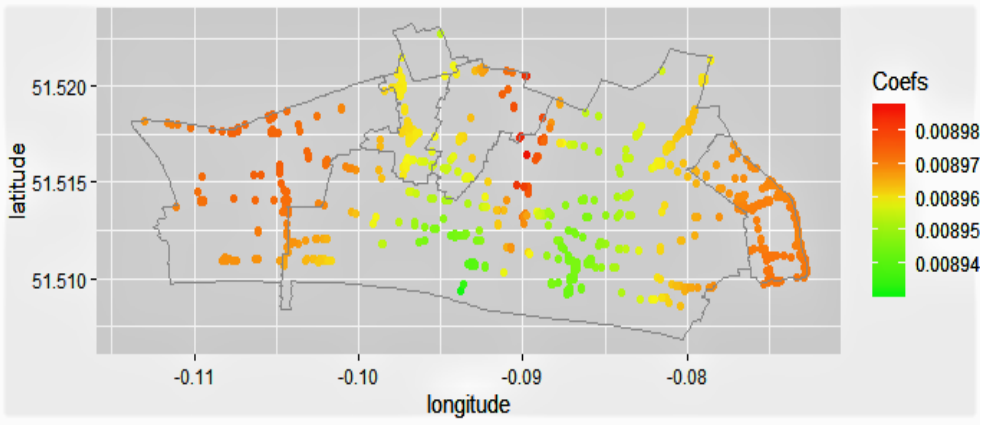


Figure 4.16: Basic GW regression coefficient estimates for *speed_limit*

Spatial Points Data Frame (SDF)-variables from *gwr* objects obtained from GWR results are further analyzed to identify existing multi-collinearity of the

variables. Both correlation analysis and pairs plot (for visual comparison) are performed. For adaptive kernel method, to analyze significant (spatial) variability of the model's parameters or coefficients Monte Carlo test (via GWmodel function *monte-carlo.gwr*) has been performed. Summary of GWR coefficient estimates and result of Monte Carlo test (for adaptive kernel method) is reported in Appendix A.2. The results indicate stationary spatial impacts of the variables. Literature (Nakaya et al., 2005; Pirdavani et al., 2014) suggest use of Geographically Weighted Logistic Regression (GWLR) and Geographically Weighted Poisson Regression (GWPR) to further explore spatially varying effect in generalised linear modeling (GLM). GWR 4.0 software (Nakaya et al., 2009) also provides user-friendly platform for geospatial researchers to analyze spatial variability of the coefficients.

Before investigating further, in this context it is noteworthy to mention that random spatial events like traffic accidents are irregularly scattered point patterns and literature (Juan et al., 2012; Karaganis & Mimis, 2006; Loo et al., 2011) show spatial point process models are useful tools to perform spatio-temporal analysis on this type of data set. Moreover, recent research works (Galgamuwa et al., 2019; Moradi & Mateu, 2019; Moradi, 2018) on spatio-temporal point processes over linear networks aids in identifying spatial-autocorrelation and to detect patterns with interactions between points in the pattern. Thus, the rest of the current study has explored spatial point process analysis and proposed spatio-temporal modeling of traffic risk mapping using Bayesian methodology with INLA-SPDE precisely on road networks. GWPR and GWLR can be investigated further in future research works. But initial exploratory data analysis and GLM (spatial and non-spatial) have refined the selection of explanatory variables that will be used in spatial point process analysis.

Chapter 5

Theoretical Framework and Model Building

This chapter presents a brief introduction to the theory of spatial point pattern analysis and Bayesian methodology with INLA and SPDE. It covers the sequential development of the spatio-temporal modeling process. The chapter is structured as follows. In the first section, a general overview of spatial distribution of the data set is given. Detailed methodologies for model design along with model fitting techniques are explored in the second section. The third section explains the methods of assessing the performance of the proposed model. The final section represents the risk map design algorithm and its implementation. Before implementing the geospatial and statistical procedures, related theoretical concepts are discussed in each section of this chapter.

5.1 Spatial Distribution of Data

In this section, the data set is represented in a 2-D plane, to visualize the distribution of traffic accidents in the study region. R package "spatstat" is used to analyze spatial point pattern of the data (Baddeley & Turner, 2005). Yearly accident data are plotted for the study period of 2005 to 2014. Figure 5.1 depicts the distribution of accident occurrence over the study region for the years 2005 and 2012 respectively. After analyzing individual year plot, it is noted that accident records from 2005 to 2012 are clustered only in two regions of the study area and are unexpectedly less with respect to the accident records from the remaining data set. In fact, 69.91% of the data are from the last two years (2013 and 2014) of the training data set.

Figure 5.1: Spatial distribution of sample traffic accident locations in 2005 and 2012

The data is cross-checked with the UK police department official portal (Transport for London, 2019) and with "stats19" (Lovelace et al., 2019) data source. But the data set is found to have same clustering issue for the mentioned time period (2005 to 2012). Thus, it is decided to discard the data for the unreliable time period. Traffic accidents data from 2013 and 2014 is considered as the updated training data set. Figure 5.2 represents the spatial distribution of training data set (2013-2014) in the study region.

Figure 5.2: Spatial distribution of sample traffic accident locations in 2013 and 2014

In the next step, kernel density plots are created to have comprehensive view of the distribution of training data set. As the traffic accidents are likely to happen mainly on the road network, kernel density plot over the complete study area as well as only on the road network are generated separately. In case of kernel density over the whole study area, objects of class *ppp* are created which represent point pattern data set in two-dimensional plane. On the other hand, for kernel density over the road network, objects of class *lpp* are created which represent point pattern only on the road network (Baddeley & Turner, 2005).

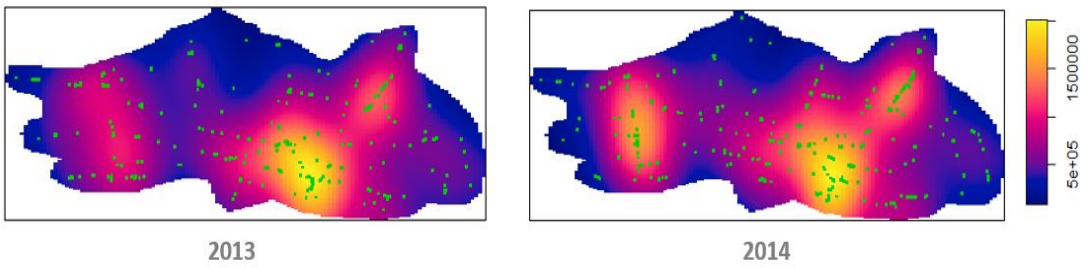


Figure 5.3: Kernel density plot of traffic accidents over study area

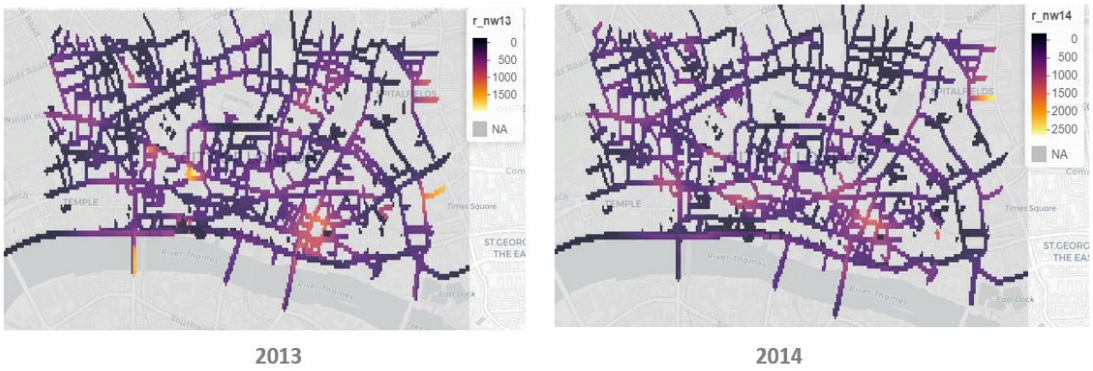


Figure 5.4: Kernel density plot of traffic accidents precisely on road network in the study area

Figure 5.3 and Figure 5.4 represents the kernel density plot over the whole study area and only on the road network respectively. Literature states that, spatial point processes are important analytical tool to investigate the spatio-temporal distribution of traffic accidents (Karaganis & Mimis, 2006). The next section of the research work is focused on spatial point process analysis with emphasis on Bayesian modeling approach.

5.2 Spatio-temporal modeling: Integrated Nested Laplace Approximation (INLA)

Literature on factors contributing traffic crash shows, most of the attempts at predicting occurrence of traffic accidents depend on spatio-temporal interacting and triggering factors (Liu & Sharma, 2017). At the same time, on the ground of spatial and temporal scales, individual accidents are considered as discrete points. Thus, it can be stated that traffic crashes are associated with their spatial coordinates, the time of occurrence and other corresponding covariates. The data can be realized as stochastic process indexed by:

$$Y(.) = \{y(s_i, t_i) \in R^2 \times R\} \tag{5.1}$$

where s_i represents spatial coordinates and t_i is for temporal instant, with both of them $Y(\cdot)$ is a spatio-temporal process defined in a subset of $R^2 \times R$. Based on these facts, the present research work is focused on Bayesian modeling framework

for the prediction of spatio-temporal occurrence of traffic accidents. Moreover, Bayesian approach facilitates representing the uncertainties related to models and inference of parameter values with the ability to incorporate prior information (Congdon, 2014; Dunson, 2001). Traditionally, Bayesian approach with Markov Chain Monte Carlo (MCMC) simulation methods can be used to fit generalized linear mixed model (GLMM) (Wikle et al., 1998). In particular, MCMC methods provide multivariate distribution that can estimate the joint posterior distribution. But there are many analysis where only marginal inference on selected parameters are required. Prior research works suggest that, for latent Gaussian models (Rue et al., 2009) and models having higher number of geo-locations (Musenge et al., 2013) the performance of MCMC methods drop significantly. Thus, a new prediction of the marginal distributions by using Laplace approximation for the integrals, is introduced by integrated nested Laplace approximation (INLA) (Rue et al., 2009). It is practically dealing with (many) univariate distributions using numerical integration techniques instead of Monte Carlo sampling and thus can have computational advantage over MCMC process. For approximation Bayesian inference, INLA is the alternative to traditional MCMC methods and it focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF) for their computational properties (Rue & Held, 2005). Principal advantages of using INLA over MCMC methods (Verdoy, 2019), are the following:

- Low computation time.
- As the basic logic is Bayesian inference, it does not require only normally distributed data set.
- Can implement both spatial and temporal effects.
- Can analyze significance of spatial and temporal effects in the model.
- Allows integrating substantially high number of covariates.
- Allows integrating new covariates at later stage of the process.
- Level of significance for each covariate can be analyzed.

Another important technique (Cameletti et al., 2012; Rue & Held, 2005) followed in this research work is the conversion of continuous scale Gaussian function (GF) to a discrete scale Gaussian Markov Random Field (GMRF) using Stochastic Partial Differential Equations (SPDE). SPDE approach has proved to be a powerful strategy for modeling and mapping complex spatial occurrence phenomena (Cameletti et al., 2012). Thus, the current study has conducted the analysis of spatial point processes by implementing INLA approach with an explicit link between GF and GMRF using SPDE. In statistical analysis, to estimate a general model it is useful to shape the mean for the additive linear predictor, defined on a suitable scale:

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m z_{mi} + \sum_{l=1}^{L} f_l(v_{li}) \tag{5.2}$$

where, β_0 is a scalar, which represents the intercept, $\beta = (\beta_1, ...\beta_M)$ are the coefficients of the linear effects of the covariates $z = (z_1, ...z_M)$ on the response, and

$f = \{f_1(.), ...f_L(.)\}$ is a collection of functions defined in terms of a set of other covariates represented as $v = v(v_1, ..., v_L)$, different from the previous covariates. In the formula above, the first function used in the current study is SPDE used to analyze the spatial effect with the Matérn covariance function (Musenge et al., 2013). The other explanatory function used is a temporal random variable function. Random walk model of the order one (RW1) is applied in the method. The current study is implementing these three-stage process to develop hierarchical Bayesian models. The rest of this section is organized as depicted in Figure 5.5. R package "INLA" (Martins et al., 2013) have been extensively used for the modeling phase.

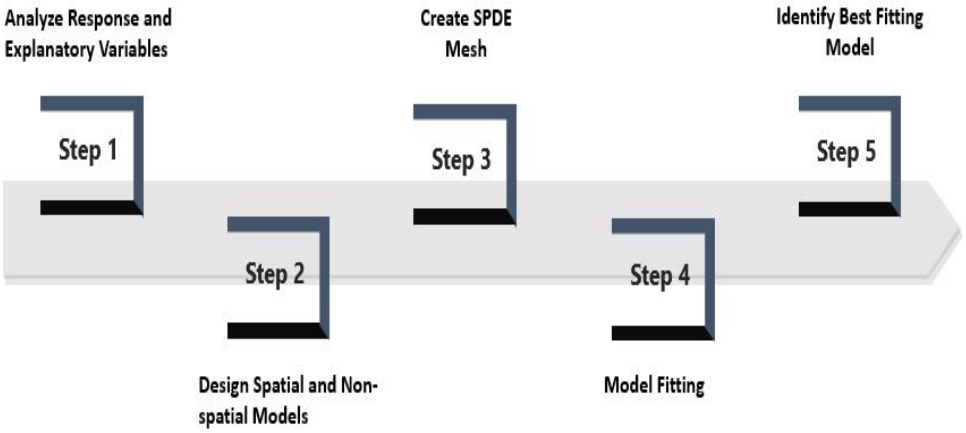


Figure 5.5: Workflow diagram: INLA-SPDE modeling phase

5.2.1 Analysis of Response and Explanatory Variables

In the spatio-temporal hierarchical Bayesian analysis the same set of eleven explanatory variables (mentioned in Table 4.1) are considered as covariates and *number_of_casualties* as the response variable (Table 4.2). In the collinearity study of the covariates (performed in Section 4.1) no pattern was detected. On the other hand, Figure 5.6 illustrates the frequency distribution of the response variable in the current data set.

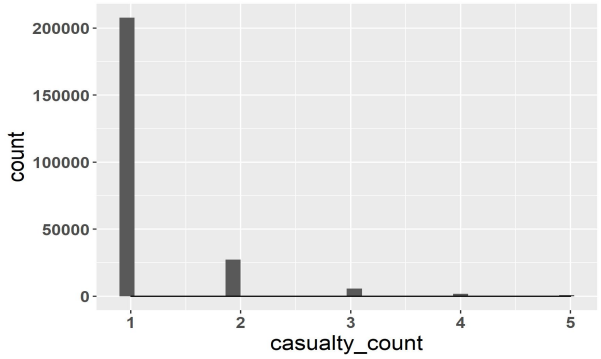


Figure 5.6: Frequency distribution of response variable

The highest frequency value (89.86%) is when the number of casualties is 1. Thus, the model is fitted with "*gamma*" control family.

5.2.2 Spatial and Non-spatial Model Design

Based on the principal objectives of the study and the sampled covariates, classes of spatio-temporal combinations are created to design the model. The combinations of spatio-temporal covariates and type of SPDE mesh are summarized into the following models referred in Table 5.1.

Table 5.1: Spatial and non-spatial training models

No	Model Code	Spatial Effect	Temporal Effect
1	M1	-	-
2	M2	-	RW1
3	M3	Region mesh	-
4	M4	Network mesh	-
5	M5	Region mesh	RW1
6	M6	Network mesh	RW1

**RW1: Random walk model of order 1*

5.2.3 Design SPDE Triangulation

INLA is computationally efficient because it uses SPDE to estimate the spatial autocorrelation of the data. But this requires using a "triangulation" or "mesh" of the discrete event locations interpolated to estimate a continuous process in space (Rue et al., 2017). Due to densely distributed nature of the road segments in the study area, initially a continuous spatial structure was chosen for modeling and triangulation was carried out on the entire study area. Figure 5.7 depicts the accident locations for the training years 2013 and 2014 within the non-convex hull boundary.

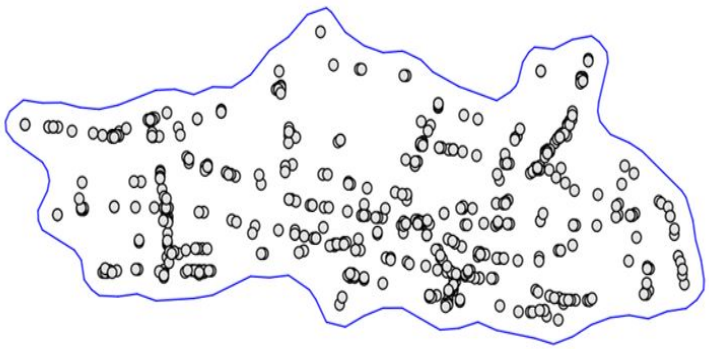


Figure 5.7: Traffic accident locations within the non- convex hull boundary

Triangle size (generated using a combination of maximum edge and cutoff) controls how precisely the equations will be tailored by the data. Using smaller

triangles increases precision but also exponentially increases computing power (Verdoy, 2019). The best fitting mesh should have enough vertices for effective predictive analysis, but the number should be within a limit to control the processing time. Following this concept, a series of meshes with or, without boundary and varying the number of vertices are created. Figure 5.8 shows examples of two constrained refined Delaunay triangulation, one designed without offset another with offset. In both the cases the number of vertices is *785* but one with offset value while the other is without offset value.

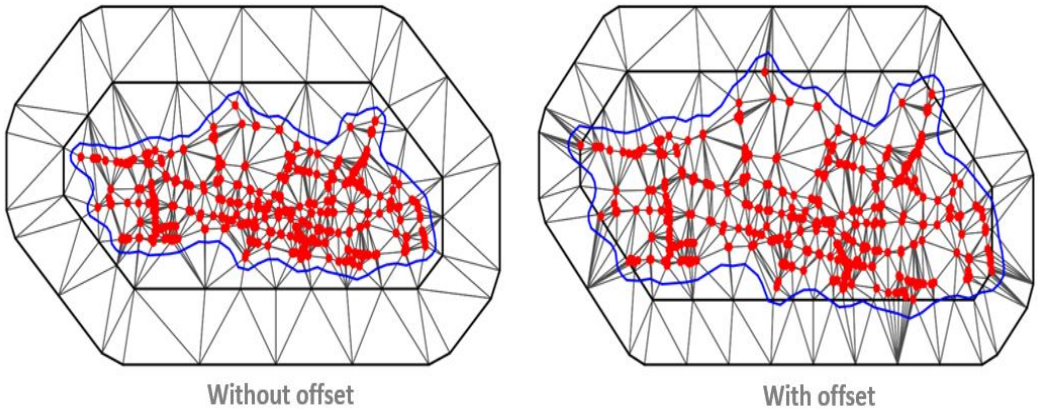


Figure 5.8: Region Mesh elements with vertices and sample points a) without offset b) with offset

Finally, the best fitting mesh without offset value and having non-convex hull boundary is selected. The number of vertices for this standard mesh is also *785*. Figure 5.9 depicts the selected mesh (with the training data accident locations) to be used for SPDE model in the current study.

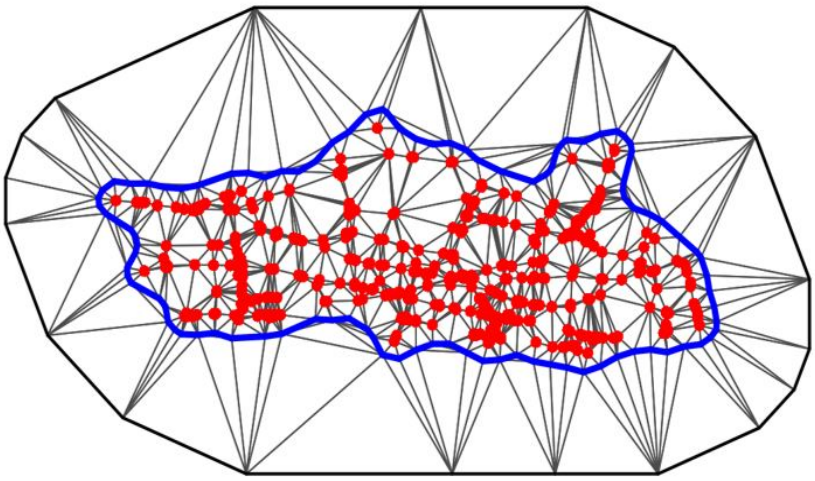


Figure 5.9: Selected region mesh with non-convex hull boundary

Introduction to SPDE *Network Triangulation* The mesh created for the entire study area is used for fitting the INLA model in the region. Prediction

involves projecting the fitted model into the mesh at precise spatial locations. While fitting the mesh a problem was noted. The sampling points (here the accidents locations) are mainly located on the road network, but the mesh was generated for the whole study area which includes road network as well as other regions. Thus, the model result might be unpreventably generalised as it is going to estimate predicted value for the regions where there is no chance of incident to happen. The next sub section introduces the novel concept of creating SPDE triangulation precisely on the road network. The steps followed to implement INLA-SPDE only on road network are illustrated in Figure 5.10.

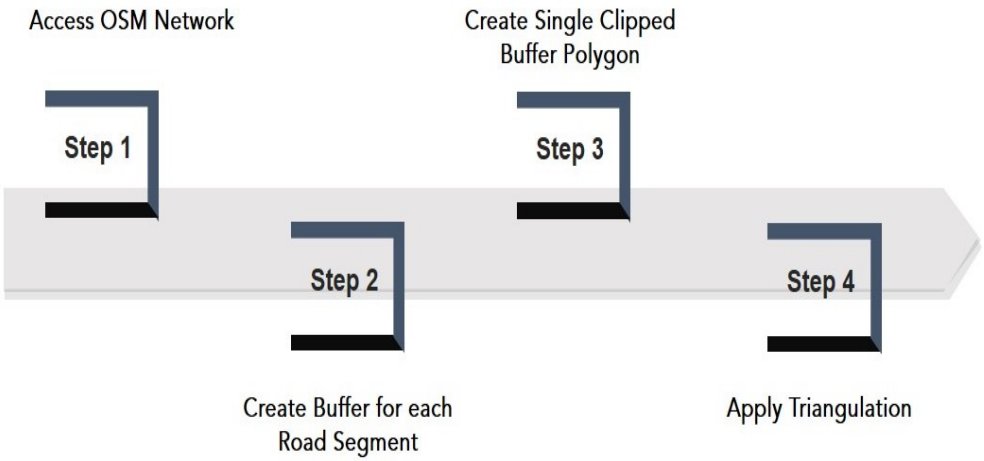


Figure 5.10: Workflow diagram: Network mesh creation

Access OSM Network OSM road network for the study region has been accessed as "sp" object using R package "osmdata" (Padgham et al., 2017). Figure 5.11 represents selected OSM road network of the study area. OSM highways of type *unclassified, bus_guideway, raceway, bridleway, path* are not used in the current study.

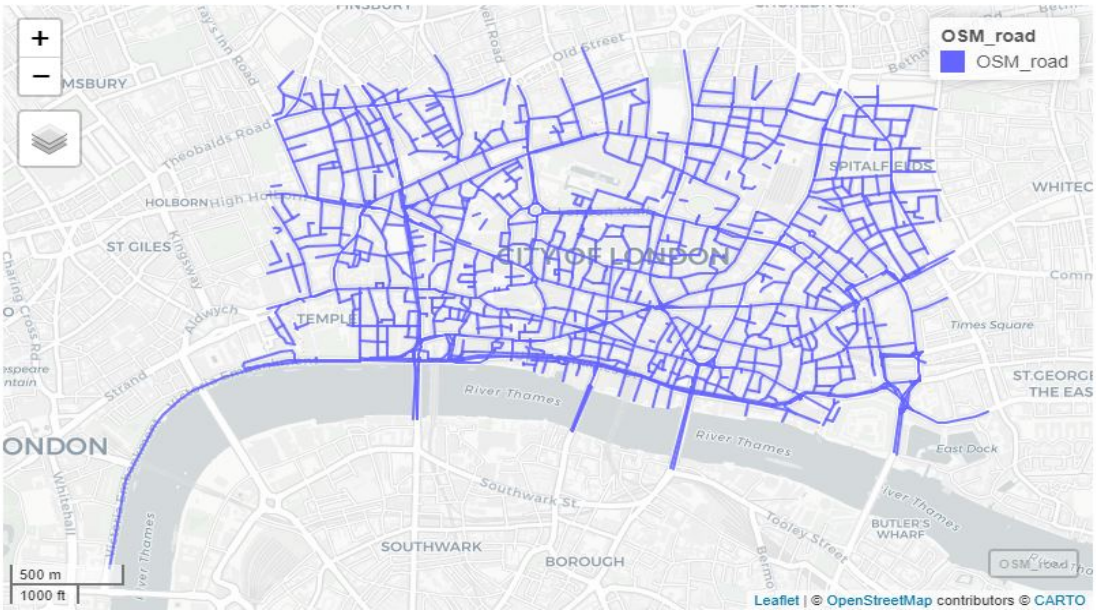


Figure 5.11: OSM road network of the study region

Buffer for each Road Segment Previous studies (Amin et al., 2014; and, 2003) have shown accuracy and reliability issues of positional data in transportation research works. There are instances where recorded data entry invariably introduces errors in both geometric and contextual attributes (Miler et al., 2016). The road traffic accident data used in the current study is compiled and maintained by the Department of Transport, UK (mentioned in Section 3.3.1). The accident locations along with the extracted OSM road networks are plotted in the same projection using QGIS (QGIS Development Team, 2009) software. Glimpse of sampled traffic accident locations (marked as red points) plotted with OSM road network is depicted in Figure 5.12 (a). These spatial locations are important components in generating the proposed SPDE network triangulation. But most of the points are identified to be located away from the road segments. In the next step, Figure 5.12 (a) and (b) depicts how the function *buff_geo* from R package "stplanr" (Robin Lovelace & Richard Ellison, 2018) has been used to identify all the traffic accidents that took place within a specified buffer region of each road network.

Figure 5.12: Traffic accident locations on road segments a) with and b) without buffer

According to (Verdoy, 2019), the best fitting SPDE triangulated mesh should have enough vertices for effective predictive analysis, but the number should be within a limit to control computational time. To achieve the optimal mesh in later phase, a series of different buffer size are applied on the OSM road segments. Examples of SPDE network mesh with buffer size 10 meter and 30 meter are illustrated accident locations in Appendix B.4. The buffer size applied in the current study is of 20 meter. For each road segment on the entire road network of the study area a buffer of size 20 meter is created. Figure 5.13 depicts OSM road network with *20 meter* added buffer.

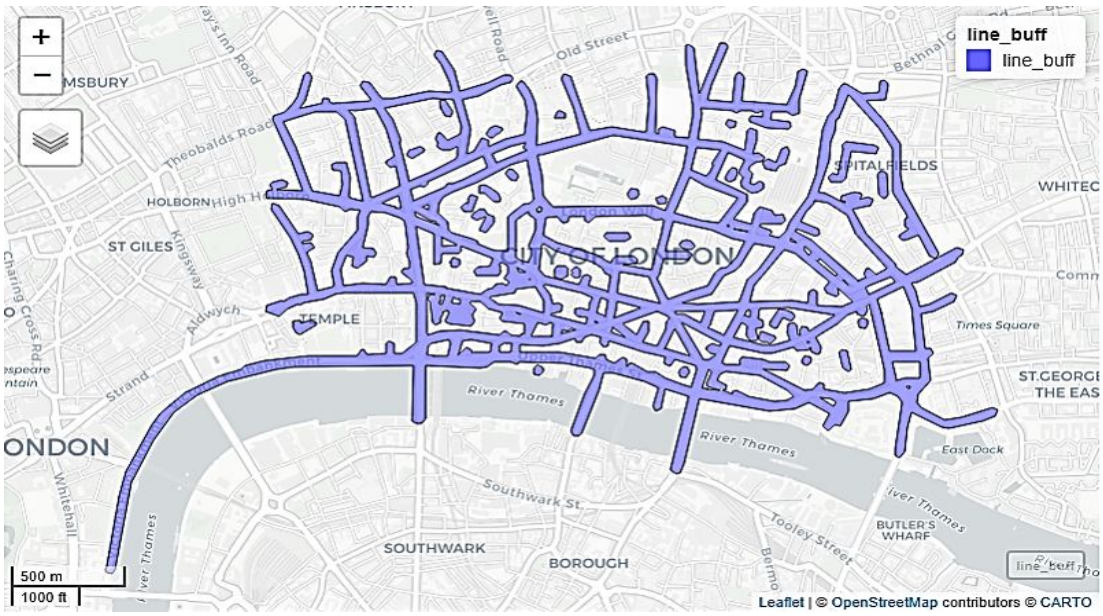


Figure 5.13: OSM road network with 20 m buffer

Create Clipped Buffer Polygon Individual buffer segments are merged and converted as a single polygon clipped within a bounding box covering the study area. Figure 5.14 illustrates the polygon of the buffered segments.

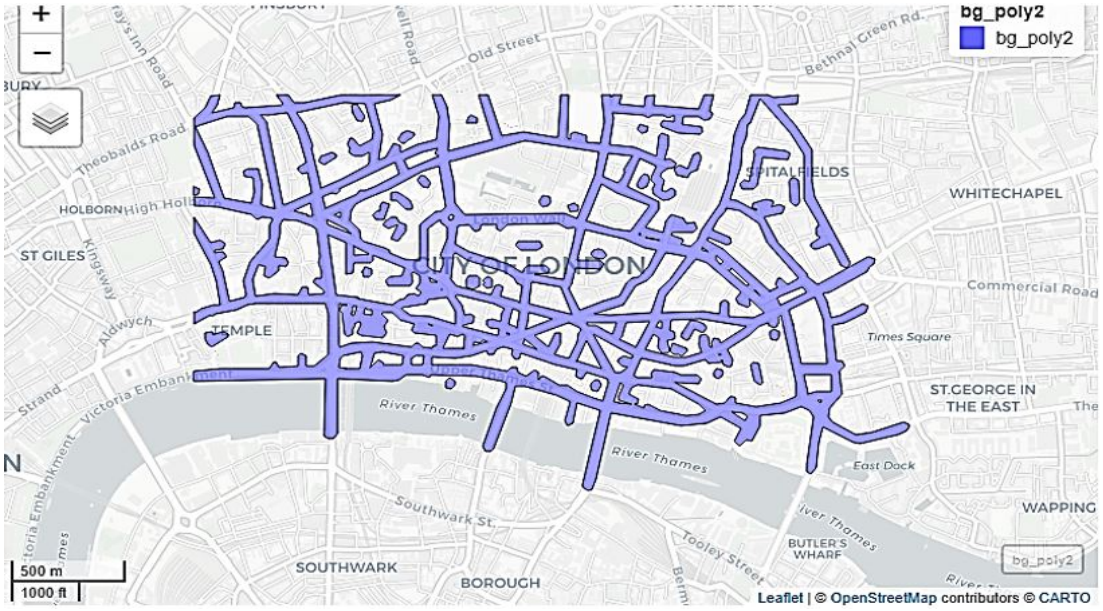


Figure 5.14: Buffer polygon clipped within bounding box of study region

Apply Triangulation Triangulation is applied on the result polygon. A series of SPDE-mesh are generated (examples are depicted in Appendix B.4). From them the best fitting mesh having buffered polygon boundary is selected. Thus, the mesh is now created only on the road network as depicted in Figure 5.15. The number of vertices for the final selected mesh is recorded as *8412*. From the figure it can be identified that; the network mesh contains almost all the

accident locations (highlighted as red marks). Only few points are found to be lying outside the mesh structure.

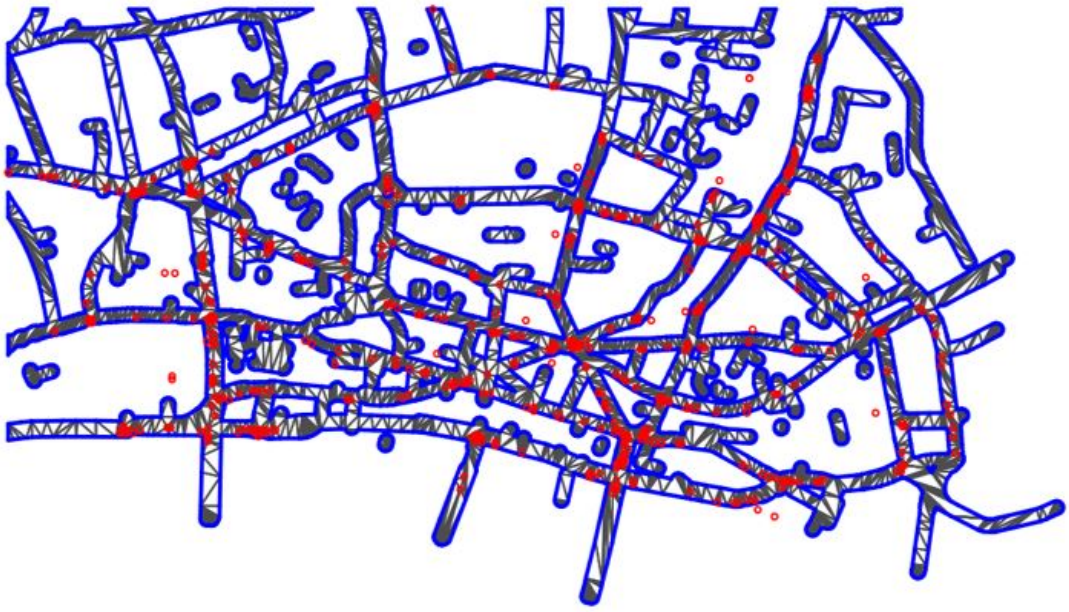


Figure 5.15: Selected network mesh with traffic accident locations

5.2.4 Model Fitting

R-INLA (Martins et al., 2013) is used to fit all the SPDE models (mentioned in Table 5.1). R code for INLA was adopted from previous studies and modified for the current analysis (Zuur et al., 2017). All the six models are executed separately for the same training data set (2013-2014) and *gamma* control family. Respective computational time for individual model is shown in Table 5.2. All the models are executed using i7 4790 processor (mentioned in Section 2.1) and Linux operating system.

Table 5.2: Computational time of individual training models

Model Code	**Comp. Time (in seconds)**
M1	1.6
M2	3.83
M3	55.1
M4	99.7
M5	136
M6	312

Except model M1 and M2 all models are having spatial effects. Significant

difference in computational time is noted between the models with and without spatial effects.

5.2.5 Identifying the Best Fitting Model

Deviance information criterion (DIC) and the Watanabe-Akaike information criterion (WAIC) are used to assess models and to select the best suitable model by balancing model accuracy against complexity (Spiegelhalter et al., 2002). Models having smaller DIC value, suggest that, in spite of the added complexity, it has a more appropriate fit to the sampled data (Blangiardo & Cameletti, 2015). Conditional predictive ordinate (CPO) value (Gelfand et al., 1992) also acts as a selection measure; smaller value of CPO indicates a better prediction quality of the model. But before comparing DIC values, parameters for each model are examined. Table 5.3 represents the precision parameters for gamma observations.

Table 5.3: Precision parameters for gamma observation

Model Code	**Mean[0.025quant, 0.975quant]**
M1	15.51 [13.87, 17.25]
M2	16.55 [14.75, 18.47]
M3	22.61 [22.36, 23.09]
M4	20.30 [19.97, 20.77]
M5	24.68 [24.36, 24.98]
M6	20.90 [20.74, 22.78]

Summary results (DIC, WAIC and CPO) related to goodness-of-fit for all the fitted models are reported in Table 5.4.

Table 5.4: Training models: DIC, WAIC and CPO values

Model Code	**DIC**	**WAIC**	**CPO**
M1	239.60	285.84	-0.406851
M2	211.16	259.54	-0.356291
M3	273.63	203.37	-2.395325
M4	267.68	261.22	-0.868227
M5	202.85	234.34	-0.801388
M6	216.76	280.28	-0.320038

DIC values of the models shown in Table 5.4 suggest the models with both spatial and temporal random effect provide better model fit. Models M5 and M6

can provide better precision though the computational time for both the models are significantly high compared to others. The selection criterion reported in Table 5.4 indicates model M5 is the best fitting model.

At this point, it is noteworthy to mention that, the sampled traffic accidents are discrete spatial points located only on the road networks. But the mesh fitted in model M5 is for the entire study area. Therefore, the predicted result locations can occur in any area with or without road networks. It is not practical that the model prediction will provide results in locations without road network where there is no chance of traffic accident occurrence. On the other hand, the mesh for model M6 is specially designed to solve the problem being faced in M5. The DIC value of M5 (*202.85*) and M6 (*216.76*) do not vary considerably. Thus, from the practical aspect of spatio-temporal prediction of traffic accidents on road networks, M6 will be a better fit than M5. The rest of the current research work is organized considering M6 as the best fitting model to enable prediction precisely on the road network of the study area.

5.3 Model Validation and Prediction

Model Validation The selected model (M6) has both non-spatial and spatial effects. The model validation is conducted by exploring the hyper-parameter values and marginal posterior distribution of the spatial effects. Trend of the random walk model is analyzed to interpret the temporal effect of the proposed model. In INLA result the fixed parameters are the regression parameters and the hyper-parameters are variance-type parameters. But, R-INLA basically works with precision and not with the variance (Zuur et al., 2017). INLA hyper-parameter values are represented in Table 5.5.

Table 5.5: Hyperparameter values of selected model

Hyperparameter	**Posterior Mean**
rw_date	45476.36
κ	211.7287
sigma$u$	0.394468
range	0.013357

Validation of the model is performed by comparing residuals between real data and the output of the executing model. Figure 5.16 depicts the residual diagnostics and the relationship with the distance. The correlation in this case, p= *0.7900711*, suggests a strong spatial correlation which decreases gradually with distance.

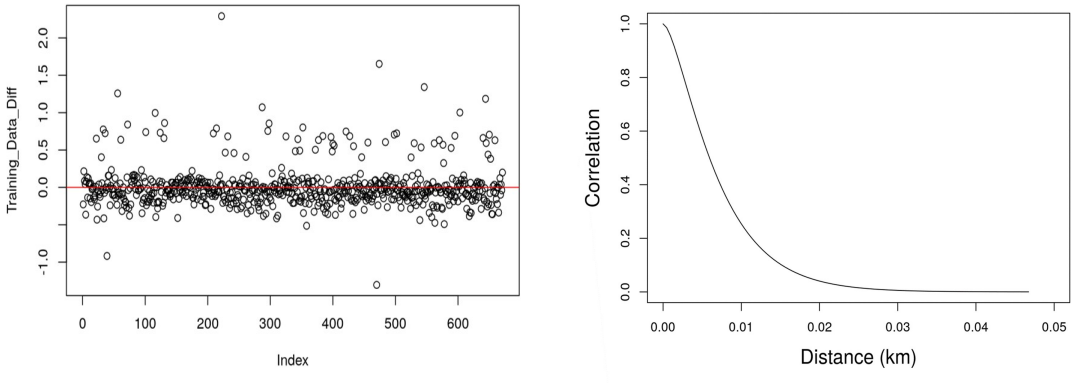


Figure 5.16: Residual diagnostics and correlation plot of selected model

With respect to the hyperparameters, the model contains a standard deviation parameter σ that is basically used for the variance σ^2 of the normal distribution. But as mentioned earlier R-INLA works with precision τ measured as $\tau = 1/\sigma^2$ (Zuur et al., 2017). In the current model four precision parameters are calculated. The marginal posterior distribution for $\tau_{\theta 1}$ and $\tau_{\theta 2}$ are depicted in Figure 5.17 and σ_ϵ and σ_v are depicted in Figure 5.18.

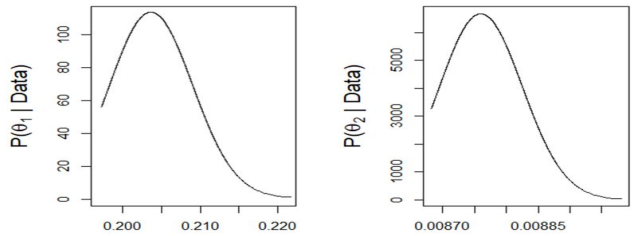


Figure 5.17: Marginal posterior distribution for $\tau_{\theta 1}$ and $\tau_{\theta 2}$

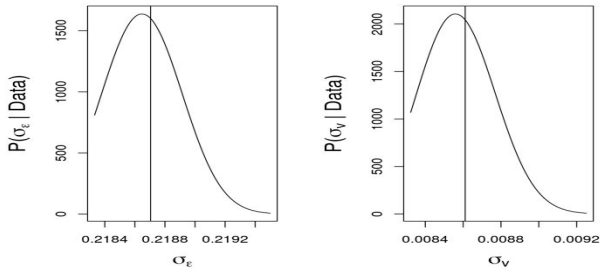


Figure 5.18: Marginal posterior distribution for σ_ϵ and σ_v

Figure 5.19 shows the estimated random walk trend (in blue) and 95% credible intervals. The trend in the plot on both sides of the zero line with fluctuating values supports the inclusion of temporal effect in the model design.

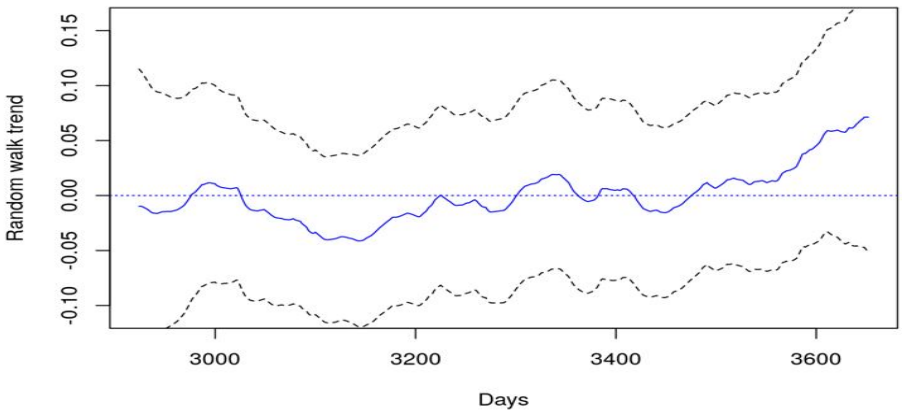


Figure 5.19: Estimated random walk trend

Model Prediction The output of INLA model provide posterior estimates of fitted values for observations, but in practice a model should be capable to predict where no observations are available. The simplest prediction strategy in R-INLA is to add response variable values with *NA* for unsampled locations (Zuur et al., 2017). But the covariates in the prediction model cannot have null values. Thus, prediction involves projecting the fitted model into the mesh at specific spatial locations of interest having demarcated predictor values. In the current study, the model is developed using the training data set (2013-2014). To assess the performance of the model, test data set (2015-2017) has been used. Next, the model is fitted for individual test year combined with the entire training data set. For example, when the test year is 2015 the model is fitted combinedly for 2015 along with 2013 and 2014. In each case the prediction results are analyzed and interpreted.

It is noteworthy to mention here that, the fitted model while executing spatio-temporal prediction has been updated from *gamma* to *Gaussian* control family. It is obvious that, the model fitted with spatial and temporal covariates can generate any real number value as predicted output. Thus, instead of constraining the result in the range of discrete input parameters, the model control family is updated to have predicted result as continuous numerical values.

5.4 Risk Map Design

One of the main objectives of the current study is to model a risk map for the entire road network in the study area. The current section will discuss in brief the methodology followed to design it. Steps to design the risk map is depicted in Figure 5.20:

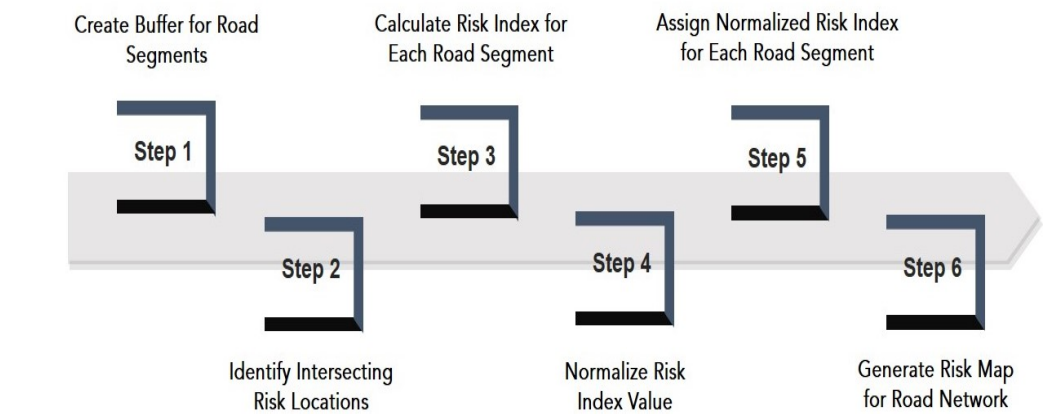


Figure 5.20: Workflow diagram: Risk map design

The proposed model has been fitted for individual test years. In each case, the predicted values are the number of casualties in the sampled accidents locations. Next, a metric system is designed using the predicted values of the test data set. The system will dynamically calculate the risk index for each road segment. Finally, these risk index values are adapted to design the risk map over the entire road network.

Create Buffer for each Road Segment: A 20 meter buffer is created for each road-segment in the OSM road network of the study area (similar technique as followed in Section 5.2.3 while creating network-mesh).

Identify risk locations lying within the buffer region of each road segment: Sampled accident locations are spatial point objects having predicted response value as a parameter. These points are named as *risk locations*. On the other hand, the buffer road segments are considered as polygons. Using intersection technique all the points (risk locations) lying within individual buffer regions are identified. Record of each buffer region with details of each point lying within it, is maintained.

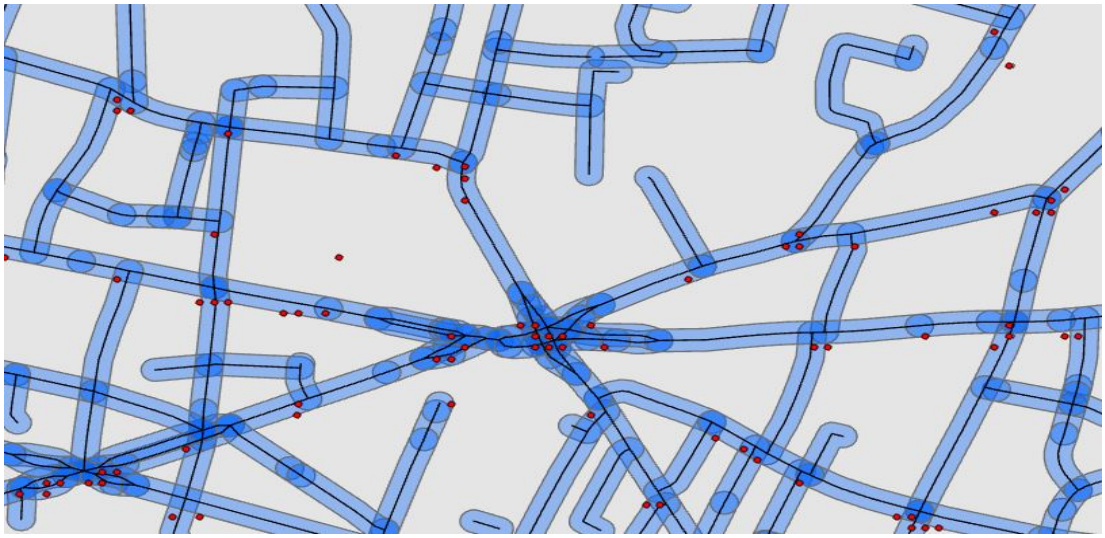


Figure 5.21: Glimpse of risk locations lying within the buffer region

Calculate Risk index value for each road segment (R_i): Figure 5.21 depicts a glimpse of the intersection result. Three categories of buffer segments have been identified from the result. Buffer segments having:

- no risk points
- only one risk point
- two or, more risk points

The risk index for each category have been calculated using the following methodology as mentioned in Table 5.6:

Table 5.6: Measure of risk index

Buffer segment with	**Risk index (R_i)**
No risk points	0
One risk point	Predicted response value of the point inside the buffer
Two or more than two risk points	Mean (predicted response value of all the points inside the buffer)

It is noteworthy to mention that, due to densely distributed nature of the road segments, a risk point can lie on a region where two or, more buffer regions are intersecting as illustrated in Figure 5.22. But the intersection method function (used in the current analysis) will return unique common elements, without having any duplication. It implies that if a risk point lies in common buffer region for more than one road segments, it has been assigned to only one of the segments, selected randomly.

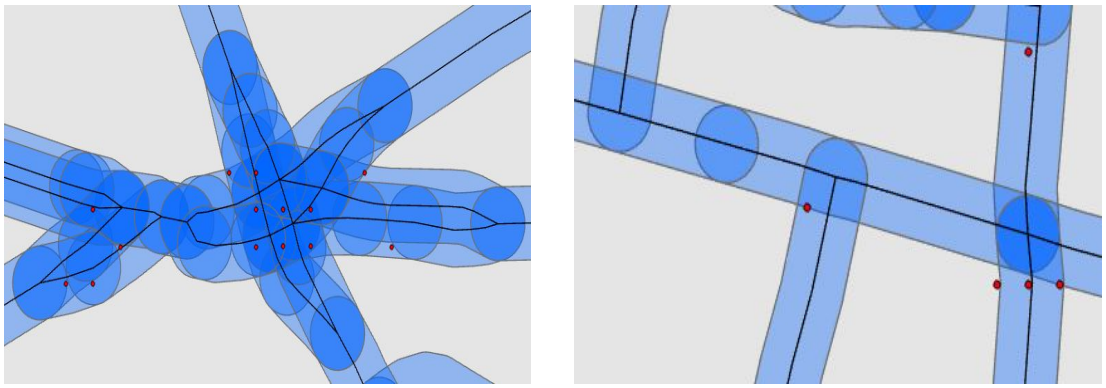


Figure 5.22: Overlapping buffer regions with common risk points

Normalize risk index value: Existing risk map for other road safety research works suggests, a predefined category range to be decided before modeling any risk map (Curran-Everett, 2013). In current analysis to create a category range, the calculated risk index values need to be normalized. The following normalization technique is implemented. Initially the risk range is calculated as follows:

$$\textit{Risk range } (R_{range}) = \frac{(max.R_{mean} - min.R_{mean})}{\textit{no. of groups in response variable field}} \tag{5.3}$$

Next, R_{range} is used to calculate the normalized values. As a relevant example, the values depicted in Table 5.7 shows that, the number of categories in the normalized scale is same as the number of groups in the response variable field. But the method can be replicated with any other updated response variable values and the normalized scale will get updated accordingly.

Table 5.7: Normalization metric for risk index values

Condition	**Nor. value**	**Safety measure**
Segment without having any risk locations	0	Low risk
$R_{mean} < R_{range}$	0	Low risk
$R_{range} \leq R_{mean} < 2 \times R_{range}$	1	Low-medium risk
$2 \times R_{range} \leq R_{mean} < 3 \times R_{range}$	2	Medium risk
$3 \times R_{range} \leq R_{mean} < 4 \times R_{range}$	3	Medium-high risk
$4 \times R_{mean} \leq R_{mean}$	4	High risk

The Safety measure mentioned in Table 5.7 follows the same category used by the European Road Assessment Programme (EuroRAP) to create the risk ratings of the motorways and other national roads in Europe ('Risk Mapping', 2016).

Assign normalized value: Individual road segments are assigned with their respective normalized risk index values.

Generate risk map: Finally, the risk map is designed in interactive geospatial platform using R package "mapview" (Appelhans, 2015).

Chapter 6

Results

This chapter presents and discusses the results of the analysis described in the previous chapter. It is structured as follows. The first section depicts the model prediction results along with residual diagnostics. It also assesses the accuracy of the prediction results. Section two illustrates the risk map for individual test years. Final section highlights few interesting findings and related discussions.

6.1 Model Prediction

The sample data set for 2013-2014 has been used to identify the best model and train it. The best-fitted model has been assessed using test data set from 2015-2017. Figure 6.1 (a), (b), (c) show the residual plots. In all three plots, the residual values are dispersed on both sides of the horizontal zero line and tending to cluster along the line with few outliers. Most of the residual values lying close to zero indicates a high precision of the predicted values for the test data set.

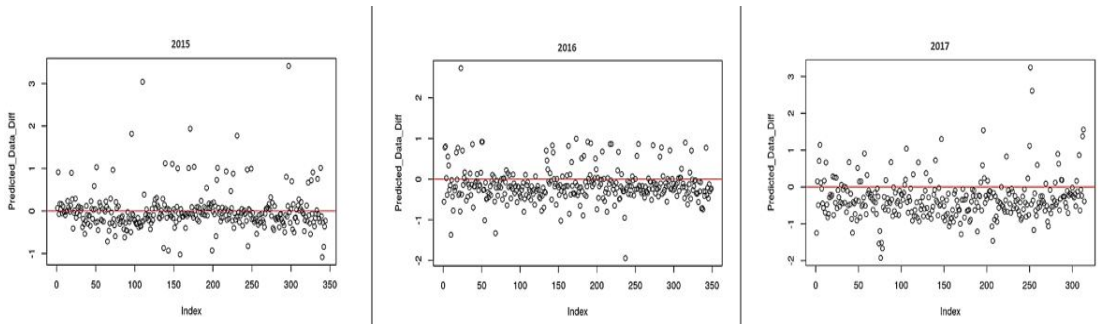


Figure 6.1: Residual plot of test data set a) 2015, b) 2016 and c) 2017

The prediction results for model fitted with individual test year combined with the training years are analyzed. Figure 6.2, 6.3 and 6.4 depict the combined prediction results for test years 2015, 2016 and 2017 respectively. Each figure illustrates the residual plot of combined data and histogram plot for the observed and predicted values. All the residual plots indicate higher accuracy of prediction as the residual values are close along the horizontal zero line. The observed and predicted values are quiet similar can be supported by respective histogram plots.

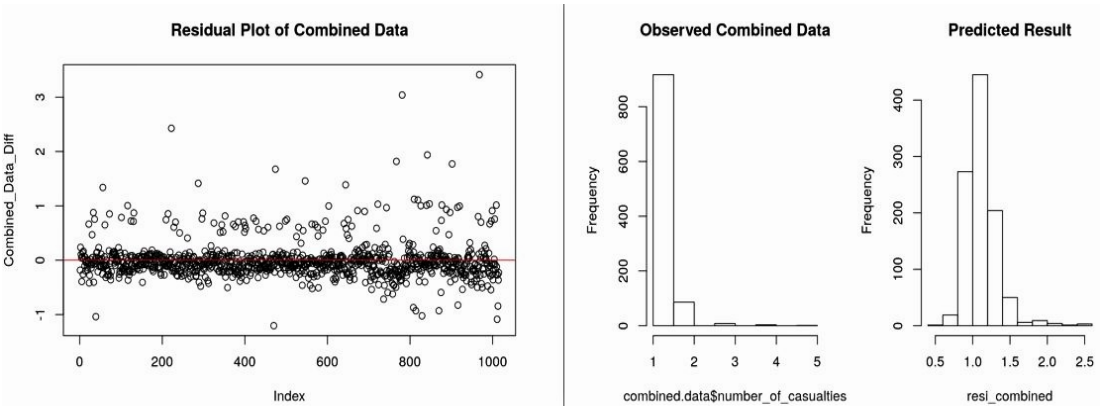


Figure 6.2: Residual analytic and predicted value comparison (2013, 2014 and 2015)

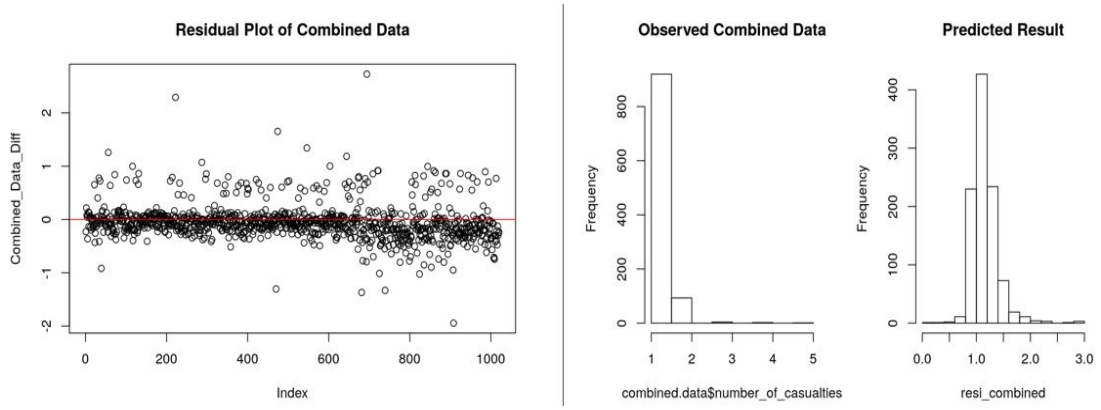


Figure 6.3: Residual analytic and predicted value comparison (2013, 2014 and 2016)

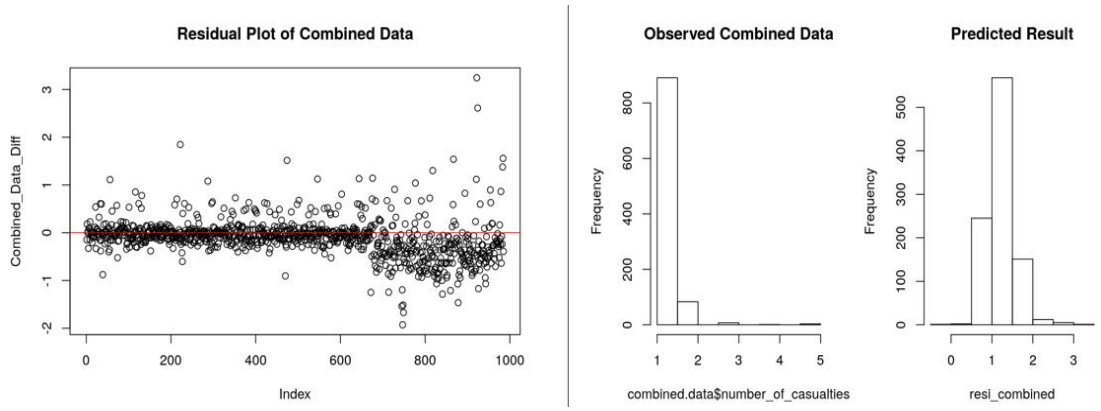


Figure 6.4: Residual analytic and predicted value comparison (2013, 2014 and 2017)

Correlation and root mean square error (RMSE) values act as indicator to assess the performance of the model. Table 6.1 shows the correlation and RMSE values in each test case.

Table 6.1: Model prediction accuracy

Test Year	Correlation	RMSE
2015	0.3860283	1.137029
2016	0.4142525	1.104775
2017	0.3850195	1.259445

From the above residual diagnostics, it can be stated that, the combined prediction result for 2016 is comparatively better than the other two predictions. Further discussion on model performance is reviewed in Chapter 7.

6.2 Risk Map

The methodology used in the current study provides the capability to develop the risk map based on Bayesian analysis, including INLA-SPDE to implement both spatial and temporal effects. Risk index for individual road segment are calculated using the predicted response values of the test data set. The normalized risk index values are used to develop the risk map for the entire road network of the study region.

The risk maps are visualized in interactive geospatial interface using R package "mapview" (Appelhans, 2015). Figure 6.5 to Figure 6.10 illustrates the risk maps and original traffic accident plots respectively for individual test years 2015, 2016 and 2017. The color scale (0 through 4) used in each map follow the same safe measure scale used in Table 5.7.

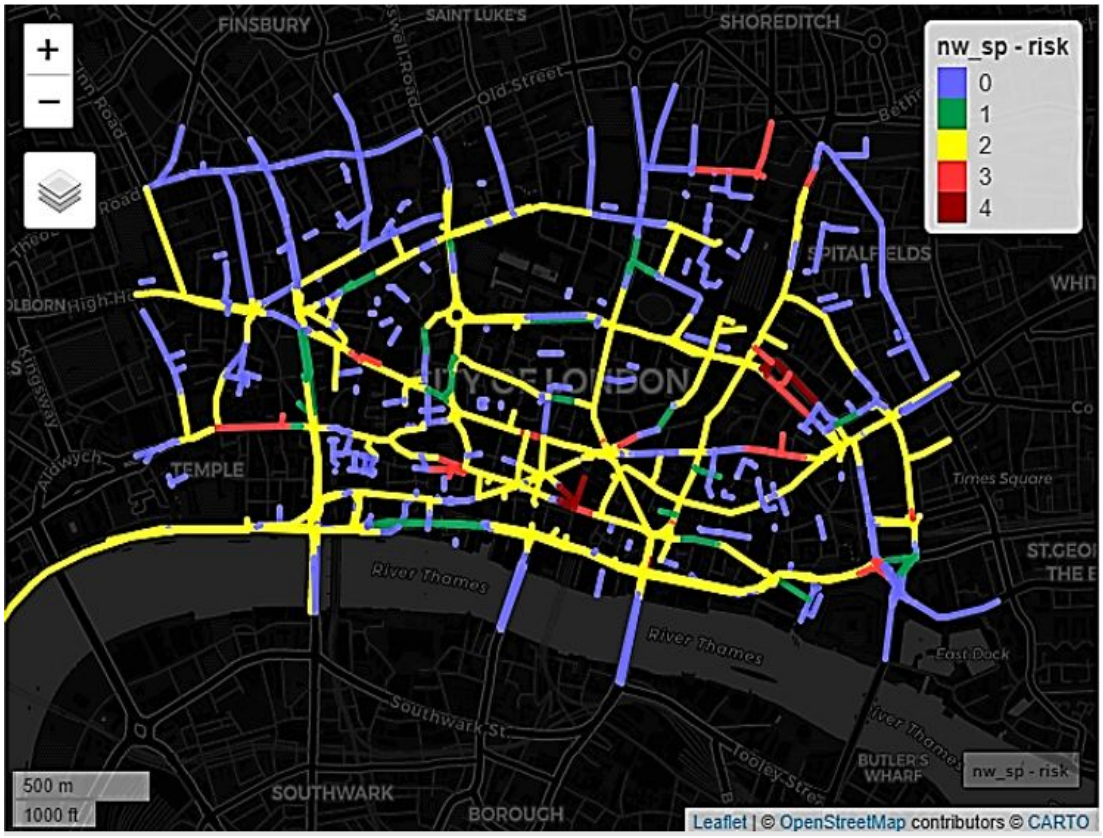


Figure 6.5: Risk map (2015)

Figure 6.6: Original sample data of traffic accident (2015)

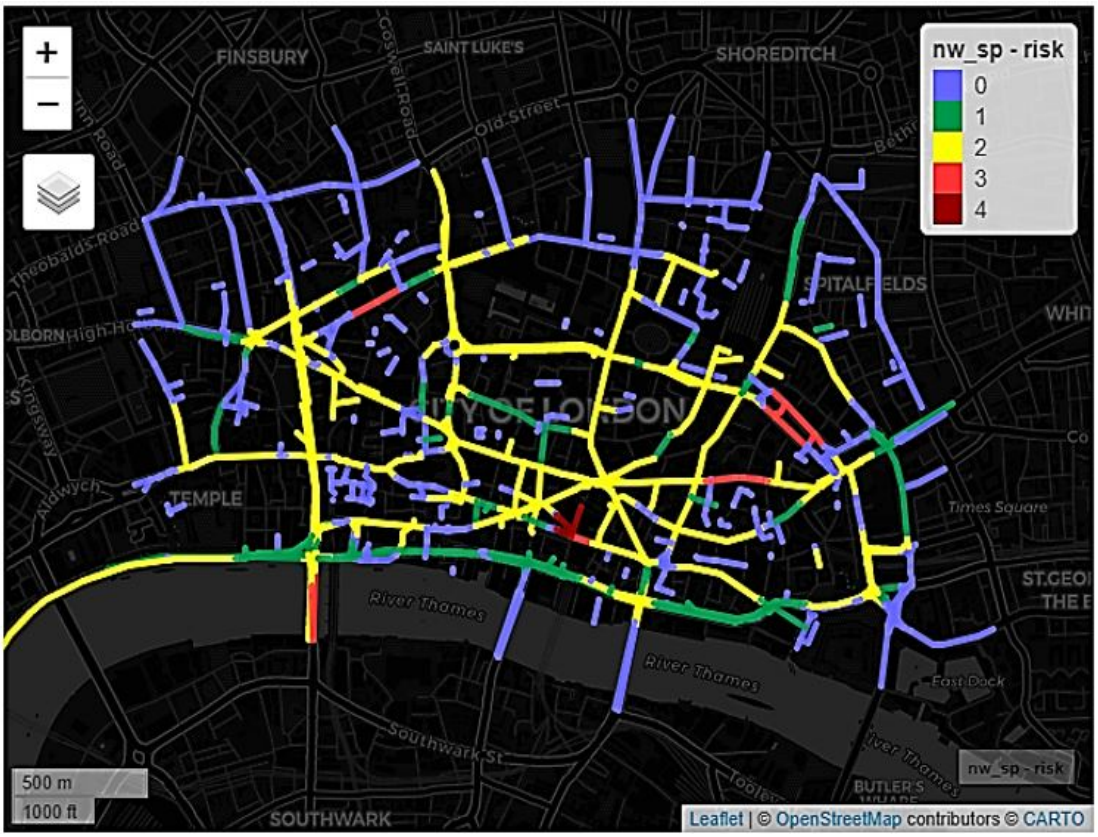


Figure 6.7: Risk map (2016)

Figure 6.8: Original sample data of traffic accident (2016)

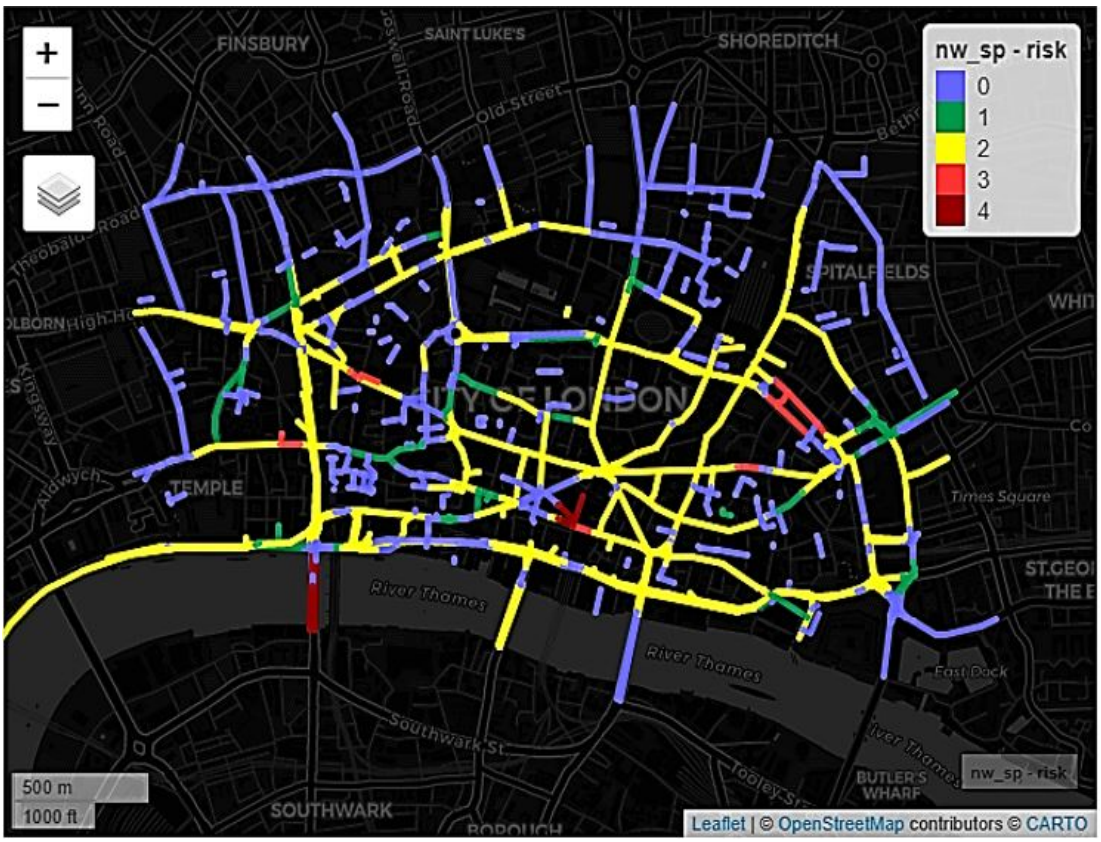


Figure 6.9: Risk map (2017)

Figure 6.10: Original sample data of traffic accident (2017)

6.3 Findings

In this section, selected findings of the current study and brief discussions on each topic with focus on future research scopes are reviewed.

The predicted risk maps are visually compared with the original record of traffic accidents during the same time span. From the above figures (Figure 6.5 to Figure 6.10), few important observations are noted.

- For all the three test years, most of the roads in the outskirt of the city are predicted to be relatively safe than those in the city center. The prediction matches with original record of accidents which can be well explained from Figure 6.6, Figure 6.8 and Figure 6.10.

- Near the city center most of the roads are predicted to be consistent with risk index value 2 during all the test years. Few annual variations are recorded due to changes in spatio-temporal effects in those specific regions. One such example is observed in the historic *Blackfriars Bridge Road, London, UK*. The model predicted that, the road segments in the bridge have gradually turned from safe to high accident-prone region from 2015 to 2017 Figure 6.11.

Figure 6.11: Example: Change point detection (2015-2017)

While comparing with original records it is found that, the number as well as severity of accidents have also increased with the same trend. Future research works to identify the factor(s) controlling trend or change pattern on linear networks can be interesting. Literature suggests similar research works for the city of London (Bhawkar, 2018), (Michalaki et al., 2015), (Curiel et al., 2018), (C. Wang et al., 2009) and identical works in other countries (Ashraf et al., 2019), (Greibe, 2003), (Liu et al., 2017).

- Two critical junctions have been identified (depicted in Figure 6.12) as the highest risk zones in the study area.

Figure 6.12: Detected highest risk zones

These two junctions have consistently highest predicted risk index values during all the test years. Original records also show high rate of severe accident occurrence in both the regions. Literature on randomness and concentration of road accidents in London, UK, "*about 5% of the road junctions are the site of 50% of the accidents*" (Curiel et al., 2018) supports the result. Moreover, the identified high risk junctions also appear as accident-prone zones in the annual reports of collisions and casualties on London roads, maintained by the Department for Transport (DfT) Collision Reporting and SHaring (CRaSH) system (Transport for London, 2019).

The risk-index algorithm implemented in the current study has intended to categorize road segments as a measure of both rate of accident occurrence as well as the severity of the accidents. As a result, segments having high rate of accidents and segments having few but severe accidents, are categorized under similar risk index levels. It is confirmed when individual risk maps are visually compared with the respective original traffic accidents records in Figure 6.6, Figure 6.8 and Figure 6.10. Similar methodology can be adapted in other traffic risk modeling algorithms.

Chapter 7

Discussion and Conclusions

This chapter highlights the final conclusion of the complete research work. It is organised as follows.The first section explores the benefits and application of the model with reference to future research scopes. The next section indicates the limitations of the fitted INLA-SPDE model and recommendations for possible improvements. The conclusions are presented at the final section of this chapter.

In the recent years, spatio-temporal modeling of road traffic accidents and risk mapping has gained attention especially in the domain of multi-dimensional road safety management. Results and findings of the current study illustrate that, the proposed model when fitted with selected covariates can generate predicted risk maps of the entire road network for any urban study area. In that sense, it is dynamic in nature. The current study is conducted in the city of London, UK using selected spatio-temporal variables. As the model is using INLA-SPDE methodology the number of covariates can be updated at any stage. Moreover, it does not require only normally distributed data set which makes the model more flexible to fit in any global study. The model can be easily replicated using different set of covariates. The level of significance for each covariates can be analyzed for further emphasis on selection of significant traffic accident causing factors. The final outcome of the proposed model is a predicted risk map for the entire road network. The maps can be generated for any study area whose prior records can be used to fit the model. At a glance, the road safety index of all the road segments including small details of each junction-points or, sharp turnings can be obtained using these maps. As illustrated in Section 6.3, another crucial application of the model is in analysing change and trend pattern of traffic accidents. Trend in increasing traffic accident risk in any road segment can be detected using risk maps for a continuous time span. Thus, it can have implications on road safety measures through an enhanced understanding of those patterns. Moreover, location of potentially dangerous roads and regions can act as baseline information for geospatial analysis on road safety metrices. The result can have strategic application in developing GIS analytical tools to identify and depict possible safe routes. As the risk map provides information about the entire road network, it can be flexible to generate possible alternative safe route(s) between any source and destinations pairs. Similar research works (Hannah et al., 2018) have been conducted. But in those analysis only spatial traffic variables like crossings, speed limits, and type of street are applied. But the current study has proposed a more flexible and statistically convincing solution

by implementing both spatial and temporal covariates in the predictive model. Besides, travel risk map is gaining popularity among business travellers, tourists and emergency service providers. In that context, the proposed model can have potential application in the spatio-temporal modeling of road traffic accidents in designing geospatial analytical tools on road safety metrices.

Few limitations were raised during the current study. Though the residual diagnostics and predicted risk maps produced by the model matches with the original test data set records; but the correlation and RMSE values of the model imply scopes for improvement. Thus, for detailed understanding of the performance of the model, it may be beneficial to analyze further the model fitting phase using INLA-SPDE, rather than at general, averaged metrics.

The proposed model has been fitted with four temporal and seven non-temporal explanatory variables (mentioned in Table 4.1). In the current study, while selecting the best-fitting model, emphasis has been given on spatial heterogeneity (spatial or, non-spatial models). Similarly, the type of mesh (region or, network mesh) also plays a significant role. But in each case, the complete set of eleven variables have been used. During model fitting process, alternative subsets and combinations of different variables are not tested. As a result, there is chance of existing one or, more covariates which are less significant in the model fitting process. It can have impact in the prediction results. Literature (Cameletti et al., 2012; Martínez-Minaya et al., 2019; Martino & Rue, 2010) suggests, model fitting using diverse subset combinations of the variables provide opportunities to improve the prediction accuracy. Moreover, fixed-effect results of INLA-SPDE models help to identify the significant explanatory variables to be the best fit for the models. Models fitting with only identified significant variables might provide a better performance accuracy.

Initial training data set was selected for a period of consecutive ten years (2005 to 2014). But during data exploration phase it is found that the final two years of the study period (2013 and 2014) comprise 69.91% of the total traffic accident records. First eight years (2005 to 2012) has only 30.09% of the sample. If continued with the same ten years of training period there will be considerable disparity between the annual, monthly, weekly and daily accident counts between these two groups. It might affect the reliability of the model prediction. To ensure this fact, three distinct groups of training years were used and fitted with the proposed INLA-SPDE model with both spatial (network mesh) and temporal effects (RW1). Performance of each group is illustrated in Table 7.1

Table 7.1: Training set results

Year	**mesh$n**	**Ex. time (sec.)**	**DIC**	**WAIC**	**CPO**
2005-2012	8541	475	331.87	366.73	-0.305830
2005-2014	8541	533	330. 64	370.47	-0.308678
2013-2014	8541	312	216.76	280.28	-0.320038

From the result it can be stated that, Group 1 and Group 2 have identical results. It implies, inclusion of first eight years of the training data set do not have

any significant impact in the model result. On the other hand, lower DIC value for Group-3 suggests it to be the best fit. Moreover, as discussed in Figure 5.1 and Figure 5.2 the spatial distribution of the accident records from 2005 to 2012 are clustered only in two regions of the study area. Thus, the unreliable training data from 2005 to 2012 was discarded. But this discard process reduced the size of training data set considerably. Literature (de Fortuny et al., 2013), (Oo & Thein, 2019) suggests, clean and larger data essentially lead to better predictive models. It reduces the probability of spurious correlations and improve the performance of the model(de Fortuny et al., 2013). But the proposed modeling process used only two years (2013 and 2014) of training data set. This might have some impact on the predictive-accuracy of the model. The problem can be easily handled in future research works. Current study shows that, size of data set for 2013 to 2017 is large and reliable. During the ongoing research work, London city accident data for the year 2018 was not available in the open repositories. Once it is available, the model can be fitted using training data set from 2013 to 2017 and test data from the year 2018. Moreover, the model can also be adapted and tested with reliable and large data set in other locations across the UK and globally. Another limitation of the proposed model is that, the sample training data used to train the model is skewed (98.97% of the observations are having response variable value less than 4; out of which 89.86% of response values are 1). As a result of this, during model design and training, *gamma* control family has been used. But while assessing the fitness of the model, the predicted result is expected to have continuous numerical value of any range and thus the control family has been updated to *Gaussian*. This might have some implication on the prediction accuracy of the model. Training data set for a longer time period can solve this issue in future research works.

As conclusions, this thesis presented a dynamic spatio-temporal analysis model predicting the occurrence of traffic accidents in urban environment. The model was used to create the risk map of road networks. To balance speed and accuracy, the current research work took advantage of the spatio-temporal nature of the data and used Bayesian methodology by including INLA and SPDE in the modeling process. The proposed model was designed and tested through a case study in the city of London, UK. The result risk maps provide the geospatial baseline to identify safe routes between source and destination points. The maps can also have implications for accident prevention and road safety measures through an enhanced understanding of the accident patterns. However, the prediction accuracy can be restored by careful inclusion of significant exogenous variables related to traffic flow and traffic control, which can explain some of the uncaptured variations. Furthermore, reliable and large training data set can improve the performance of the proposed model.
The novelty of the proposed model was introducing "network triangulation" or, "network mesh" in SPDE to estimate the spatial autocorrelation of discrete events. As such, it took a new step in INLA-SPDE modeling to perform spatio-temporal predictive analysis only on selected areas (especially for road networks) instead of performing for entire continuous region. In the broader picture, the thesis contributes to the relatively small amount of literature on spatio-temporal analysis using INLA-SPDE of spatial events precisely on road networks. The methodology can be adapted and applied to other locations globally.

Bibliography

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

Aghajani, M. A., Dezfoulian, 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. https://doi.org/10.1016/j.trpro.2017.05.409

Amin, M. S., Reaz, M. B. I., Bhuiyan, M. A. S. & Nasir, S. S. (2014). Kalman filtered gps accelerometer-based accident detection and location system: A low-cost approach. *Current Science*, *106*(11), 1548–1554. http://www.jstor.org/stable/24102459

and. (2003). *Quality and accuracy of positional data in transportation*. Transportation Research Board. https://doi.org/10.17226/21953

Anderson, T. K. (2009). Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, *41*(3), 359–364. https://doi.org/10.1016/j.aap.2008.12.014

Appelhans, T. (2015). Mapview - basic interactive viewing of spatial data in r. https://doi.org/10.13140/RG.2.1.1044.2727

Ashraf, I., Hur, S., Shafiq, M. & Park, Y. (2019). Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis (Y. Guo, Ed.). *PLOS ONE*, *14*(10), e0223473. https://doi.org/10.1371/journal.pone.0223473

Azuike, E. (2018). The causes and prevalence of road traffic accident.

Baddeley, A. & Turner, R. (2005). Spatstat: AnRPackage for analyzing spatial point patterns. *Journal of Statistical Software*, *12*(6). https://doi.org/10.18637/jss.v012.i06

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Krainski, E., Simpson, D. & Lindgren, F. (2018). Spatial modelling with r-inla: A review.

Bhawkar, A. (2018). Severe traffic accidents in united kingdom. https://www.researchgate.net/publication/330676135_Severe_Traffic_Accidents_in_United_Kingdom

Bivand, R. (2017). Geographically weighted regression [r package spgwr version 0.6-32]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/spgwr/index.html

Bivand, R. S., Pebesma, E. & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, second edition*. Springer, NY. http://www.asdar-book.org/

Bivand, R., Keitt, T. & Rowlingson, B. (2019). R interface to gdal, ogr and proj.4: Project home – r-forge. https://r-forge.r-project.org/projects/rgdal/

Bivand, R. & Lewin-Koh, N. (2019). Tools for handling spatial objects [r package maptools version 0.9-9]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/maptools/index.html

Blangiardo, M. & Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models with r-INLA*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118950203

Boulieri, A., Liverani, S., de Hoogh, K. & Blangiardo, M. (2016). A space-time multivariate bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(1), 119–139. https://doi.org/10.1111/rssa.12178

Briz-Redón, Á., Martínez-Ruiz, F. & Montes, F. (2019). Identification of differential risk hotspots for collision and vehicle type in a directed linear network. *Accident Analysis & Prevention*, *132*, 105278.

Bruin, J. (2011). *Newtest: Command to compute new test @ONLINE*. https://stats.idre.ucla.edu/stata/ado/analysis/

Brunsdon, C., Fotheringham, A. S. & Charlton, M. (1998). Spatial nonstationarity and autoregressive models. *Environment and Planning A: Economy and Space*, *30*(6), 957–973. https://doi.org/10.1068/a300957

Buchanan computing collision map. (2019). https://www.collisionmap.uk/

Cameletti, M., Lindgren, F., Simpson, D. & Rue, H. (2012). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, *97*(2), 109–131. https://doi.org/10.1007/s10182-012-0196-3

Cantillo, V., Garcés, P. & Márquez, L. (2016). Factors influencing the occurrence of traffic accidents in urban roads: A combined GIS-empirical bayesian approach. *DYNA*, *83*(195), 21–28. https://doi.org/10.15446/dyna.v83n195.47229

Castro, M., Paleti, R. & Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B: Methodological*, *46*(1), 253–272. https://doi.org/10.1016/j.trb.2011.09.007

Chalabi, Y., Mächler, M. & Würtz, D. (2011). Rmetrics - timeDate Package. *The R Journal*, *3*(1), 19–24. https://doi.org/10.32614/RJ-2011-001

Chambers, J. M., Cleveland, W. S., Tukey, P. A. & Kleiner, B. (1983). *Graphical methods for data analysis (wadsworth & brooks/cole statistics/probability series)*. Duxbury Press. https://www.amazon.com/Graphical-Analysis-Wadsworth-Statistics-Probability

Chaudhuri, S. (2020). *Exploratory analysis of stats19 package data* [Unpublished working package]. Unpublished working package.

Congdon, P. (2014, July 4). *Applied bayesian modelling*. Wiley-Blackwell. https://www.ebook.de/de/product/22311333/peter_congdon_applied_bayesian_modelling.html

Curiel, R. P., Ramırez, H. G. & Bishop, S. R. (2018). A novel rare event approach to measure the randomness and concentration of road accidents (Y. Deng, Ed.). *PLOS ONE*, *13*(8), e0201890. https://doi.org/10.1371/journal.pone.0201890

Curran-Everett, D. (2013). Explorations in statistics: The analysis of ratios and normalized data. *Advances in Physiology Education*, *37*(3), 213–219. https://doi.org/10.1152/advan.00053.2013

de Fortuny, E. J., Martens, D. & Provost, F. (2013). Predictive modeling with big data: Is bigger really better? *Big Data*, *1*(4), 215–226. https://doi.org/10.1089/big.2013.0037

Demasi, F., Loprencipe, G. & Moretti, L. (2018). Road safety analysis of urban roads: Case study of an italian municipality. *Safety*, *4*(4), 58. https://doi.org/10.3390/safety4040058

Deublein, M., Schubert, M., Adey, B. T., Köhler, J. & Faber, M. H. (2013). Prediction of road accidents: A bayesian hierarchical approach. *Accident Analysis & Prevention*, *51*, 274–291. https://doi.org/10.1016/j.aap.2012.11.019

Dunson, D. B. (2001). Commentary: Practical advantages of bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, *153*(12), 1222–1226. https://doi.org/10.1093/aje/153.12.1222

Everitt, B. S. (2006). *An r and s-plus® companion to multivariate analysis (springer texts in statistics)*. Springer.

Farmer, C. M. (2005). Temporal factors in motor vehicle crash deaths. *Injury Prevention*, *11*(1), 18–23. https://doi.org/10.1136/ip.2004.005439

Galgamuwa, U., Du, J. & Dissanayake, S. (2019). Bayesian spatial modeling to incorporate unmeasured information at road segment levels with the INLA approach: A methodological advancement of estimating crash modification factors. *Journal of Traffic and Transportation Engineering (English Edition)*. https://doi.org/10.1016/j.jtte.2019.03.003

Gelfand, A. E., Smith, A. F. M. & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532. https://doi.org/10.1080/01621459.1992.10475235

Gitelman, V., Vis, M., Weijermars, W. & Hakkert, S. (2014). Development of road safety performance indicators for the european countries. *Advances in Social Sciences Research Journal*, *1*(4), 138–158. https://doi.org/10.14738/assrj.14.302

Gollini, I., Lu, B., Charlton, M., Brunsdon, C. & Harris, P. (2015). GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software*, *63*(17), 1–50. http://www.jstatsoft.org/v63/i17/

GoogleMaps. (2019). City of london. Google. http://www.google.com/maps

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention*, *35*(2), 273–285. https://doi.org/10.1016/s0001-4575(02)00005-2

Guo, Y., Osama, A. & Sayed, T. (2018). A cross-comparison of different techniques for modeling macro-level cyclist crashes. *Accident Analysis & Prevention*, *113*, 38–46. https://doi.org/10.1016/j.aap.2018.01.015

Hannah, C., Spasić, I. & Corcoran, P. (2018). A computational model of pedestrian road safety: The long way round is the safe way home. *Accident Analysis & Prevention*, *121*, 347–357. https://doi.org/10.1016/j.aap.2018.06.004

Hezaveh, A. M., Arvin, R. & Cherry, C. R. (2019). A geographically weighted regression to estimate the comprehensive cost of traffic crashes at a zonal level. *Accident Analysis & Prevention*, *131*, 15–24. https://doi.org/10.1016/j.aap.2019.05.028

Hoffman, J. I. (2019). Logistic regression, In *Basic biostatistics for medical and biomedical practitioners*. Elsevier. https://doi.org/10.1016/b978-0-12-817084-7.00033-4

Huang, J., Malone, B. P., Minasny, B., McBratney, A. B. & Triantafilis, J. (2017). Evaluating a bayesian modelling approach (INLA-SPDE) for environmental mapping. *Science of The Total Environment*, *609*, 621–632. https://doi.org/10.1016/j.scitotenv.2017.07.201

Jegede, F. (1988). Spatio-temporal analysis of road traffic accidents in oyo state, nigeria. *Accident Analysis & Prevention*, *20*(3), 227–243. https://doi.org/10.1016/0001-4575(88)90007-3

Jin, R., Li, J. & Shi, J. (2007). Quality prediction and control in rolling processes using logistic regression, In *Transactions of the north american manufacturing research institution of sme*.

Juan, P., Mateu, J. & Saez, M. (2012). Pinpointing spatio-temporal interactions in wildfire patterns. *Stochastic Environmental Research and Risk Assessment*, *26*, 1131–1150.

Karacasu, M., Ergül, B. & Yavuz, A. A. (2013). Estimating the causes of traffic accidents using logistic regression and discriminant analysis. *International*

Journal of Injury Control and Safety Promotion, *21*(4), 305–313. https://doi.org/10.1080/17457300.2013.815632

Karaganis, A. & Mimis, A. (2006). A spatial point process for estimating the probability of occurrence of a traffic accident. *European Regional Science Association, ERSA conference papers*.

Khulbe, D. & Sourav, S. (2019). Modeling severe traffic accidents with spatial and temporal features.

Liu, C. & Sharma, A. (2017). Exploring spatio-temporal effects in traffic crash trend analysis. *Analytic Methods in Accident Research*, *16*, 104–116. https://doi.org/10.1016/j.amar.2017.09.002

Liu, C., Zhang, S., Wu, H. & Fu, Q. (2017). A dynamic spatiotemporal analysis model for traffic incident influence prediction on urban road networks. *ISPRS International Journal of Geo-Information*, *6*(11), 362. https://doi.org/10.3390/ijgi6110362

Loo, B. P. Y., Yao, S. & Wu, J. (2011). Spatial point analysis of road crashes in shanghai: A GIS-based network kernel density method, In *2011 19th international conference on geoinformatics*, IEEE. https://doi.org/10.1109/geoinformatics.2011.5980938

Lord, D. & Persaud, B. N. (2000). Accident prediction models with and without trend: Application of the generalized estimating equations procedure. *Transportation Research Record: Journal of the Transportation Research Board*, *1717*(1), 102–108. https://doi.org/10.3141/1717-13

Lovelace, R., Morgan, M., Hama, L. & Padgham, M. (2019). Stats19: A package for working with open road crash data. *Journal of Open Source Software*, *4*(33), 1181. https://doi.org/10.21105/joss.01181

Manley, E. (2015). Estimating urban traffic patterns through probabilistic interconnectivity of road network junctions (T. Preis, Ed.). *PLOS ONE*, *10*(5), e0127095. https://doi.org/10.1371/journal.pone.0127095

Martínez-Minaya, J., Lindgren, F., López-Quílez, A., Simpson, D. & Conesa, D. (2019). The integrated nested laplace approximation for fitting models with multivariate response.

Martino, S. & Rue, H. (2010). Case studies in bayesian computation using inla. In P. Mantovan & P. Secchi (Eds.), *Complex data modeling and computationally intensive statistical methods* (pp. 99–114). Milano, Springer Milan. https://doi.org/10.1007/978-88-470-1386-5_8

Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, *67*, 68–83. https://doi.org/10.1016/j.csda.2013.04.014

Miaou, S.-P. (1993). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions.

Michalaki, P., Quddus, M. A., Pitfield, D. & Huetson, A. (2015). Exploring the factors affecting motorway accident severity in england using the generalised ordered logistic regression model. *Journal of Safety Research*, *55*, 89–97. https://doi.org/10.1016/j.jsr.2015.09.004

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

Mohanty, M. & Gupta, A. (2015). Factors affecting road crash modeling. *Journal of Transport Literature*, *9*(2), 15–19. https://doi.org/10.1590/2238-1031.jtl.v9n2a3

Moradi, M. M. & Mateu, J. (2019). First- and second-order characteristics of spatio-temporal point processes on linear networks. *Journal of Computational and Graphical Statistics*, 1–21. https://doi.org/10.1080/10618600.2019.1694524

Moradi, M. M. (2018). *Spatial and spatio-temporal point patterns on linear networks* (Doctoral dissertation). Universitat Jaume I. https://doi.org/10.6035/14123.2018.685382

Musenge, E., Chirwa, T. F., Kahn, K. & Vounatsou, P. (2013). Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural north east south africa. *International Journal of Applied Earth Observation and Geoinformation*, *22*, 86–98. https://doi.org/10.1016/j.jag.2012.04.001

Nakaya, T., Fotheringham, A. S., Brunsdon, C. & Charlton, M. (2005). Geographically weighted poisson regression for disease association mapping. *Statistics in Medicine*, *24*(17), 2695–2717. https://doi.org/10.1002/sim.2129

Nakaya, T., Fotheringham, S., Charlton, M. & Brunsdon, C. (2009). Semiparametric geographically weighted generalised linear modelling in gwr 4.0 (Lees, Ed.). In Lees (Ed.). http://mural.maynoothuniversity.ie/4846/

Oh, J., Washington, S. P. & Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention*, *38*(2), 346–356. https://doi.org/10.1016/j.aap.2005.10.004

Oo, M. C. M. & Thein, T. (2019). An efficient predictive analytics system for high dimensional big data. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.09.001

Padgham, M., Rudis, B., Lovelace, R. & Salmon, M. (2017). Osmdata. *The Journal of Open Source Software*, *2*(14). https://doi.org/10.21105/joss.00305

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pikūnas, A., Pumputis, V. & Sadauskas, V. (2004). The influence of vehicles speed on accident rates and their consequences. *Transport*, *19*(1), 15–19. https://doi.org/10.3846/16484142.2004.9637946

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. https://doi.org/10.1061/(asce)te.1943-5436.0000680

Poul Thyregod, H. M. (2010, November 9). *Introduction to general and generalized linear models*. Taylor & Francis Inc. https://www.ebook.de/de/product/8332588/poul_thyregod_henrik_madsen_introduction_to_general_and_generalized_linear_models.html

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. https://doi.org/10.1016/j.sbspro.2011.07.020

Pulugurtha, S. S. & Sambhara, V. R. (2011). Pedestrian crash estimation models for signalized intersections. *Accident Analysis & Prevention*, *43*(1), 439–446. https://doi.org/10.1016/j.aap.2010.09.014

QGIS Development Team. (2009). *Qgis geographic information system*. Open Source Geospatial Foundation. http://qgis.osgeo.org

Qiu, L. & Nixon, W. A. (2008). Effects of adverse weather on traffic crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2055*(1), 139–146. https://doi.org/10.3141/2055-16

Risk mapping. (2016). https://www.eurorap.org/protocols/risk-mapping/

Robin Lovelace & Richard Ellison. (2018). stplanr: A Package for Transport Planning. *The R Journal*, *10*(2). https://doi.org/10.32614/RJ-2018-053

RStudio Team. (2015). *Rstudio: Integrated development environment for r*. RStudio, Inc. Boston, MA. http://www.rstudio.com/

Rue, H. & Held, L. (2005). *Gaussian markov random fields: Theory and applications (chapman & hall/crc monographs on statistics and applied probability)*. Chapman; Hall/CRC.

Rue, H. & Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of Statistical Planning and Inference*, *137*(10), 3177–3192. https://doi.org/10.1016/j.jspi.2006.07.016

Rue, H., Martino, S. & 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. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. & Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, *4*(1), 395–421. https://doi.org/10.1146/annurev-statistics-060116-054045

Salifu, M. (2004). Accident prediction models for unsignalised urban junctions in ghana. *IATSS Research*, *28*(1), 68–81. https://doi.org/10.1016/s0386-1112(14)60093-5

Sampson, M. L., Gounden, V., van Deventer, H. E. & Remaley, A. T. (2016). CUSUM-logistic regression analysis for the rapid detection of errors in clinical laboratory test results. *Clinical Biochemistry*, *49*(3), 201–207. https://doi.org/10.1016/j.clinbiochem.2015.10.019

Sawalha, Z. & Sayed, T. (2003). Statistical issues in traffic accident modeling.

Scott, M. C., Roy, S. S. & Prasad, S. (2016). Spatial patterns of off-the-system traffic crashes in miami–dade county, florida, during 2005–2010. *Traffic Injury Prevention*, *17*(7), 729–735. https://doi.org/10.1080/15389588.2016.1144878

Shafabakhsh, G. A., Famili, A. & Bahadori, M. S. (2017). GIS-based spatial analysis of urban traffic accidents: Case study in mashhad, iran. *Journal of Traffic and Transportation Engineering (English Edition)*, *4*(3), 290–299. https://doi.org/10.1016/j.jtte.2017.05.005

Shahid, S., Minhans, A., Puan, O. C., Hasan, S. A. & Ismail, T. (2015). Spatial and temporal pattern of road accidents and casualties in peninsular malaysia. *Jurnal Teknologi*, *76*(14). https://doi.org/10.11113/jt.v76.5843

Song, J., Ghosh, M., Miaou, S. & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, *97*(1), 246–273. https://doi.org/10.1016/j.jmva.2005.03.007

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & 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. https://doi.org/10.1111/1467-9868.00353

Transport for London. (2019). Road safety. https://tfl.gov.uk/corporate/publications-and-reports/road-safety

Verdoy, P. J. (2019). Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. choosing the best mesh for each database. *Communications in Statistics - Simulation and Computation*, 1–34. https://doi.org/10.1080/03610918.2019.1618473

Wang, C., Quddus, M. A. & Ison, S. G. (2009). Impact of traffic congestion on road accidents: A spatial analysis of the m25 motorway in england. *Accident Analysis & Prevention*, *41*(4), 798–808. https://doi.org/10.1016/j.aap.2009.04.002

Wang, W., Yuan, Z., Yang, Y., Yang, X. & Liu, Y. (2019). Factors influencing traffic accident frequencies on urban roads: A spatial panel time-fixed ef-

fects error model (Y. Guo, Ed.). *PLOS ONE*, *14*(4), e0214539. https://doi.org/10.1371/journal.pone.0214539

Wards. (2018). Mayor; Commonalty; Citizens of the City of London. https://web.archive.org/web/20110612211932/http://www.cityoflondon.gov.uk/Corporation/LGNL_Services/Council_and_democracy/Councillors_democracy_and_elections/ward_boundaries.htm

WHO. (2019, January 10). *Global status report on road safety 2018*. WORLD HEALTH ORGN. https://www.ebook.de/de/product/36226085/world_health_organization_global_status_report_on_road_safety_2018.html

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wikle, C. K., Berliner, L. M. & Cressie, N. (1998). Hierarchical bayesian space-time models. *Environmental and Ecological Statistics*, *5*, 117–154.

Williamson, A. M. & Feyer, A.-M. (1995). Causes of accidents and the time of day. *Work & Stress*, *9*(2-3), 158–164. https://doi.org/10.1080/02678379508256550

Xu, P. & Huang, H. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention*, *75*, 16–25. https://doi.org/10.1016/j.aap.2014.10.020

Zamani, H. & Ismail, N. (2013). Score test for testing zero-inflated poisson regression against zero-inflated generalized poisson alternatives. *Journal of Applied Statistics*, *40*(9), 2056–2068. https://doi.org/10.1080/02664763.2013.804904

Zeng, Q. & Huang, H. (2014). Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis & Prevention*, *67*, 105–112. https://doi.org/10.1016/j.aap.2014.02.018

Zhang, Z. (2016). Model building strategy for logistic regression: Purposeful selection. *Annals of Translational Medicine*, *4*(6), 111–111. https://doi.org/10.21037/atm.2016.02.15

Zheng, L., Robinson, R. M., Khattak, A. J. & Wang, X. (2011). All accidents are not equal: Using geographically weighted regressions models to assess and forecast accident impacts.

Zhong-xiang, F., Shi-sheng, L., Wei-hua, Z. & Nan-nan, Z. (2014). Combined prediction model of death toll for road traffic accidents based on independent and dependent variables. *Computational Intelligence and Neuroscience*, *2014*, 1–7. https://doi.org/10.1155/2014/103196

Zuur, A. F., leno, E. N. & Saveliev, A. A. (2017). *Beginner's guide to spatial, temporal and spatial-temporal ecological data analysis with r-inla: Using glm and glmm volume i* (Vol. 1). Highland Statistics Ltd.

Appendices

Appendix A

Generalized Linear Model Results

A.1 Linear Regression without Spatial Effect

```
## Analysis of Deviance Table (Poisson Regression)
```

	Df	Dev Resid.	Df Resid.	Dev	Pr(>Chi)
NULL			956	106.471	
factor(accident_severity)	2	0.5908	954	105.880	0.7442
factor(day_of_week)	6	0.8851	948	104.995	0.9896
factor(time_slot)	23	4.4440	925	100.551	1.0000
factor(road_type)	4	0.1378	921	100.413	0.9977
factor(speed_limit)	3	0.4001	918	100.013	0.9402
factor(junction_detail)	7	0.3955	911	99.617	0.9997
factor(light_conditions)	3	0.1119	908	99.505	0.9904
factor(weather)	5	1.1480	903	98.357	0.9498
factor(road_surface)	3	0.1289	900	98.229	0.9882
factor(week_end_night)	1	0.1327	899	98.096	0.7156
factor(month)	11	2.8551	888	95.241	0.9925

```
## Analysis of Deviance Table (Logistic Regression)
```

	Df	Dev Resid.	Df Resid.	Dev	Pr(>Chi)	
NULL			956	51.803		
factor(accident_severity)	2	2.9493	954	48.854	0.22885	
factor(day_of_week)	6	6.2597	948	42.594	0.39473	
factor(time_slot)	23	19.1321	925	23.462	0.69353	
factor(road_type)	4	2.4746	921	20.987	0.64918	
factor(speed_limit)	3	0.0000	918	20.987	1.00000	
factor(junction_detail)	7	12.6214	911	8.366	0.08189	.
factor(light_conditions)	3	0.0000	908	8.366	1.00000	
factor(weather)	5	0.0000	903	8.366	1.00000	
factor(road_surface)	3	0.6174	900	7.749	0.89244	
factor(week_end_night)	1	0.0000	899	7.749	0.99995	
factor(month)	11	4.9761	888	2.773	0.93232	

A.2 Linear Regression with Spatial Effect

```
## Analysis of Variance Table

Brunsdon, Fotheringham & Charlton (1999) ANOVA
(BFC99.gwr.test)

data:  gwr.model
F = 1.0372, df1 = 392.64, df2 = 922.91, p-value = 0.3296
sample estimates:
SS GWR improvement    SS GWR residuals
     1.109748           131.247797
```

```
## Summary of GWR coefficient estimates at data points
```

	Min.	1st Qu.	Median	3rd Qu.	Max.
X.Intercept.	0.64857762	1.03071391	1.23738079	1.49621869	1.65267806
accident_severitySerious	−0.09531246	−0.04363329	−0.01965426	0.02166168	0.06869012
accident_severitySlight	−0.14143519	−0.09701780	−0.05983502	−0.00410425	0.05702785
day_of_week2	−0.09085192	−0.06940330	−0.05182906	−0.02916635	−0.00362105
day_of_week3	−0.05875321	−0.04887087	−0.04079060	−0.03476395	−0.02947215
day_of_week4	−0.06177150	−0.04629270	−0.02821297	−0.01222954	0.03458330
day_of_week5	−0.11668209	−0.09360916	−0.07821058	−0.04910188	−0.00947967
day_of_week6	−0.06556488	−0.05287494	−0.04499987	−0.02944667	0.00425858
day_of_week7	−0.15495381	−0.12772829	−0.10365444	−0.06257349	−0.01870418
time_slot	−0.00438555	−0.00317329	−0.00220077	−0.00146281	−0.00040885
road_typeOne.way.street	−0.14831112	−0.08188264	−0.01848508	0.02619426	0.09677703
road_typeOthers	−0.34782526	−0.28487502	−0.20449759	−0.13116129	−0.01797145
road_typeRoundabout	−0.15470224	−0.12562286	−0.08773098	−0.05955850	−0.03318179
road_typeSingle.carriageway	−0.12960136	−0.09979389	−0.04554478	−0.00088860	0.07638665
speed_limit	−0.01654109	−0.01000962	−0.00297348	0.00223244	0.01166842
junction_detail	0.00062380	0.00370641	0.00816587	0.01083718	0.01292832
light_conditions	−0.00360080	0.00066020	0.00138213	0.00224045	0.00402299
weatherFine.with.high.winds	0.11812803	0.16278715	0.19556277	0.24313657	0.47336945
weatherOthers	−0.16408960	−0.10833619	−0.07809014	−0.02505083	0.00258981
weatherRaining.no.high.winds	−0.08875691	0.00097201	0.03179800	0.06435971	0.11871426
weatherRaining.with.high.winds	−0.29164349	−0.20287107	−0.17510184	−0.13957479	−0.11534307
weatherSnowing.no.high.winds	0.07713725	0.14826159	0.18585716	0.20805807	0.39391872
road_surface	0.00090961	0.01035567	0.03554538	0.05425864	0.12948519
Week_end_nightYes	0.12114488	0.15624952	0.16592965	0.17984823	0.20303085
month2	0.06509932	0.08974839	0.09502335	0.10251499	0.11120934
month3	0.10097247	0.10945531	0.11376441	0.12124825	0.13620456
month4	0.08017112	0.09753617	0.09990712	0.10494277	0.13305155
month5	−0.01863718	−0.01405118	0.00606497	0.03282309	0.07465956
month6	0.02745640	0.04437882	0.05641911	0.07399111	0.10697332
month7	−0.04031489	−0.02261832	0.01027657	0.03890939	0.08350560
month8	0.00416234	0.01316789	0.01604245	0.01975757	0.03463327
month9	0.06763515	0.07543859	0.08377721	0.10481541	0.14417479
month10	0.05285774	0.07021458	0.08847650	0.10226748	0.12318191
month11	−0.03379872	−0.02443184	−0.00763820	0.01556685	0.02714648
month12	0.09043723	0.09863573	0.11195909	0.13192740	0.15795174

```
## Monte Carlo test for significance of GWR parameter variability
```

	p−value
(Intercept)	0.06
accident_severitySerious	0.79
accident_severitySlight	0.57
day_of_week2	0.62
day_of_week3	0.88
day_of_week4	0.47
day_of_week5	0.44
day_of_week6	0.68
day_of_week7	0.40
time_slot	0.25
road_typeOne way street	0.12
road_typeOthers	0.07
road_typeRoundabout	0.35
road_typeSingle carriageway	0.07
speed_limit	0.05
junction_detail	0.15
light_conditions	0.67
weatherFine with high winds	0.70
weatherOthers	0.12
weatherRaining no high winds	0.08
weatherRaining with high winds	0.09
weatherSnowing no high winds	0.67
road_surface	0.24
week_end_nightYes	0.77
month2	0.98
month3	0.76
month4	0.61
month5	0.28
month6	0.56
month7	0.18
month8	0.68
month9	0.59
month10	0.61
month11	0.46
month12	0.70

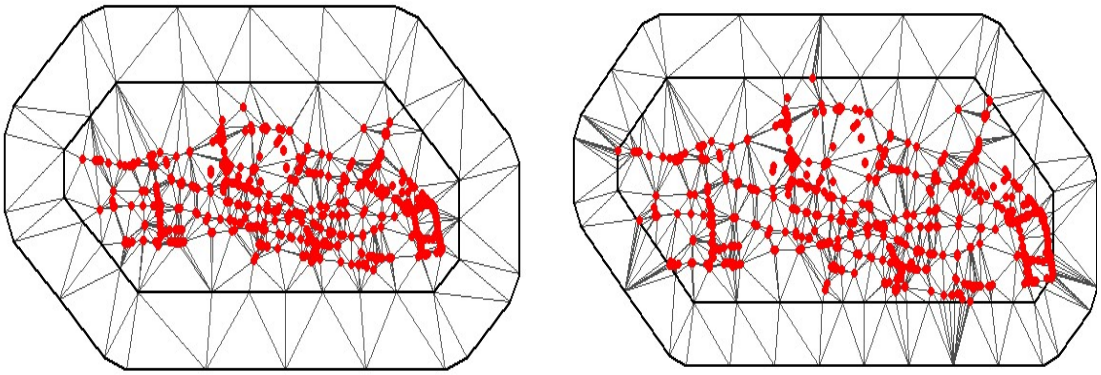
Appendix B

INLA-SPDE Model Results

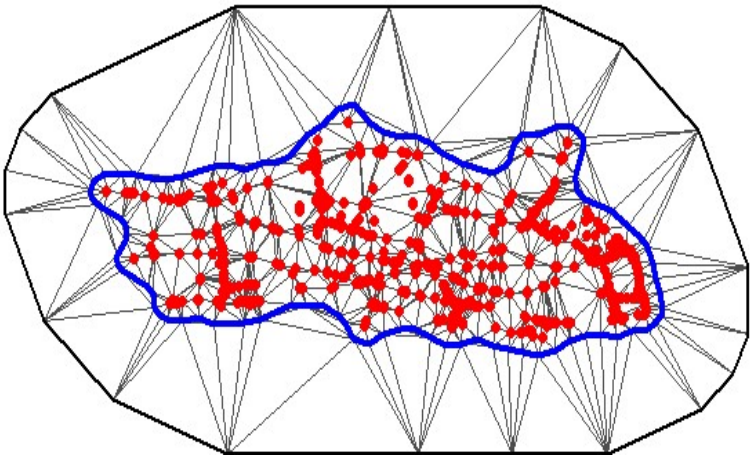
This chapter illustrates the detailed results of all the training data sets. SPDE triangulation results and the model assessment results for each test data set are represented as follows:

B.1 Training Data (2005 - 2014)

SPDE Triangulation Result with 2005-2014 Accident Locations



SPDE region mesh (2005-2014) a) without offset b) with offset



SPDE region mesh (2005-2014) with non-convex hull boundary

SPDE network mesh (2005-2014)

Model Results

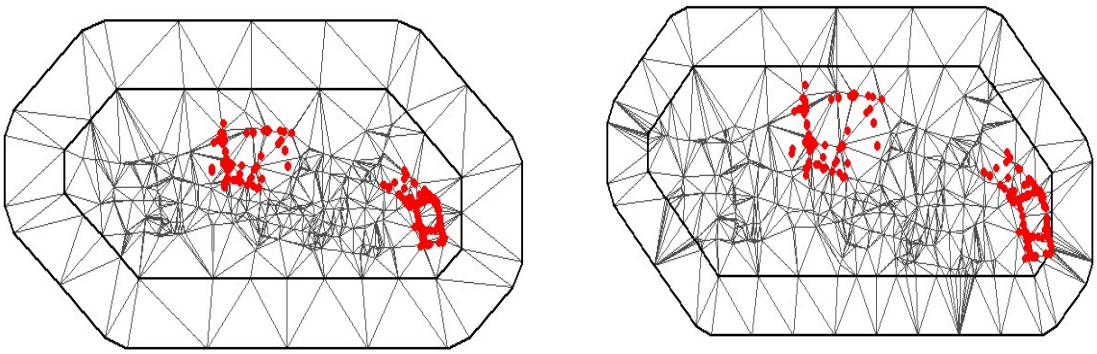
Model Results (2005-2014)

Model	DIC	WAIC	CPO	Mean [0.025quant, 0.975quant]
M1	313.44	367.23	-0.251863	15.78[14.38,17.25]
M2	311.81	369.51	-0.254919	15.99[14.54,17.54]
M3	360.75	296.42	-1.507253	22.50[22.31,22.84]
M4	320.82	367.9	-0.396691	18.98[18.80,35.62]
M5	252.61	319.85	0.553167	26.62[26.01,27.56]
M6	330.87	366.73	-0.305829	16.89[16.78,30.44]

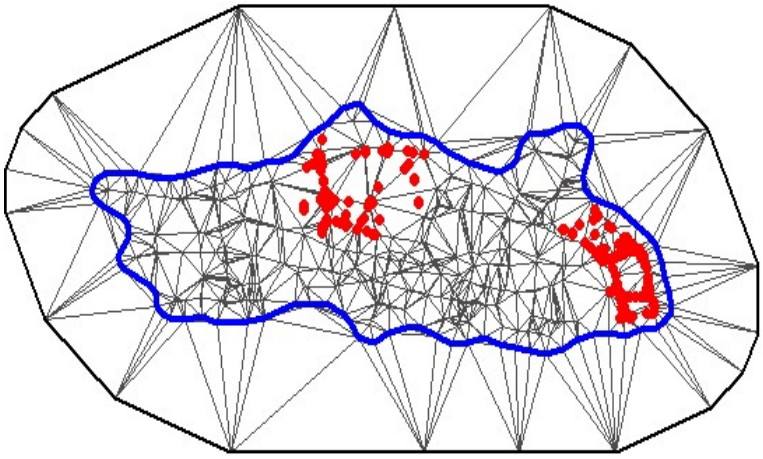
B.2 Training Data (2005 - 2012)

Model fitting using training data set from 2005 to 2012. From the triangulation plot it is clear that, the number of accident plots for this time period is relatively low compared to the previous training data set.

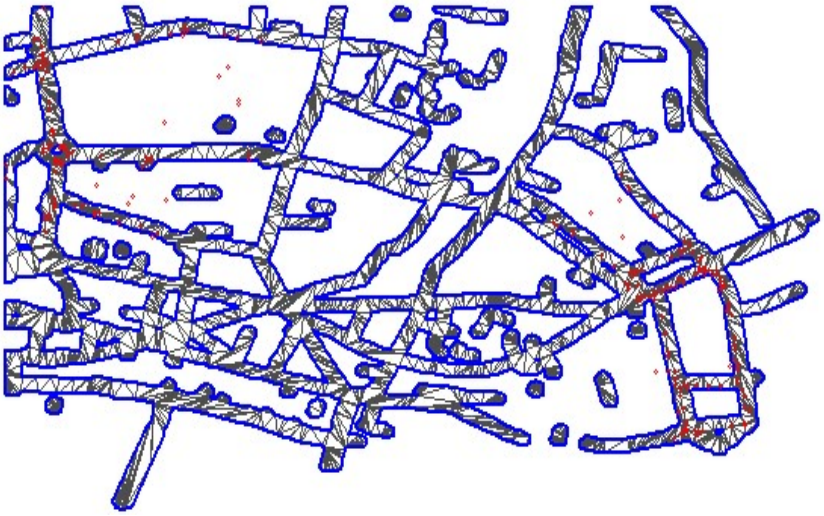
SPDE Triangulation Result with 2005-2012 Accident Locations



SPDE region mesh (2005-2012) a) without offset b) with offset



SPDE region mesh (2005-2012) with non-convex hull boundary



SPDE network mesh (2005-2012)

Model Results

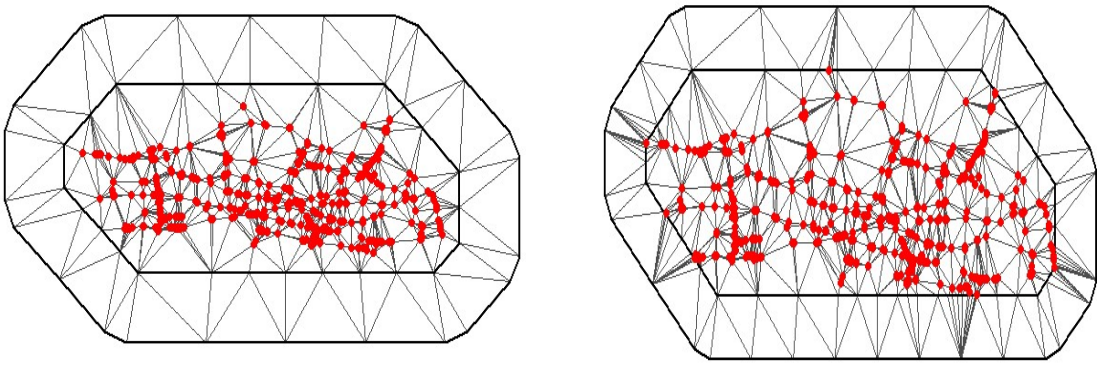
Model Results (2005-2012)

Model	DIC	WAIC	CPO	Mean [0.025quant, 0.975quant]
M1	313.44	367.23	-0.251878	15.78[14.38,17.25]
M2	311.81	369.51	-0.254918	15.99[14.54,17.54]
M3	248.08	348.35	-0.498941	28.42[28.26,28.57]
M4	320.82	367.9	-0.396691	18.98[18.80,35.62]
M5	305.44	313.43	-0.537538	21.29[21.13,45.82]
M6	330.87	366.73	-0.305829	16.89[16.78,30.44]

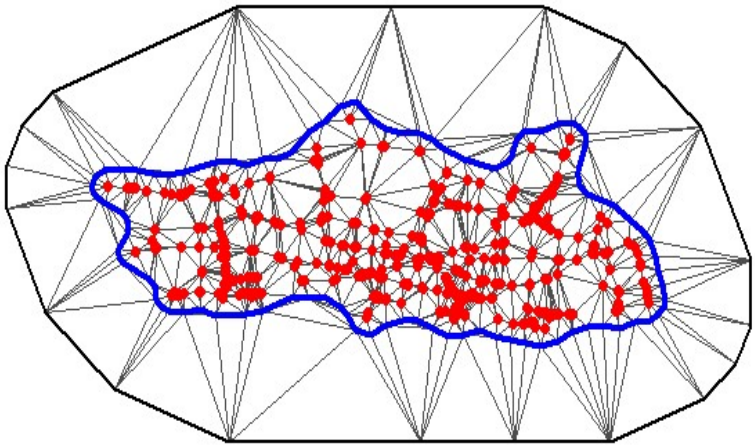
B.3 Training Data (2013 - 2014)

Model fitting using training data set from 2013 to 2014. From the previous two triangulation plot it is clear that, the number of accident plots for the time period 2005 to 2012 period is relatively low and clustered in to regions only. Thus, the final training data set selected is traffic the accident record for the years 2013 and 2014.

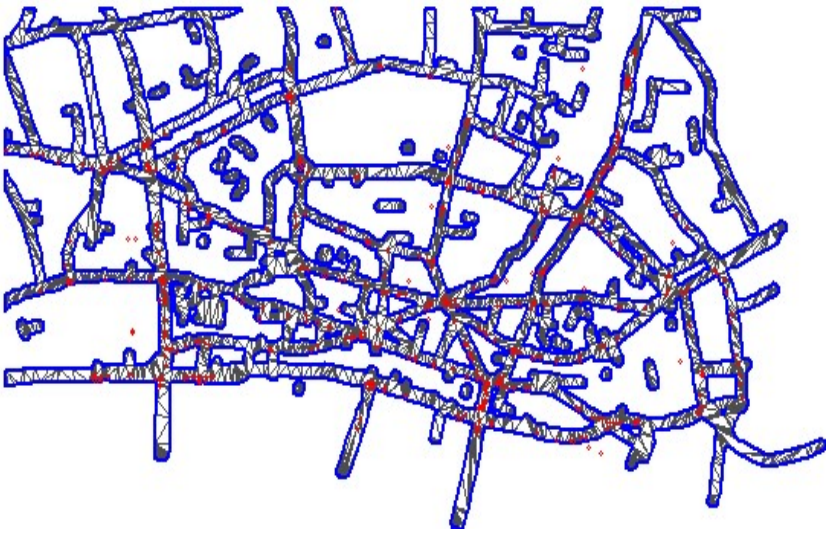
SPDE Triangulation Result with 2013-2014 Accident Locations



SPDE region mesh (2013-2014) a) without offset b) with offset



SPDE region mesh (2013-2014) with non-convex hull boundary



SPDE network mesh (2013-2014)

Model Results (2013-2014)

Model	**DIC**	**WAIC**	**CPO**	**Mean [0.025quant, 0.975quant]**
M1	239.60	285.84	-0.406851	15.51[13.87,17.25]
M2	211.16	259.54	-0.356291	16.55[14.75,18.47]
M3	273.63	203.37	-2.395325	22.61[22.36,23.09]
M4	267.68	261.22	-0.868227	20.30[19.97,20.77]
M5	202.85	234.34	-0.801388	24.68[24.36,24.98]
M6	216.76	280.28	-0.320038	20.90[20.74,22.78]

To analyze the significance of temporal and spatial effects, the model 6 (M6) was fitted with temporal and spatial covariates separately. The result is as follows:

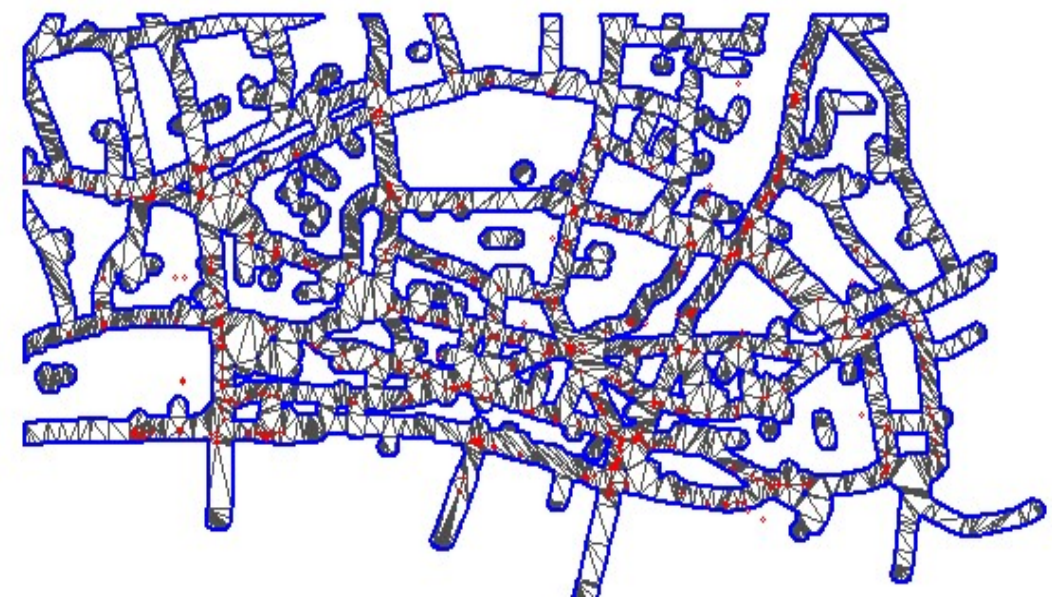
Result of Model 6 Using Temporal and Spatial Effects Separately on training data 2013-2014

Model	**DIC**	**WAIC**	**CPO**	**Mean [0.025quant, 0.975quant]**
Temporal	216.12	269.15	-0.565096	22.86[22.57,47.20]
Spatial	241.88	260.42	-0.490794	18.76[18.46,39.73]

B.4 SPDE Network Triangulation

SPDE network triangulation (network mesh) has been explored with different buffer size applied on the OSM road segments. The best fitting mesh should have enough vertices for effective predictive analysis, but the number should be within a limit to control the processing time. Examples of SPDE network mesh with buffer size 10m and 30m are illustrated with training data set of 2013-2014. The optimal buffer size applied in the current study is of 20m.

SPDE network mesh (2013-2014) with 10m buffer



SPDE network mesh (2013-2014) with 30m buffer

Appendix C

Code

```
## Poisson Regression
acc_posn <- glm(number_of_casualties ~
            factor(accident_severity)+
            factor(day_of_week)+
            factor(time_slot)+
            factor(road_type)+
            factor(speed_limit)+
            factor(junction_detail)+
            factor(light_conditions)+
            factor(weather)+
            factor(road_surface)+
            factor(week_end_night)+
            factor(month),
            data = acc.test, family = poisson)


## Logistic Regression
acc_log <- glm(logi_reg_var ~
           factor(accident_severity)+
           factor(day_of_week)+
           factor(time_slot)+
           factor(road_type)+
           factor(speed_limit)+
           factor(junction_detail)+
           factor(light_conditions)+
           factor(weather)+
           factor(road_surface)+
           factor(week_end_night)+
           factor(month),
           data = acc.test, family = binomial)
```

```
## Geographically Weighted Regression (GWR)

## GWR (Fixed Kernel Method)

    # Calculate fixed bandwidth
    DM <- gw.dist(dp.locat= coordinates(acc.points))
    bw_fixed <- bw.gwr(number_of_casualties~
            accident_severity+
            day_of_week+
            time_slot+
            road_type+
            speed_limit+
            junction_detail+
            light_conditions+
            weather+
            road_surface+
            week_end_night+
            month,
            data=acc.points,
            adaptive=FALSE, dMat=DM,
            approach="CV", p=2, kernel="gaussian")

    # Implement basic GWR
    gwr.res <- gwr.basic(number_of_casualties~
            accident_severity+
            day_of_week+
            time_slot+
            road_type+
            speed_limit+
            junction_detail+
            light_conditions+
            weather+
            road_surface+
            week_end_night+
            month,
            data=acc.points,
            adaptive=FALSE, dMat=DM,
            bw=bw_fixed, p=2, kernel="gaussian")


## GWR (Adaptive Kernel Method)

    # Calculate adaptive bandwidth
    DM <- gw.dist(dp.locat= coordinates(acc.points))
    bw_adapt <- bw.gwr(number_of_casualties~
            accident_severity+
            day_of_week+
            time_slot+
            road_type+
            speed_limit+
            junction_detail+
```

```
        light_conditions+
        weather+
        road_surface+
        week_end_night+
        month,
        data=acc.points,
        adaptive=TRUE, dMat=DM,
        approach="AICc", p=2, kernel="gaussian")

# Implement basic GWR
gwr.res <- gwr.basic(number_of_casualties~
        accident_severity+
        day_of_week+
        time_slot+
        road_type+
        speed_limit+
        junction_detail+
        light_conditions+
        weather+
        road_surface+
        week_end_night+
        month,
        data=acc.points,
        adaptive=TRUE, dMat=DM,
        bw=bw_adapt, p=2, kernel="gaussian")


# Monte Carlo Method
mm.gwr <- gwr.montecarlo(number_of_casualties ~
        accident_severity+
        day_of_week+
        time_slot+
        road_type+
        speed_limit+
        junction_detail+
        light_conditions+
        weather+
        road_surface+
        week_end_night+
        month,
        data=acc.points,
        adaptive=TRUE,
        bw=bw_adapt, p=2, kernel="gaussian")
```

```
## INLA Model
# Building SPDE mesh
mesh <- inla.mesh.2d(loc = coords, boundary=bg_poly2, max.edge=.1)
mesh$n
plot(mesh)
points(coords, col = "red", cex=0.4)

# SPDE Model
spde <- inla.spde2.matern(mesh, alpha = 2)
str(spde2)

# SPDE Model Index
s.index <- inla.spde.make.index(name="i", n.spde = mesh$n)
str(s.index)

# Projector matrix
A <- inla.spde.make.A(mesh, loc = coords)
str(A)
dim(A)

# Building INLA Stack
stack <- inla.stack(tag="est",
         data=list(y=number_of_casualties),
         effects=list(c(s.index,list(m=1)),
         list(accident_severity = accident_severity,
         day_of_week = day_of_week,
         time_slot = time_slot,
         road_type = road_type,
         speed_limit = speed_limit,
         junction_detail = junction_detail,
         light_conditions = light_conditions,
         road_surface = road_surface,
         weather = weather,
         week_end_night = Week_end_night,
         month = month,
         rw_date = rw_date)),
         A = list(A, 1)
        )
dim(inla.stack.A(stack))

# Model  with all covariates and temporal and spatial effects
  fn_inla <- y ~ -1 +
         factor(accident_severity)+
         factor(day_of_week)+
         factor(time_slot)+
         factor(road_type)+
         factor(speed_limit)+
         factor(junction_detail)+
         factor(light_conditions)+
         factor(weather)+
```

```
          factor(road_surface)+
          factor(Week_end_night)+
          factor(month) +
          f(rw_date, model = "rw1")+
          f(i, model = spde)

# INLA Model Fitting
res_inla <- inla( fn_inla,
          data=inla.stack.data(stack),
          control.predictor= list(
          A = inla.stack.A(stack), compute=T),
          family = "gamma", control.compute = list(
          config=T, dic=T, cpo=T, waic=T),
          keep=FALSE, verbose=TRUE)
```

Masters Program in Geospatial Technologies

Spatio-temporal Modeling of Traffic Risk Mapping on Urban Road Networks

Somnath Chaudhuri

Dissertation submitted in partial fulfilment of the requirements for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS