MONIQUE MARTINS GOMES

MODELOS ESPACIAIS DE PREVISÃO DE ACIDENTES: UMA AVALIAÇÃO DO DESEMPENHO DOS MODELOS A PARTIR DA INCORPORAÇÃO DE INFORMAÇÕES APRIMORADAS E A ADEQUAÇÃO DE DIFERENTES ABORDAGENS DE MODELAGEM ESPACIAL

Programa de Pós-graduação em Engenharia de Transportes da EESC-USP

Exemplar definitivo (corrigido). O exemplar original está disponível na CPG da EESC-USP

São Carlos, **31/01/2019**

RESOLUÇÃO CoPGr N º 6018, DE 13 DE OUTUBRO DE 2011, artigo 5º

Tese de Doutorado apresentada ao Departamento de Engenharia de Transportes da Escola de Engenharia de São Carlos, da Universidade de São Paulo e ao *Transportation Research Institute* da Hasselt University como parte dos requisitos para a obtenção do título de Doutor em Ciências (Programa de Pós-graduação em Engenharia de Transportes, área de concentração: Planejamento e Operação de Transportes) e *Doctor of Transportation Sciences*, respectivamente.

Orientadores: Profª. Cira Souza Pitombo

Prof. Tom Brijs

Coorientador: Prof. Ali Pirdavani

São Carlos
Dezembro
2018

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues
Fontes da EESC/USP

G633m

Gomes, Monique Martins
Modelos espaciais de previsão de acidentes : uma avaliação do desempenho dos modelos a partir da incorporação de informações aprimoradas e a adequação de diferentes abordagens de modelagem espacial / Monique Martins Gomes; orientadores Cira Souza Pitombo; Tom Brijs; coorientador Ali Pirdavani. -- São Carlos; Hasselt, 2018.

Tese (Doutorado) - Programa de Pós-Graduação em Engenharia de Transportes e Área de Concentração em Planejamento e Operação de Sistemas de Transporte da Escola de Engenharia de São Carlos da Universidade de São Paulo e Transportation Institute Hasselt University, 2018.

1. Modelos de previsão de acidentes. 2. Regressão geograficamente ponderada. 3. Segurança no trânsito. 4. Geoestatística. 5. Modelos espaciais de predição. 6. *Holdout* repetido. I. Titulo.

Elaborado por Elena Luzia Palloni – CRB 4464

FOLHA DE JULGAMENTO

Candidato: Tecnólogo **MONIQUE MARTINS GOMES.**

Título da tese: "Modelos espaciais de previsão de acidentes: uma avaliação do desempenho dos modelos a partir da incorporação de informações aprimoradas e a adequação de diferentes abordagens de modelagem espacial".

Data da defesa: 04/12/2018.

Comissão Julgadora:	**Resultado:**
Profa. Associada **Cira Souza Pitombo** **(Orientadora)** (Escola de Engenharia de São Carlos/EESC)	Aprovado
Prof. Dr. **Gustavo Garcia Manzato** (Universidade Estadual Paulista "Júlio de Mesquita Filho"/UNESP - Bauru)	Aprovado
Profa. Dra. **Barbara Stolte Bezerra** (Universidade Estadual Paulista "Júlio de Mesquita Filho"/UNESP - Bauru)	APROVADO
Prof. Dr. **Ali Pirdavani** (Hasselt University)	Aprovado
Prof. Dr. **Tom Brijs** (Hasselt University)	Aprovado

Coordenadora do Programa de Pós-Graduação em Engenharia de Transportes:
Profa. Associada **Ana Paula Camargo Larocca**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Luís Fernando Costa Alberto**

MONIQUE MARTINS GOMES

SPATIAL CRASH PREDICTION MODELS: AN EVALUATION OF THE IMPACTS OF ENRICHED INFORMATION ON MODEL PERFORMANCE AND THE SUITABILITY OF DIFFERENT SPATIAL MODELING APPROACHES

Doctorate thesis presented to the Department of Transportation Engineering of the School of Engineering of São Carlos at University of São Paulo, and to the Transportation Research Institute of Hasselt University as part of the requirements for obtaining the degrees of Doctor in Sciences and Doctor in Transportations Sciences, respectively.

Promoters: Prof. Cira Souza Pitombo

Prof. Dr. Tom Brijs

Co-Promoter: Prof. Ali Pirdavani

São Carlos
December
2018

Catalog card prepared by Patron Service at "Prof. Dr. Sergio Rodrigues Fontes" Library at EESC/USP

G633s

Gomes, Monique Martins
Spatial crash prediction models : an evaluation of the impacts of enriched information on model performance and the suitability of different spatial modeling approaches / Monique Martins Gomes; promoters Cira Souza Pitombo; Tom Brijs; co-promoter Ali Pirdavani. -- São Carlos; Hasselt, 2018.

Doctorate (Thesis) - Graduate Program in Transportation Engineering and Research Area in Planning and Operation of Transport Systems of the School of Engineering of São Carlos at University of São Paulo and to the Transportation Research Institute of Hasselt University, 2018.

1. Crash prediction models. 2. Geographically weighted regression. 3. Road safety. 4. Geostatistics. 5. Spatial prediction models. 6. Repeated Holdout. I. Title.

Elena Luzia Palloni Gonçalves – CRB 4464

2018 | School for Transportation Sciences

UHASSELT

KNOWLEDGE IN ACTION

EESC · USP

Doctoral dissertation submitted to obtain the degrees of
- Doctor in Sciences for Universidade de São Paulo
- Doctor of Transportation Sciences for Hasselt University

Monique Martins Gomes

DOCTORAL DISSERTATION

Spatial crash prediction models: an evaluation of the impacts of enriched information on model performance and the suitability of different spatial modeling approaches

Promoters: Prof. Dr Tom Brijs | UHasselt
Prof. Dr Cira Souza Pitombo | University of São Paulo

Co-promoter: Prof. Dr Msc Ali Pirdavani | UHasselt

D/2018/2451/93

DEDICATION

To the memory of my mom.
God has you in his keeping; I have you in my heart.

To my father and all other victims of road crash,
I believe in a safer tomorrow.

ACKNOWLEDGMENTS

I would like to acknowledge many people that contributed to the preparation of this joint doctorate thesis along these years.

First, my special thanks to my promoters, Dr. Cira Souza Pitombo, Dr. Tom Brijs and Dr. Ali Pirdavani, for the guidance and trustworthy. Without their support and constructive comments, this research would never have been possible.

Second, my extensive gratitude to Dr. Bárbara Bezerra, Dr. José Alberto Quintanilha, Dr. Gustavo Garcia Manzato, Dr. Simone Becker Lopes, Dr. Rita Salgueiro, Dr. André Cunha and Dr. Cassiano Isler, for the wise advices in the early stages of my PhD, which were essential to build solid basis and determine the direction of the research.

My sincere appreciation further goes to Dr. Paulo Cesar Lima Segantine and Dr. Ana Paula Camargo Larocca, for making this joint doctorate possible. I also would like to thank both USP and UHasselt staff for their collaboration in this process, especially Cesar Derisso, Heloísa Belo, Elizabeth Ortega, Sabrina de Brito, Kristel Hertogs and Edith Donders.

I thank my colleagues at USP and UHasselt, for their company and friendship, especially those who were more closely involved in making this journey more pleasant along these years: Murilo Santos, Andressa NG, Artur Piatti, Fernando Piva, Thalita Nascimento, Marcela Navarro, Joicy Poloni, Isabela Fornaciari, Viviani Antunes Gomes, Marilia Gabriela Morais, Aurenice Figueira Yasaí, Loana Sanchez, Maria José Zagatto, Bruno Medeiros and many others.

My gratitude also to some colleagues who were more closely involved in helping me with this research: Renaude Carneiro, Rafael Cavalcanti, Fabio Vieira, Tiago Torquato, Lucas Assirati and Samille Rocha. I also thank Jorge Cossío for helping me with the practical arrangements.

My acknowledgment to CNPq – *Conselho Nacional de Desenvolvimento Científico e Tecnológico* and the Science without Borders Program for the financial support that enable this fruitful experience.

Last but not least, my profound gratitude to Yannick, for being understanding, supporting me emotionally, and always motivating me along these years; brothers and “families” in Bauru, Botucatu and Belgium, whose always support me in my decisions and for being the source of infinite love.

ABSTRACT

GOMES, M .M. ***Spatial crash prediction models: an evaluation of the impacts of enriched information on model performance and the suitability of different spatial modeling approaches.*** São Carlos, 2018, 170 p. Doctorate thesis – Double degree between School of Engineering of São Carlos at University of São Paulo and Instituut voor Mobiliteit (Universiteit Hasselt)

The unavailability of crash-related data has been a long lasting challenge in Brazil. In addition to the poor implementation and follow-up of road safety strategies, this drawback has hampered the development of studies that could contribute to national goals toward road safety. In contrast, developed countries have built their effective strategies on solid data basis, therefore, investing a considerable time and money in obtaining and creating pertinent information. In this research, we aim to assess the potential impacts of supplementary data on spatial model performance and the suitability of different spatial modeling approaches on crash prediction. The intention is to notify the authorities in Brazil and other developing countries, about the importance of having appropriate data. In this thesis, we set two specific objectives: (I) to investigate the spatial model prediction accuracy at unsampled subzones; (II) to evaluate the performance of spatial data analysis approaches on crash prediction. Firstly, we carry out a benchmarking based on Geographically Weighted Regression (GWR) models developed for Flanders, Belgium, and São Paulo, Brazil. Models are developed for two modes of transport: active (i.e. pedestrians and cyclists) and motorized transport (i.e. motorized vehicles occupants). Subsequently, we apply the repeated holdout method on the Flemish models, introducing two GWR validation approaches, named GWR holdout1 and GWR holdout2. While the former is based on the local coefficient estimates derived from the neighboring subzones and measures of the explanatory variables for the validation subzones, the latter uses the casualty estimates of the neighboring subzones directly to estimate outcomes for the missing subzones. Lastly, we compare the performance of GWR models with Mean Imputation (MEI), K-Nearest Neighbor (KNN) and Kriging with External Drift (KED). Findings showed that by adding the supplementary data, reductions of 20% and 25% for motorized transport, and 25% and 35% for active transport resulted in corrected Akaike Information Criterion (AICc) and Mean Squared Prediction Errors (MSPE), respectively. From a practical perspective, the results could help us identify hotspots and prioritize data collection strategies besides identify, implement and enforce appropriate countermeasures. Concerning the spatial approaches,

GWR holdout2 outperformed all other techniques and proved that GWR is an appropriate spatial technique for both prediction and impact analyses. Especially in countries where data availability has been an issue, this validation framework allows casualties or crash frequencies to be estimated while effectively capturing the spatial variation of the data.

Keywords: Crash Prediction Models, Geographically Weighted Regression, Road Safety, Geostatistics, Spatial Prediction Models, Repeated Holdout.

RESUMO

GOMES, M. M. ***Modelos espaciais de previsão de acidentes: uma avaliação do desempenho dos modelos a partir da incorporação de informações aprimoradas e a adequação de diferentes abordagens de modelagem espacial***. São Carlos, 2018, 170 p. Duplo diploma entre Escola de Engenharia de São Carlos (Universidade de São Paulo) e Instituut voor Mobiliteit (Universiteit Hasselt)

A indisponibilidade de variáveis explicativas de acidentes de trânsito tem sido um desafio duradouro no Brasil. Além da má implementação e acompanhamento de estratégias de segurança viária, esse inconveniente tem dificultado o desenvolvimento de estudos que poderiam contribuir com as metas nacionais de segurança no trânsito. Em contraste, países desenvolvidos tem construído suas estratégias efetivas com base em dados sólidos, e portanto, investindo tempo e dinheiro consideráveis na obtenção e criação de informações pertinentes. O objetivo dessa pesquisa é avaliar os possíveis impactos de dados suplementares sobre o desempenho de modelos espaciais, e a adequação de diferentes abordagens de modelagem espacial na previsão de acidentes. A intenção é notificar as autoridades brasileiras e de outros países em desenvolvimento sobre a importância de dados adequados. Nesta tese, foram definidos dois objetivos específicos: (I) investigar a acurácia do modelo espacial em subzonas sem amostragem; (II) avaliar o desempenho de técnicas de análise espacial de dados na previsão de acidentes. Primeiramente, foi realizado um estudo comparativo, baseado em modelos desenvolvidos para Flandres (Bélgica) e São Paulo (Brasil), através do método de Regressão Geograficamente Ponderada (RGP). Os modelos foram desenvolvidos para dois modos de transporte: ativos (pedestres e ciclistas) e motorizados (ocupantes de veículos motorizados). Subsequentemente, foi aplicado o método de *holdout* repetido nos modelos Flamengos, introduzindo duas abordagens de validação para GWR, denominados RGP holdout1 e RGP holdout2. Enquanto o primeiro é baseado nas estimativas de coeficientes locais derivados das subzonas vizinhas e medidas das variáveis explicativas para as subzonas de validação, o último usa as estimativas de acidentes das subzonas vizinhas, diretamente, para estimar os resultados para as subzonas ausentes. Por fim, foi comparado o desempenho de modelos RGP e outras abordagens, tais como Imputação pela Média de dados faltantes (IM), K-vizinhos mais próximos (KNN) e Krigagem com Deriva Externa (KDE). Os resultados mostraram que, adicionando os dados suplementares, reduções de 20% e 25% para o transporte motorizado, e 25% e 35% para o transporte ativo, foram resultantes em termos de Critério de

Informação de Akaike corrigido (AICc) e Erro Quadrático Médio da Predição (EQMP), respectivamente. Do ponto de vista prático, os resultados poderiam ajudar a identificar *hotspots* e priorizar estratégias de coleta de dados, além de identificar, implementar e aplicar contramedidas adequadas. No que diz respeito às abordagens espaciais, RGP holdout2 teve melhor desempenho em relação a todas as outras técnicas e, provou que a RGP é uma técnica espacial apropriada para ambas as análises de previsão e impactos. Especialmente em países onde a disponibilidade de dados tem sido um problema, essa estrutura de validação permite que as acidentes sejam estimados enquanto, capturando efetivamente a variação espacial dos dados.

Palavras-chave: Modelos de Previsão de Acidentes, Regressão Geograficamente Ponderada, Segurança no Trânsito, Geoestatística, Modelos Espaciais de Predição, *Holdout* repetido.

SAMENVATTING

GOMES, M .M. ***Ruimtelijke ongevalspredictiemodellen: een evaluatie van de impact van verrijkte informatie op modelperformantie en de geschiktheid van verschillende ruimtelijke modelleertechnieken***. São Carlos, 2018, 170 p. Double degree tussen Engenharia de São Carlos (Universidade de São Paulo) en Instituut voor Mobiliteit (Universiteit Hasselt)

De onbeschikbaarheid van ongevalsgerelateerde data is al geruime tijd een uitdaging in Brazilië. In combinatie met het gebrekkig implementeren en opvolgen van verkeersveiligheidstrategieën heeft dit nadeel ervoor gezorgd dat het ontwikkelen van studies die kunnen bijdragen aan nationale doelstellingen met betrekking tot verkeersveligheid werden belemmerd. Anderzijds hebben ontwikkelde landen tijd en financiële middelen vrijgemaakt om data te verzamelen en te valideren zodat effectieve strategieën op basis van deze data konden worden ontwikkeld. In dit onderzoek bestuderen we de potentiële impact van supplementaire data op de performantie van ruimtelijke modellen en de geschiktheid van verschillende ruimtelijke modellen om voorspellingen van ongevallen te doen. Het is de bedoeling om de autoriteiten in Brazilië en andere ontwikkelingslanden het belang van kwaliteitsvolle data te doen inzien. In deze thesis zetten we twee specifieke doelstellingen voorop: (I) onderzoeken hoe accuraat de voorspellingen van ruimtelijke modellen zijn op subzones die geen deel uitmaken van de steekproef (II) de performantie van ruimtelijke data analyse methodes op het voorspellen van ongevallen evalueren. Ten eerste voeren we een vergelijking uit gebaseerd op Geographically Weighted Regression (GWR) modellen die ontwikkeld zijn voor Vlaanderen, België en São Paulo, Brazilië. De modellen zijn ontwikkeld voor twee transportmodi: actief (voetgangers en fietsers) en gemotoriseerd transport (inzittenden van gemotoriseerde voertuigen). Vervolgens passen we de repeated holdout methode toe op de Vlaamse modellen; zo introduceren we twee GWR validatietechnieken, namelijk GWR holdout1 en GWR holdout2. Terwijl de eerste methode gebaseerd is op de schattingen van lokale coëfficiënten die zijn afgeleid van naburige subzones en de waardes van verklarende variabelen voor de validatie subzones, gebruikt de tweede methode slachtofferschattingen van naburige subzones om uitkomsten voor ontbrekende subzones rechtstreeks te schatten. Tenslotte vergelijken we de performantie van GWR modellen met Mean Imputation (MEI), K-Nearest Neighbor (KNN) en Kriging with External Drift (KED). Resultaten tonen aan dat door toevoeging van additionele data, reducties van 20 tot 25% voor gemotoriseerd transport en 25 tot 35% voor actief transport bekomen worden voor de gecorrigeerde Akaike Information Criterion (AICc) en Mean Squared Prediction Errors (MSPE). Vanuit een praktisch standpunt kunnen de

resultaten helpen om gevaarlijke locaties te identificeren en om dataverzamelingsstrategiën te prioritiseren, alsook om tegenmaatregelen te identificeren, implementeren en op te leggen. Met betrekking tot de ruimtelijke technieken presteerde de GWR holdout2 beter dan alle andere technieken en bewijst dit dat GWR een gepaste ruimtelijke techniek is voor zowel voorspellings- als impactanalyse. Voornamelijk in landen waar beschikbaarheid van data een struikelblok is, levert dit validatieframework een methode om slachtoffers of ongevalsfrequenties te schatten en tegelijkertijd de ruimtelijke variatie in de data effectief te vatten.

Sleutelwoorden: Ongevalsvoorspellingsmodellen, Geographically Weighted Regression, Verkeersveiligheid, Geostatistiek, Ruimtelijke voorspellingsmodellen, Repeated Holdout.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AADT	Average Annual Daily Traffic
AIC	Akaike Information Criterion
AICc	corrected Akaike Information Criterion
AT	Active Transport
BLUP	Best Linear Unbiased Predictor
CCA	Canonical Correlation Analysis
CONTRAN	*Conselho Nacional de Trânsito*
CPM	Crash Prediction Models
CV	Cross-Validation
DENATRAN	*Departamento Nacional de Trânsito*
DNIT	*Departamento Nacional de Infraestrutura de Transporte*
DVMT	Daily Vehicle-Miles Traveled
DWF	Distance Weighted function
GIS	Geographic Information Systems
GLM	Generalized Linear Models
GLS	Generalized Least Squares
GWPR	Geographically Weighted Poisson Regression
GWR	Geographically Weighted Regression
HSM	Highway Safety Manual
IBGE	*Instituto Brasileiro de Geografia e Estatística*
ICD	International Classification of Diseases
IQR	Inter-Quartile Range
KED	Kriging with External Drift
KNN	K-Nearest Neighbor
MAUP	Modifiable Areal Unit Problem
MEI	Mean Imputation
MSPE	Mean Squared Prediction Error
MT	Motorized Transport
NB	Negative Binomial

NOTs	Number of Trips
OK	Ordinary Kriging
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
REML	Restricted Maximum Likelihood
RGP	*Regressão Geograficamente Ponderada*
RQ	Research Questions
RV	Regionalized Variables
SIM	*Sistema de Informações de Mortalidade*
SK	Simple Kriging
SP	São Paulo
SS	Sub-Sample
TAZ	Traffic Analysis Zone
VHT	Vehicle Hours Traveled
VIF	Variance Inflation Factor
VKT	Vehicle Kilometers Traveled
VMT	Vehicle Miles Traveled
WHO	World Health Organization
WLS	Weighted Least Squares

LIST OF CONTENTS

1 INTRODUCTION

Despite the efforts toward road safety, an estimated 1.25 million victims of road crashes still die every year, imposing a heavy burden on households and national economies. Developing countries, which have low and middle incomes, account for 90 percent of this number (World Health Organization [WHO], 2015), which is likely to rise even more if proper safety countermeasures and investments are not foreseen.

Among countries struggling to prevent road fatalities, Brazil alone is responsible for up to fifty thousand deaths and five hundred thousand injured every year. They are casualties of over one million crashes per year in the country (WHO, 2015). Unfortunately, associations between Brazil and high rates of road fatalities are commonplace in safety reports (such as in WHO, 2015; Job et al., 2015; AMBEV, 2017). As in other developing countries, this problem has been attributed to an insufficient development of supportive road infrastructure, policy changes and enforcement, which have not taken into account urban intensification and the steady increase in vehicle use. In spite of the growing awareness on the urgency to reverse these trends and efforts put into programs and campaigns toward road safety, the country's performance remains below expectations leading to an exponential rise in the number of casualties.

Road safety has long been a priority in developed nations, commonly ranked as “high road safety-performing countries”. In addition to ongoing efforts regarding the implementation and good practice of successful countermeasures involving infrastructure, vehicle and road user behaviors, developed countries have invested a great amount of time and money strengthening their road safety strategies at the planning level, e.g. by collecting and making comprehensive sources of reliable data available. Specifically in Europe, these strategies at both safety-planning and operational levels have led to a steady reduction in the number of deaths in most countries, therefore allowing the European fatality rates to decrease far below

the global average (9.3 per 100,000 population, relative to the global rate of 17.4) (WHO, 2015).

In developing countries, data unavailability has impaired model performance still at the planning level and, consequently, the tradeoff between input and outcomes. Unfortunately, this challenging condition has discouraged researchers and policy makers, restricting Crash Prediction Models (CPM) to data availability. However, even considering efforts made in associating crashes with the available explanatory variables in such circumstances, this drawback leads to higher modelling errors, and thus unreliable predictions. Besides being not statistically reliable, they fail in terms of impact analysis that could further help to implement appropriate safety countermeasures. One explanation could be the correlation among variables, i.e. multicollinearity and the existence of omitted variable bias, which in particular plays an important role in the reliability of CPM, generating biased and inconsistent estimates and coefficient signs (Washington, Karlaftis, & Mannering, 2010; Mitra & Washington, 2012). Besides, the spatial dependence of the data could be an important factor and must be considered. The investigation of different spatial data analysis approaches on crash prediction is desirable as it can lead to improvements toward safety-planning studies, and help policy makers target the best suitable techniques.

1.1 AIM AND MOTIVATION

In view of the above, this thesis aims to assess the potential improvements of supplementary data on spatial model performance, therefore highlighting the importance of a more diverse and comprehensive set of explanatory variables for crash modelling in both, prediction and impact analysis. Moreover, the suitability of different spatial modeling approaches is also evaluated. Hence, two specific objectives and five Research Questions (RQ) have been established, as follows:

I. To investigate the spatial model accuracy at unsampled subzones;
II. To evaluate the performance of spatial data analysis approaches on crash prediction.

RQ1. Based on a benchmarking exercise, what potential improvements in spatial model performance can be obtained by including additional explanatory variables?

RQ2. What are the statistical individual contributions of variables to the developed models?

RQ3. Are these models reliable?

RQ4. In case of data unavailability, would the produced models be suitable to estimate unsampled unit of areas?

RQ5. Considering geostatistics, by means of Kriging with External Drift (KED), what is the most suitable method to explore the spatial dependence of crash data and solve issues involving missing information?

Spatial CPM are a critical component in safety planning considering both prediction and impact analysis purposes. This argument is valid as CPM enable the estimation of values while providing an insight of the spatially varying relationship between crashes and several related factors (Kononov, 2002; Yau, 2004; Elvik, 2007). Specifically for researchers and planners from Brazil and other developing countries, this task still at the planning-level, has been a long lasting challenge as essential data is often unavailable. This drawback has restrained potential studies that could help to scrutinize the phenomena, for instance by identifying potential hotspots and influential factors, or by turning macro into local-level investigations, and thereby prioritizing countermeasures. Therefore, a study highlighting the importance of appropriate modelling techniques and input information, hence exploring both statistical and practical considerations of data within spatial model performance, is desirable toward road safety analysis and promotion.

In order to accomplish the established goals, analyses are firstly conducted based on a case study with crash/fatality-related available information from São Paulo (Brazil) and Flanders (Belgium). This enables us characterize the differences found in terms of road safety planning aspects in developing and developed countries. Macro-level CPM are developed within the

GWR framework with Poisson distribution of errors (GWPR). Subsequently, models are developed based on the Flemish dataset only, accounting for different multivariate spatial data analysis approaches, i.e. Mean Imputation (MEI), K-nearest neighbor (KNN) imputation and KED. Section 1.2 provides a brief overview of how this thesis is structured.

1.2 THESIS STRUCTURE

After contextualizing the problem about “road unsafety” and outlining the objectives and motivation for writing this manuscript, this introductory chapter ends by presenting the thesis structure (Figure 1.1). Hence, Chapter 2 gives a brief overview of spatial data analysis methods and crash-related information commonly used to develop CPM. In Chapter 3, we describe the study areas and data, as well as the proposed method framework. Chapters 4, 5 and 6 address the core chapters of this thesis, in which the in-depth analysis involving the specific objectives and corresponding research questions are put forward. Our conclusions and suggestions for future studies are drawn in Chapter 7.

Figure 1.1 – Thesis structure

CHAPTER 1 Introduction	▪ Contextualization ▪ Aim and motivation ▪ Thesis structure
CHAPTER 2 Literature Review	▪ Spatial statistics methods on crash prediction ▪ Crash-related explanatory variables
CHAPTER 3 Material and Method	▪ Study areas ▪ Data preparation ▪ Software packages ▪ Method framework
CHAPTER 4 Study of the impacts of enriched information on spatial model performance	▪ Potential improvements in model performance ▪ Sensitivity analysis ▪ Model's Validation ▪ Chapter discussion and conclusions
CHAPTER 5 Investigation of spatial model predictions accuracy at unsampled subzones	▪ GWR interpolation approach within the holdout method ▪ Missing-data imputation ▪ Chapter discussion and conclusions
CHAPTER 6 Evaluation of spatial data analysis approaches on crash prediction	▪ Kriging with External Drift ▪ Chapter discussion and conclusions
CHAPTER 7 Conclusions and recommendations	▪ Conclusions ▪ Suggestions for future research

2 LITERATURE REVIEW

This chapter describes the theoretical basis related to the main topics and spatial statistics models used in this thesis.

2.1 SPATIAL STATISTICS METHODS ON CRASH PREDICTION

Crash Prediction Models (CPM) play a significant role in traffic safety analysis, enabling for instance the identification of hotspots and sites where implementing countermeasures should be a priority. In this context, thanks to the scientific and technological advances and the availability of geocoded information, spatial analysis has emerged, leading to great prospects towards road safety.

On account of the first law of geography, which states that "Everything is related with everything else, but closer things are more related than distant things" (Tobler, 1970), spatial models have enabled a better understanding of the spatially varying relationship between road crashes and potential related information. Some of these models have also appealed as powerful tools to estimate values of an attribute at unsampled sites, accounting for known information only. In this respect, the appropriate choice of the triplet: modelling technique, Geographic Information Systems (GIS) and input information, are crucial tasks, as they directly affect the model performance, thus playing a significant role in planning, risk assessment and decision-making. Particularly concerning the modelling technique, Geographically Weighted Regression (GWR) outstands other approaches, which are based on the inherent spatial autocorrelation characteristics of the geographic observations only (e.g., kernel density interpolation, inverted distance weighted interpolation and univariate kriging interpolation methods). In this respect, GWR addresses both inherent characteristics of spatial data: spatial autocorrelation and spatial heterogeneity, thus accounting for the fact that variables are also correlated in the feature space (Fotheringham, Brunsdon, & Charlton, 2002). In contrast, geostatistical tools (e.g. kriging) by means of their intrinsic characteristics (e.g. semivariogram), are able to provide the Best Linear Unbiased Predictors (BLUP), meaning

estimates with minimum error and variance (Journel & Huijbregts, 1978; Matheron, 1963; Stein, 1999). Hence, implementing a geostatistical multivariate method could be a potential alternative to GWR and other spatial multivariate methods.

Geostatistical tools are usually applied to data with apparent spatial continuity, e.g. temperature, rainfall and land composition, in the fields of geology (Lee, Carle, & Fogg, 2007; Orton, Pringle, & Bishop, 2016; Tamayo-Mas, Mustapha, & Dimitrakopoulos, 2016), hydrology (Guven & Kitanidis, 1988; Goovaerts, 2000) and mining (Coburn, 2012), for example. However, over the last decades, its implementation on spatially discrete data (Goovaerts, 2006; Goovaerts, 2008) has proven to be a potential alternative when adapted to such spatial continuity problems. Therefore, geostatistics has become increasingly explored in different fields, e.g. health studies, where kriging techniques have been used for instance to identify areas of contamination or risk of mortality (Goovaerts, 2004, 2005, 2006, 2008, 2009). In transportation studies, its implementation has been explored in studies on traffic engineering (Ciuffo, Punzo, & Quaglietta, 2011; Mazzella, Piras, & Pinna, 2011; Zou, Yue, Li, & Yeh, 2012; Zhang & Wang, 2013), vehicle emission gases (Pearce, Rathbun, Aguilar-Villalobos, & Naeher, 2009; Kassteele & Velders 2006; Kassteele & Stein, 2006), and, more recently, to travel demand forecasting problems (Pitombo, Salgueiro, Costa, & Isler, 2015; Lindner, Pitombo, Rocha, & Quintanilha, 2016; Gomes, Pitombo, Rocha, & Salgueiro, 2016; Lindner & Pitombo, 2018). Specifically in traffic data, geostatistical tools have been implemented to analyze the spatial structure of the data under explanatory purposes (Majumdar, Noland, & Ochieng, 2004; Mcmillan, Hanson, & Lapham, 2007; Lascala, Johnson, & Gruenewald, 2001) or toward confirmatory analysis (Manepalli & Bham, 2011; Matsumono & Flores, 2013; Gundogdu, 2014; Molla, Stone, & Lee 2014).

In the next subsections, we provide a more detailed overview of these spatial approaches, together with a literature review addressing the input variables commonly implemented when estimating crashes.

2.1.1 GEOGRAPHICALLY WEIGHTED REGRESSION

GWR was developed by Fotheringham et al. (2002) intending to address the non-stationary relationship between variables found in Generalized Linear Models (GLM). In essence, GWR models capture this spatial variation by fitting a regression model, using a series of distance-related weights at each sample point. The result of this process is a set of local spatial parameters, described by Equation 2.1, varying over space, thus independent spatial error terms.

$$ln[E(C)(l_i)] = ln\big(\beta_0(l_i)\big) + \beta_1(l_i)ln(Exposure) + \beta_2(l_i)x_1 + \cdots + \beta_n(l_i)x_n \qquad (2.1)$$

Where $E(C)$ is the expected crash frequency, β_0, β_1, β_2, and β_n are model parameters for a determined location l_i. Exposure is the exposure variable, and x_1 and x_n correspond to other explanatory variables.

Motivated by Tobler's assumption, GWR assumes that the closer the observed data is from the location from the location of the parameter to be estimated, the greater the influence on the estimation of β at location i compared to those that are far from it. This influence is determined based on geographic weights, which are assigned in function of all neighboring observations using a kernel function (Fotheringham et al., 2002), e.g., Gaussian (Equation 2.2) and bi-square (Equation 2.3), which are the two most common choices (Hadayeghi, Shalaby, & Persaud, 2010).

$$W_{ij} = e^{-0.5(\frac{w_{ij}}{b})^2} \qquad (2.2)$$

$$W_{ij} = \begin{cases} (1-(\frac{d_{ij}}{b})^2)^2 & if\ d_{ij} < b \\ 0 & otherwise \end{cases} \qquad (2.3)$$

Where W_{ij} is the measure of contribution of location j when calibrating the model for location i. d_{ij} is the Euclidian distance between locations i and j, and b is the bandwidth size defined by a distance metric measure.

In GWR, the bandwidth controls the size of the kernel, i.e. the number of observations around each data point, and the rate at which weights decay with increasing distances. Thus, similar to the weighting scheme, the choice of the bandwidth size plays an important role in the performance of the GWR models, as it involves a trade-off between bias and variance. The size of the bandwidth is optimized either by distance (fixed kernel), or by the number of neighboring observations (adaptive kernel) (Fotheringham et al., 2002; Guo, Ma, & Zhang, 2008; Hadayeghi et al., 2010).

An optimum bandwidth can be found by minimizing the Cross-Validation (CV) score (Cleveland, 1979; Bowman, 1984) or the Akaike Information Criterion (AIC) (Akaike, 1973). Hence, a corrected version of the AIC (AICc) can be used, which unlike basic AIC is a function of sample size (Hurvich, Simonoff, & Tsai, 1998). While CV is given by the difference between observed and estimated values, AICc additionally to the statistical goodness-of-fit, rewards the complexity of the model, by imposing a penalty for increasing the number of estimated parameters (Fotheringham et al., 2002), expressed by the formulation in Equation (2.4).

$$AICc = D(b) + 2K(b) + 2\,\frac{K(b)(K(b)+1)}{n-K(b)-1} \tag{2.4}$$

Where D and K denote the deviance and the effective number of parameters in the model with bandwidth b, respectively. And n denotes the number of observations.

2.1.2 GEOSTATISTICS

Geostatistics refers to a set of spatial statistical methods, which enables the estimation of a variable value in locations where it is unknown. Developed by Matheron (1963), geostatistics is based on the theory of Regionalized Variables (RV) (Matheron, 1971; Wackernagel, 2003), which consists of a spatial structured and random component (Matheron, 1971). In general, geostatistics is better described by the following three procedures: (1) variographic analysis, (2) cross validation and (3) kriging, which are discussed as follows.

2.1.2.1 VARIOGRAPHIC ANALYSIS

The main motivation of the variographic analysis is to study the spatial structure of the RV. This inspection is conducted based on two key points: calculating the experimental semivariogram and adjusting the theoretical model.

In this respect, the primary task is to construct the semivariogram for the graphical representation of the spatial structure of the RV. The semivariogram function is defined as the arithmetic average of all squares of the differences between the pair's values separated by a distance h and a direction (Journel & Huijbregts, 1978), given by Equation 2.5.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{n(h)} [z\,(x_i) - z\,(x_i + h)]^2 \tag{2.5}$$

Where $N(h)$ the set of all pairwise data values $z\,(x_i)$ and $z\,(x_i + h)$ at spatial locations i and $i + h$, respectively (Matheron, 1963).

Therefore, the set of semi-variances derived from the function $\gamma(h)$ is plotted as a function of h (i.e. experimental semivariogram calculation). In order to calculate the experimental semivariogram, the establishment of some graphical aspects (e.g. lag distance, cut distance, lag tolerance and angular direction) is required (Matheron, 1971), for which:

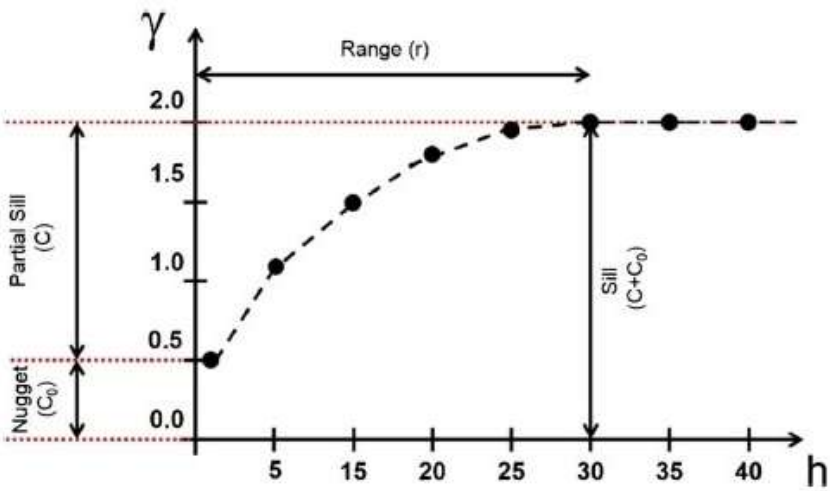
- lag distance: distance between neighboring samples;
- cut distance (h): the value of the distance at which the semivariogram is calculated, meaning that from that distance on, pairs of points are not considered. As a rule of thumb, the value adopted is generally half of the greatest distance between points in the sample;
- lag tolerance (Δh): values within the limits (h + Δh) and (h - Δh) are considered within h;
- angular direction (Θ): direction at which semivariograms represent the better spatial variability of the RV. Semivariograms may be calculated accounting for specific directions or not. In the first case, a semivariogram with an anisotropic structure implies

in different spatial variability in different directions, and a main direction, in which the variability is higher, compared to the others (Clark, 1979). In the second case, if there is isotropy, the experimental semivariograms are similar for all directions, thus "omnidirectional".

Thereafter, from the experimental semivariogram, a theoretical model is fitted enabling the representation of the continuous regional variability, which describes the overall trend. This practice is possible by manual or automatic fitting. However, while in the former the choice of the best model is usually given by the visual appearance of the experimental semivariograms, the latter, accounts for analytical methods. These methods can be summarized in two categories: (1) Maximum likelihood methods, e.g. Restricted Maximum Likelihood (REML); and (2) Least Square methods, e.g. Ordinary Least Squares (OLS), Weighted Least Squares (WLS), Generalized Least Squares (GLS) and others to mention (Cressie, 1985, 1993; Wackernagel, 2003).

From the various theoretical models for adjustment of the experimental semivariograms, the most frequently used are Spherical, Gaussian and Exponential. Such models can be described based on the following important parameters: nugget effect (C_o), partial sill (C) and range (r), illustrated on Figure 2.1 and described as follows (Matheron, 1963, 1971; Wackernagel, 2003, Lindner & Pitombo, 2018).

Figure 2.1 - Graphical parameters of a semivariogram



(Lindner & Pitombo, 2018 adapted from Wackernagel, 2003)

By definition, the nugget effect reflects the residual of the variance of sampling errors and the spatial variance at short distances. The point at which the increasing function of the semivariogram stabilizes is called Sill, equivalent to the sample variance. Hence, the value of the distance at which the sill is reached is called range and it represents the average distance around a point, to which there is still some degree of spatial autocorrelation, thus from that point on, values are no longer spatially correlated (Wackernagel, 2003; Li & Heap, 2008; Duarte, Calvo, Borges, & Scatoni, 2015). Thereafter, the defined values of C_0, C_1, and range for the theoretical semivariograms are used in the weighting scheme, at the geostatistical modelling step where the interpolations are carried out.

2.1.2.2 CROSS VALIDATION

Cross Validation (CV) compares various assumptions, either concerning the model (e.g. type of function to be adjusted, variogram parameters) or the data. In the cross validation procedure, each sample value $Z(x_i)$ is removed in turn from the dataset and a value $z^*(x_i)$ at the location is estimated using the remaining $n-1$ samples. The difference between a data value and the estimated value $(Z(x_i) - Z^*(x_i))$ gives an indication of how well the data value fits into the neighborhood of the surrounding data values (Journel & Huijbregts, 1978; Wackernagel, 2003).

2.1.2.3 KRIGING ESTIMATION

Kriging is the interpolation method used in geostatistics to estimate one or more variables, appealing to provide the BLUP for variables that have the tendency to vary over space (Journel & Huijbregts, 1978; Matheron, 1963). Kriging presupposes that points that are spatially close tend to have values that are more similar to points, which are far apart. Such influence is determined on account of geographic weights, which are produced based on the graphical parameters of the theoretical semivariograms, within an area established by an ellipsoid with radii determined by the ranges of the major and minor directions (Matheron, 1971; Wackernagel, 2003). Such kriging estimators are defined by Equation 2.6 as follows:

$$\hat{z}\,(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i) \tag{2.6}$$

Where $\hat{Z}$ is the estimated value of an attribute at the point of interest x_0, λ_i are the weights and z is the observed value at the sample point x_i. In order to ensure unbiasedness of the estimators, the sum of the weights must be 1 (Webster & Oliver, 2007), expressed in Equation 2.7.

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{2.7}$$

Simple Kriging (SK) and Ordinary Kriging (OK) are the most usual univariate kriging methods. SK is used when the average is assumed to be statistically constant in the sample area, and OK for the contrary (Goovaerts, 1997; Armstrong, 1998). Considering the purposes of the analyses stated in this thesis, we implement a multivariate interpolation tool, namely Kriging with External Drift (KED). KED enables the use of a secondary variable to co-estimate a correlated one. Hence, considering the integration of two correlated variables (Z(x) and Y(x)), which express the same attribute, estimates at new locations are made as a function of the linear function in Equation 2.8 (Armstrong, 1998; Wackernagel, 2003).

$$E[Z(x_0)] = a_0 + b_1 Y(x_0) \tag{2.8}$$

Where $Y(x_0)$ is the external drift function to estimate the primary variable $Z(x_0)$ based on the estimated values x_0. The variance of the estimation and their corresponding weights are enabled by the matrix shown in Equation 2.9 (Wackernagel, 2003).

$$\begin{bmatrix} C & 1 & Y \\ 1^T & 0 & 0 \\ Y^t & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ -\mu \\ -\mu \end{bmatrix} = \begin{bmatrix} C_0 \\ 1 \\ Y_0 \end{bmatrix} \tag{2.9}$$

Where C is the covariance function, and μ is the Lagrange multiplier that minimizes the variance of the estimation and both constraints in Equations 2.6 and 2.7.

2.2 CRASH-RELATED EXPLANATORY VARIABLES

Besides estimation purposes, spatial CPM are powerful tools able to provide a full understanding of the spatially varying relationship of crashes and related variables across the study area. Therefore, selecting a comprehensive and correct set of independent variables is a crucial task and has been a particular concern to decision makers and researchers. As a result, road crashes can be attributed to six major groups of risk factors (Kononov, 2002; Valent et al., 2002; Yau, 2004; Shankar, Mannering, & Barfield, 1995; Delen, Sharada, & Bessonov, 2006; Elvik, 2007):

- **human factors (driver behavior)** - commonly associated to driver behavior (e.g. alcohol and drug use, negligent and careless operation of the vehicle, failure to properly use protection devices, use of telephone or texting while driving and fatigue) (see Petridou & Moustaki, 2000; Odgen, 1996; Redelmeier & Tibshirani, 1997; Movig et al., 2004);
- **vehicle related factors** - refers to the characteristics of the vehicle and safety design standards for its performance, e.g. Active (activated before the road crash takes place, so that they could avoid accidents) and Passive Vehicle Safety Systems (used to avoid or mitigate injuries and their severity), such as air bags and safety belts (see Harvey & Durbin, 1986; Robertson, 1996; Langley, Mullin, Jackson, & Norton, 2000; Bédard, Guyatt, Stones, & Hirdes, 2002; Richter, Pape, Otte, & Krettek, 2005);
- **traffic volumes** - commonly represented by the Average Annual Daily Traffic (AADT) or the Vehicle Miles Traveled (VMT). Both parameters are often used as exposure variables in CPM (see Hauer, 1995; Zhou & Sisiopiku, 1997; Martin, 2002; Qin, Ivan, & Ravishanker, 2004; Pei, Wong, & Sze, 2012; Xu, Liu, Wang, & Li, 2012; Ahmed, Abdel-Aty, & Yu, 2012; Pirdavani, Brijs, Bellemans, Kochan, & Wets 2013);
- **road design (geometry)** – refers to road geometries and roadside conditions. Variables within this group are commonly related to the forgiving and self-explanatory character of roads (Brookhuis et al., 2006; Lotz et al., 2006; Wiethoff et al., 2012) such as appropriate road categorization, well-designed curves and grades, wide lanes, adequate sight distance, appropriate design speeds, clearly visible striping, flared guardrails, good quality shoulders, roadsides free of obstacles, well-located crash attenuation devices and well-planned use of traffic signals (Miaou, 1994; Taylor, Lynam,

Baruya, 2000; Amundsen & Ranes, 2000; Kloeden, Ponte, & McLean, 2001; Karlaftis & Golias, 2002; Nilsson, 2004; Aarts & van Schagen, 2006; Rengarasu, Hagiwara, & Hirasawa, 2007);

- **environmental related factors** - weather and light conditions, for example (see Shankar et al., 1995; Andrey & Knapper, 2003; Golob & Recker, 2003; Ahmed et al., 2012; Brijs, Karlis, & Wets, 2008);
- **time factors** – related to the season of the year, the month of the year, weekdays and the time of crash occurrence (see Doherty, Andrey, & MacGregor, 1998; Qin, Ivan, & Ravishanker, 2006; Hao, Kamga, & Wan, 2016).

On account of this and given the challenges to obtain most of these variables, the following recommendations have been suggested in previous literature and can help to provide an adequate model. According to Elvik (2007), CPM should include variables that:

- have been found in previous studies;
- have a major influence on the dependent variable;
- are not correlated to other variables in the dataset;
- that can be measured in a valid and reliable way;
- are not endogenous;
- and above all, consider what people are exposed to that could result in an accident as absence of an exposure variable can lead to biased results, since other variables, correlated with exposure will suffer from omitted variable bias. This supports previous findings in the literature (Carroll, 1971, Chapman, 1973; Hauer, 1982, 1995; Hauer, Ng, & Lovell, 1996; Qin et al., 2004, de Guevara, Washington, & Oh, 2004; Elvik, 2007). Examples of traffic crash risk exposure measures are AADT or Vehicle Kilometers Traveled (VKT), commonly used in CPM (Miaou, Song, & Mallick, 2003; Fristrøm, Ifver, Ingebrigtsen, Kulmala, & Thomsen, 1995; Jovanis & Chang, 1986; Ahmed et al., 2012; Pirdavani, Brijs, Bellemans, Kochan, & Wets, 2012; Pirdavani et al., 2013a).

Certainly, detailed variables, such as those related to driving data (e.g. acceleration, braking and steering information, driver response to stimuli), if incorporated in the models would help

to better identify cause and effect relationships regarding crash occurrence (Lord & Mannering, 2010). However, the meticulous details of these variables make them expensive to collect and obtain. As a result, these variables are hardly available for consultation, thus rarely implemented in CPM. In this context, crash analysis has been performed at different levels of aggregation, depending on their purposes, e.g. microscopic, for a specific road segment or intersection, or at macroscopic-level for a larger area, such as a municipality (Pirdavani, Daniels, van Vlierden, Brijs, & Kochan, 2017). We suggest Huang et al. (2016) for more detailed information concerning both aggregation levels. Specifically for macro-level analysis, efforts have been made to associate crashes and predictive variables that have macro-level characteristics, such as socioeconomic, exposure and network variables (e.g. area, population, traffic volume variables, for instance AADT and VMT, road length, ratio or number of vehicles, degree of urbanization, speed limit, income, gender, level of education, age, trip generation, employment rate, poverty, etc.). A more compiled review of these variables can be found in Pulugurtha, Duddu and Kotagiri (2013) and Rhee, Kim, Lee and Ulfarsson (2016).

In order to support the findings and arguments in this research, we conducted a search of the scientific literature from the last decades. Results of this investigation are presented below together with a description of variables and methods, which have been used to predict road crashes.

Golob and Recker (2003) applied linear and nonlinear multivariate statistical analyses to relate crashes on freeways in Southern California to traffic flow, weather and lighting conditions. Principal Component Analysis (PCA) was used to identify the most significant variables from a set of original traffic flow variables, and a Canonical Correlation Analysis (CCA) was used to relate the identified principal components to weather and lighting conditions. Findings include associations between left lane collisions and dry roads during day light, while off road to driver’s right collisions were associated to wet road at night. Moreover, results emphasized the relationship between type of collision and variations in speed for left and interior lanes. Concerning the severity of crashes, results indicated a higher influence of volume, rather than speed, when controlling weather and lighting conditions.

Hadayeghi, Shalaby and Persaud (2003) developed a series of macro-level prediction models to estimate the number of accidents in 463 Traffic Analysis Zones (TAZs) in the city of Toronto as a function of zonal characteristics. Firstly, the authors used a GLM (Negative Binomial) separately for total accidents and for severe (fatal and non-fatal injury) accidents as a function of socioeconomic/demographic, traffic demand and network data variables. Subsequently, authors used GWR to explore the spatial variations in the estimated parameters from the zones. Results revealed that the number of accidents per zone in a year increased as the zonal VKT, major and minor road kilometers, total employed labor force, household population and intersection density increased, and decreased with higher posted speed and higher congestion in the zone.

Aguero-Valverde & Jovanis (2006) related both fatal and injury crash data from Pennsylvania to socioeconomic, transportation infrastructure-related and environmental related factors. In the models developed for fatal crashes, only three transportation related variables were found to be significant: Daily Vehicle-Miles Traveled (DVMT), percentage of travel on federal aid roads, and infrastructure mileage. The first two were found to have a negative correlation with crash frequency, while infrastructure mileage was found to have a positive correlation. Concerning the socioeconomic variables, information regarding area deprivation, percentage of population under poverty, as well as persons younger than 15 years were found to be significant and positively correlated with crash frequency. Within the environmental variables, only precipitation was found to be significant, and positively correlated. In the models developed for the total injury crashes, four transportation-related variables were found to be significant, i.e. DVMT, infrastructure mileage, mileage density and percentage of federal aid roads. In these models, only DVMT was found to have a negative correlation with the dependent variable, therefore suggesting an increasing risk of injury crashes at a decreasing rate.

Caliendo, Guida and Parisi (2007) modelled crash frequency on multilane roads in Italy as a function of traffic flow, infrastructure geometry, pavement surface and rainfall information. Authors used Poisson, Negative Binomial (NB) and Negative Multinomial regression models, which were applied separately to tangents and curves. Analyses included both total crashes

and injury crashes including fatalities and were conducted based on results of a 5-year monitoring period on a four lane median-divided highway.

Brijs et al. (2008) studied the effect of weather conditions on daily crash counts using a discrete time-series model (Poisson Integer-Valued Autoregressive - INAR). Analyses were made for three large cities in the Netherlands (Dordrecht, Haarlemmermeer and Utrecht). Thereafter, the authors compared the results of model performance to classical models (Poisson regression and negative binomial regression). The influence of weather conditions on road crashes, can also be found in Andrey & Knapper (2003).

Quddus (2008) associated crashes with variables related to traffic and road characteristics, and socio-demographic factors. Ward-level casualty data were splitted by severity of casualties (i.e. fatalities, serious injuries and slight injuries) and by severity of the casualties related to various road users. Their spatial units of analysis were the 633 census wards from the Greater London metropolitan area. Based on non-spatial negative binomial models and spatial Bayesian hierarchical models at census ward level, results revealed that households with no cars and total employment were statistically significant variables to predict crashes. Results from the Bayesian hierarchical modeling showed they were more consistent with the literature and more coherent in all cases. Findings from this analysis also include a positive association found between traffic flow and casualties.

Hadayeghi et al. (2010) developed GWPR models to investigate the local spatial variations in the relationship between the number of zonal collisions for total and severe (i.e. fatal and injury) collisions and potential transportation planning predictors such as traffic volume, road network characteristics, socioeconomic and demographic features, land use, dwelling unit and employment type. Thereafter, the authors compared the accuracy of these models to that of GLMs. VKT was used as the exposure variable in all the models. Results revealed positive correlations between VKT and the dependent variables in most subzones and models, for both categories. Positive correlations were also found for the number of schools, total arterial road kilometers, total expressway road kilometers, total collector kilometers, total number of

signalized intersections, while negative correlations were found for total rail kilometers and total local road kilometers, for example.

Matkan, Mohaymany, Mirbagheri and Shahri (2011) used GWPR to develop a local safety model for Mashhad, Iran. Models were developed with trip production and attraction information, as well as VKT, considered as the exposure variable. Subsequently, authors compared the goodness of fit of these models to those obtained with GLM. Findings indicated positive correlations between crashes and the three variables for most of the 253 TAZs included in the analyses.

Ahmed et al. (2012) investigated the impact of geometrical, traffic and weather variables on the occurrence of crashes on a mountainous freeway in Colorado (United States). A Bayesian logistic regression model was used. The modelling results revealed that the geometric factors were significant during the dry and snowy seasons, but that during the snow season, low visibility, high precipitation and speed variation increased the likelihood of accidents while for the dry season low average speeds and low visibility increased the odds of an accident.

Pirdavani et al. (2012) developed CPM to associate the number of injury crashes with different exposure, network, and sociodemographic variables at the zonal level for 2,200 TAZs in Flanders, Belgium. To this, NB models were developed within the GLM framework, on the basis of different measures of exposure used as independent variables, i.e. Number of Trips (NOTs), Vehicle Hours Traveled (VHT) and VKT. Models were categorized into the following groups: (a) flow-based models, (b) trip-based models, and (c) a combination of the two. Results revealed that, the models that contained the combination of two exposure variables outperformed the models calibrated with only one of the exposure variables.

Xu et al. (2012) used conditional logistic regression models to examine the relationships between crash risks and traffic characteristics (termed by the authors as traffic states). Results of that investigation revealed that traffic flow parameters had different effects on safety for every traffic state. For instance, the average downstream occupancy seemed to reduce

accident risk in two traffic states (in congested traffic, as well as in transition from free flow to congested flow) but caused an increase in the overall model.

Li, Wang, Liu, Bigham and Ragland (2013) used GWPR models to capture spatially varying relationships between fatal crashes and traffic patterns, road network attributes, and socio-demographic characteristics of 58 counties in California (United States). Thereafter, the performance of GWPR was compared to a traditional GLM. Findings revealed negative correlation between the risk of fatal crashes in most subzones, for the percentage of freeway mileage, road density, traffic intensity, percentage of urban traffic. On the contrary, positive correlations were found for population density and percentage of truck and trailers.

Pulugurtha et al. (2013) developed NB count models (with log-link) at TAZ level. Models were developed as a function of several land use variables, based on land use characteristics (e.g. mixed use development, urban residential, single-family residential, multi-family residential, business and office districts) from the city of Charlotte, Mecklenburg County, North Carolina (United States). Except for the single-family residential area, positive correlations were found between land use characteristics and the total number of crashes. According to the authors, such a negative correlation, was possibly due to different behaviors adopted by drivers (such as cautious driving) or lower travel speed in these areas.

Shariat-Mohaymany, Shahri, Mirbagheri and Matkan (2015) related the total number of crashes to other traffic volume, network characteristics and trip generation variables of 253 TAZs in Mashhad (Iran). GWPR and GLM approaches were carried out in that investigation. The spatial models suggested positive correlation for all variables, which were included in the models in most subzones. These variables were VKT (used as the exposure variable in its logarithm form), total main street length, NOTs and number of non-signalized intersections.

Lee, Abdel-Aty, Choi and Huang (2015) used Bayesian Poisson Lognormal Simultaneous Equations Spatial Error Model (BPLSESEM) to identify the contributing factors for “pedestrian crashes per crash location ZIP code area” and “crash-involved pedestrians per residence ZIP”. The set of variables included information of population, VMT, income, proportion measures

of high-speed roads, children for a specific group age, people working at home, households without a vehicle and households below poverty level, and other facility/attraction information.

Rhee et al. (2016) used GWR to identify factors related to crashes in Seoul, Korea. Traffic, road networks, demographic and socioeconomic information were used. Findings showed that increasing ratio of a central bus-only has the major contribution to the number of crashes. Moreover, results revealed a decrease in crash frequency for an increase of roads with speed limits below 30 km/h.

Xu, Huang, Dong and Wong (2017) investigated the spatially varying relationships between crash frequency and related risk factors using a fully Bayesian approach. Explanatory variables included road and traffic-related factors, such as the DVMT, trip production and attraction, intersections and road segment lengths with various speed limits, and a number of factors reflecting the demographic and socioeconomic features. The authors conducted a case study using a three-year crash dataset from the Hillsborough County, Florida (United States). DVMT, NOTs and population were included as exposure variables in the model. Results revealed that the coefficients were all significantly positive, implying that more severe crashes were expected in zones with higher concentrations of traffic volumes, travel demand and residents. These findings were in line with other previous studies, as reported by Huang et al. (2016), Pirdavani et al. (2012) and Lee et al. (2015).

Table 2.1 details the explanatory variables and techniques, which were investigated in the above mentioned previous studies. Based on this investigation, it was possible to identify some research gaps, which helped us outline the goals of this thesis. We noted, for instance, the predominance of micro-level studies. Few studies have tended to focus on the relationship between crashes and explanatory variables at a more aggregated level. Especially for countries where data availability is an issue, the development of macro-level models (e.g. zonal and municipal), could assist in a long-term transportation planning processes by the identification of potential hotspots together with their major influential factors. This practice could be useful, as it would help policy makers to prioritize hot zones and data collection

simultaneously. Moreover, there is a considerable amount of studies on the performance of spatial models (e.g. GWR) over tradition non-spatial models (e.g. GLM). In light if the evident spatial character of road crashes, we believe that research on the spatial correlation of the data is needed. Furthermore, there has been little discussion on spatial data analysis approaches, other than GWR, for example. Lastly, we were able to identify some typical variables used at macro-level models, and which were found to be significant predictors of crashes (e.g. traffic volume, VHT, VKT, population, employment, level of income, urbanization degree, traffic intensity, number of intersections and intersection density, speed, and road length). This helped us later on support the findings and arguments in this research.

Table 2.1 - Summary of explanatory variables used in the CPM of the previous studies

Study	Techniques	Explanatory variables
Golob and Recker (2003)	PCA and CCA	**Traffic flow** (e.g. variation in volume and occupancy of the lanes), **weather** (dry or wet weather), **lighting conditions** (daylight, darkness, and dusk–dawn)
Hadayeghi et al. (2003)	GLM (NB) and GWR	**Socioeconomic and demographic** (total population, population density, number of households, household density, full-time employed, part-time employed, total employed, employment density, number of vehicles, number of vehicles per household); **network or supply** (number of intersections, intersection density, major road kilometers, minor road kilometers, total road kilometers, area); **traffic demand** (e.g. speed, vehicles flow, VKT)
Aguero-Valverde & Jovanis (2006)	Full Bayes (FB) hierarchical spatial models and Traditional NB	**Socioeconomic** (population accounting for age, sex, level of poverty and level of drunk driving); **transportation-related** (e.g. daily VMT - DVMT, road, road density); **environment-related** (precipitation, number of rainy days, snowfall, number of days with snow)
Caliendo et al. (2007)	Poisson, Negative Binomial and Negative Multinomial regression models	**Infrastructure geometry**: horizontal alignment (tangent or curve), vertical alignment (upgrade or downgrade), **weather and pavement surface conditions** (dry or wet), **number of vehicles and persons involved,** and a short description of the **accident dynamics**
Brijs et al. (2008)	Discrete time-series model and GLM (Poisson regression and NB regression)	**Weather** (Precipitation, temperature, sunshine, city specific dummies); **traffic** (daily vehicle counts, VKT)
Quddus (2008)	Non-spatial NB models and Spatial Bayesian hierarchical models	**Traffic** (traffic flow, speed); **road characteristics** (link length); so**ciodemographic** (age, employment population and car ownership)
Hadayeghi et al. (2010)	GWR and GLMs with Negative Binomial and Poisson error structures	**Traffic** (speed, VKT, average volume over capacity); **road network** (number of rail stations, total rail kilometer, number of schools, total arterial kilometers, total expressway kilometers, total collector kilometers, total collector kilometers, total local road kilometers, total ramp kilometers, number of signalized intersections); **employment types** (full and part time employees, number of not employed, employment sector); **demographic** (age, gender, in possession of driver's license or not, in possession of a transit pass or not); **land use** (commercial, governmental, institutional, residential, open area, parks and recreation, resource and industrial, water body), **dwelling unit** (number of houses, number of apartments, and number of townhouses)

Continue

Conclusion

Study	Techniques	Explanatory variables
Matkan et al. (2011)	GWR and GLM	**Traffic** (trip production and attraction information, VKT)
Ahmed et al. (2012)	Bayesian logistic regression model	**Geometrical** (grade, degrees of curvature, width); **traffic** (speed); **weather** (visibility)
Pirdavani et al. (2012)	GLM (NB)	**Exposure** (VKT, VHT, NOTs); **road network** (speed, capacity, number of intersections, TAZ in urban or suburban area); **sociodemographic** (income level, population)
Xu et al. (2012)	Conditional logistic regression models	**Traffic** (number of vehicles, speed and traffic occupancy)
Li et al. (2013)	GWR and GLM	**Traffic** (DVMT, traffic intensity, percentage of urban DVMT, percentage of trucks and trailers); **road network** (percentage of freeway mileage, road density); **sociodemographic** (population density, age, income)
Pulugurtha et al. (2013)	Negative binomial (with log-link)	**Land use** (e.g. mixed use development, urban residential, industrial, business, single-family residential, multi-family residential, office districts, institutional, neighborhood service development, right-of-way, commercial center, innovative, planned unit development, research district)
Shariat-Mohaymany et al. (2015)	GWR and GLM	**Traffic and road network** (VKT, total main street length, NOTs and number of non-signalized intersections
Lee et al. (2015)	Bayesian Poisson lognormal simultaneous equations spatial error model (BPLSESEM)	**Demographic** (population, proportion of children for an specific age); **roadway/traffic** (VMT; proportion of high-speed roads); **commute** (proportion of people working at home); **socioeconomic** (proportion of households without available vehicle, proportion of households below poverty level, income); **facility/attraction** (number of rail and bus stations, number of hotels, motels, and guest houses, number of marina/ferry terminals, number of schools)
Rhee et al. (2016)	GWR	**Traffic** (VKT); **road network** (speed limits, number of subways stations and bus stops, ratio of length of central bus lane versus total road length, number of access points with over 30 km/h absolute difference in posted speed limit); **demographic** (population age); **socioeconomic and neighborhood information** (income, number of elementary and lower schools, mixed use of land, ratio of large condominium/apartment complexes)
Xu et al. (2017)	Fully Bayesian approach	**Road and traffic-related** (DVMT, trip production and attraction, intersections and road segment lengths with various speed limits; **demographic and socioeconomic** (total population, population for specifics age groups, income, travel mode)

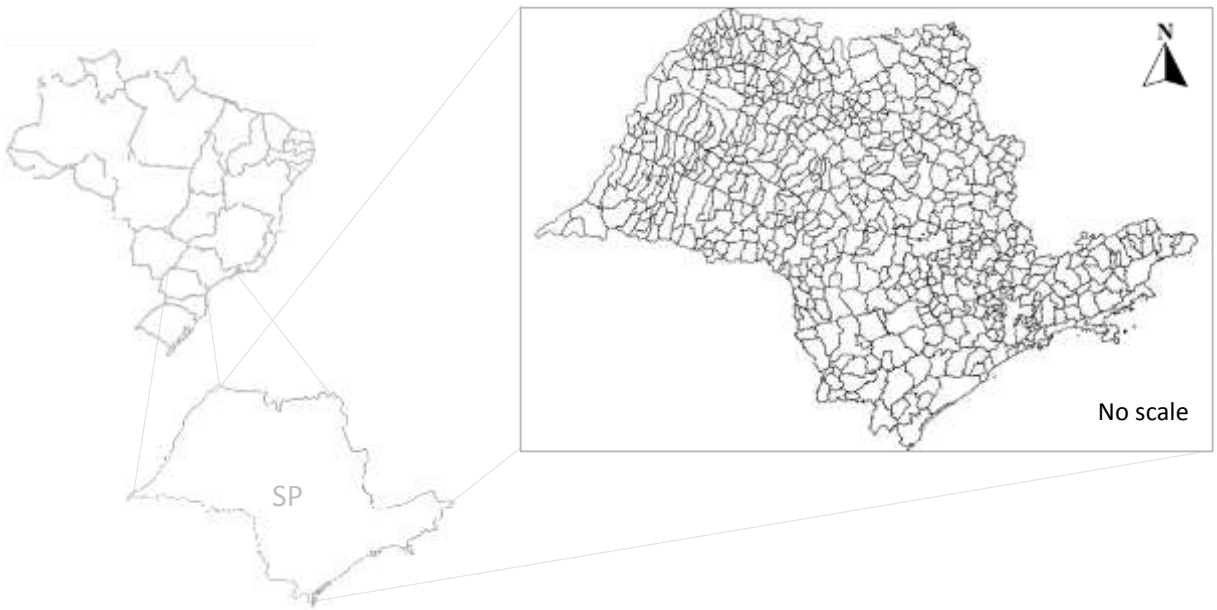
3 MATERIALS AND METHOD

This chapter provides an overview of both study areas, and the data and software packages used to construct the prediction models. Moreover, the methodological framework to achieve the aims of this thesis is presented.

3.1 STUDY AREAS

São Paulo (SP) is one of the 26 states that comprises the Federative Republic of Brazil. Located in the southeast region of the country, São Paulo has a total of 645 municipalities and the greatest population in the country. In the last census, in 2010, statistics pointed to around 41 million inhabitants in a land area close to 248,000 square kilometers. However, in 2017, this number had already exceeded 45 million (IBGE, 2018). In terms of road networks, the state has more than 35,000 kilometers of roads that transport thousands of passengers and freight vehicles every day. Unfortunately, in 2015, more than six thousand people died in São Paulo, who were victims of road crashes on these roads, (IBGE, 2018; DATASUS, 2018). Figure 3.1 shows the division of Brazil by states (in the upper-left), as well as the study area (in the upper-right).

Figure 3.1 – Study area in Brazil



As can be seen in Figure 3.2, Belgium is divided in three regions, i.e. Flanders, Wallonia and Brussels. Subsequently, the Flemish and the Walloon regions are each subdivided in five provinces. In this study, Crash Prediction Models (CPM) for Belgium are developed based on available information of Flanders, which is a Dutch-speaking region in northern Belgium (Figure 3.3).

Figure 3.2 – Administrative regions and provinces in Belgium

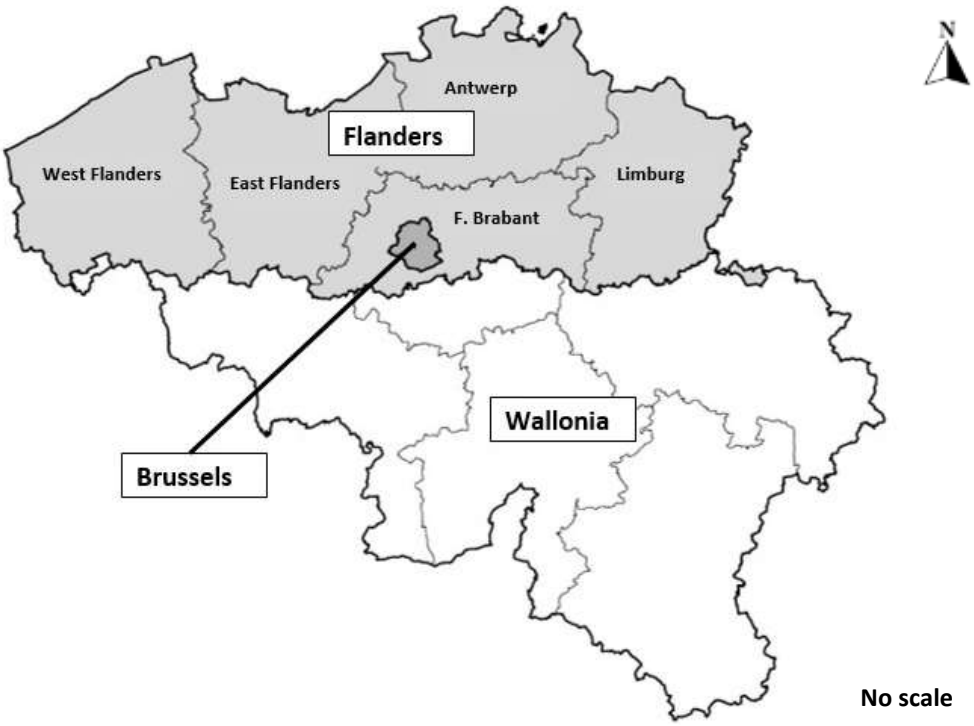
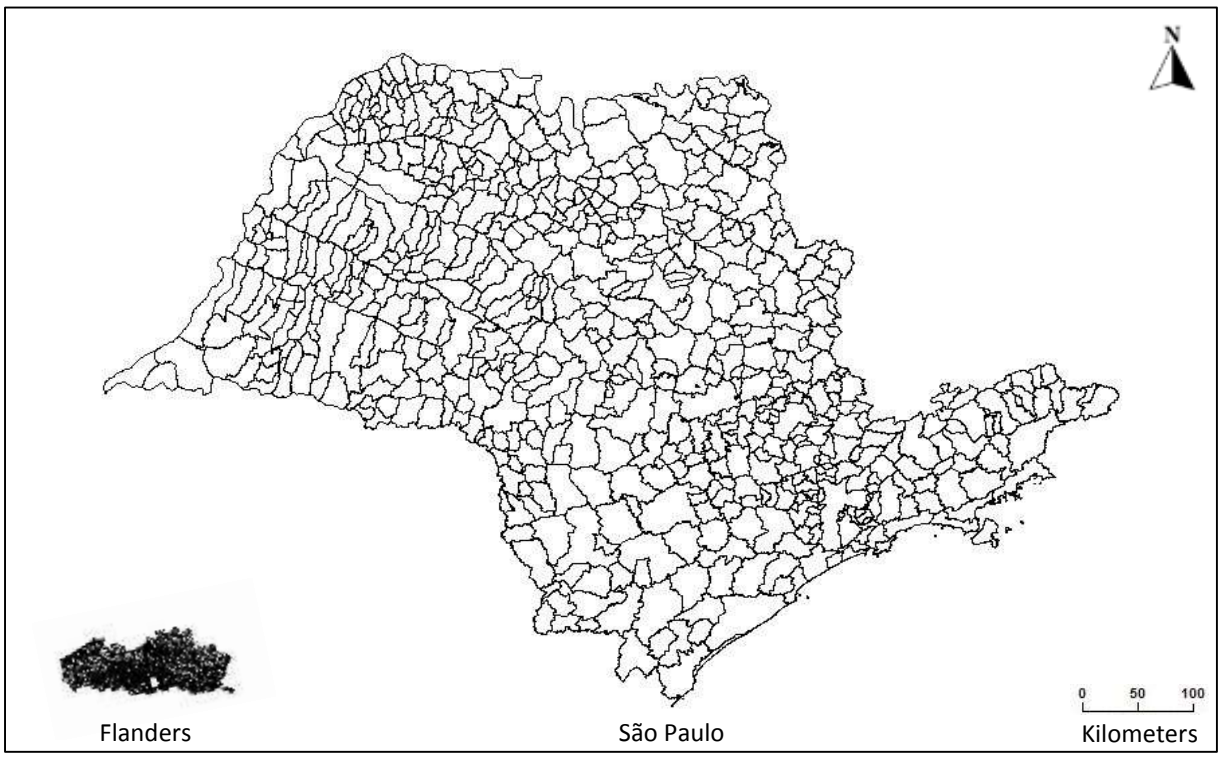


Figure 3.3 - Study area in Belgium

Compared to São Paulo, Flanders is considerably smaller in population and size, with a population of 6.5 million (about 60% of the population of the country) in a land area of approximately 14,000 square kilometers (VLAANDEREN, 2018). However, in terms of road networks, due to its central location in Europe, Flanders is recognized for having one of the densest road networks worldwide (5.08 km per square km). From a total of 71,500km, 883km are highways, and 6,040km comprise regional roads. Figure 3.4 presents both study regions in the same scale. In this respect, São Paulo is about 18 times as big as Flanders.

Figure 3.4 - Study areas in scale



In the next subsection, a brief overview of the prospect of road crashes in Brazil and Belgium is given, adding to a better understanding of the historical evolution of road safety in both regions of study and their linkages with the statistics nowadays. More detailed information can be found in the road safety reports made available by WHO (2015) and OECD (2016), in which major information used in this research was collected.

3.1.1 PROSPECT OF ROAD CRASHES AND SAFETY IN BRAZIL

"Road unsafety" has long been a huge problem in Brazil. Despite the growing awareness concerning the urgency to reverse trends and put efforts into programs and campaigns toward road safety promotion, the country has not managed to lessen the number of road fatalities, showing even increasingly higher numbers (WHO, 2015; Job et al., 2015; AMBEV, 2017).

In general, up to fifty thousand people die and five hundred thousand are injured every year, who are victims of the more than one million accidents on Brazilian roads (WHO, 2015). Historically, this began after World War II when motorized vehicles were introduced on to Brazilian roads, and thereafter it was reinforced by a rapid economic growth. In addition to this, in the 1960s, Brazil experienced important political and social changes, which granted rapid industrial expansion and agricultural modernization of the country. As a consequence, several million people left the rural fields and moved to urban areas, leading to rapid urban intensification. In 1990, there were only 20.6 million vehicles in Brazil, including 1.5 million motorcycles (IBGE, 2018; DENATRAN, 2018). From the 1990s, encouraged by the federal government, manufacturing, acquisition and use of motorized vehicles increased, especially motorcycles, given the fact that they were inexpensive and versatile. Since then, motorization rates have not stopped increasing, shifting the mobility pattern of Brazil's low and middle-income populations from public to private transport.

Unfortunately, infrastructural developments, policy changes and level of enforcement have not kept pace with vehicle use, and it has led to a chaotic situation. The number of fatalities involving motorcyclists, for instance, which was 725 in 1996, increased to 12,604 in 2014, meaning an increase of approximately 1640%. Fatalities involving passenger car occupants had an increase of 166%, approximately, for the same period (DATASUS, 2018), as presented in Table 3.1 below.

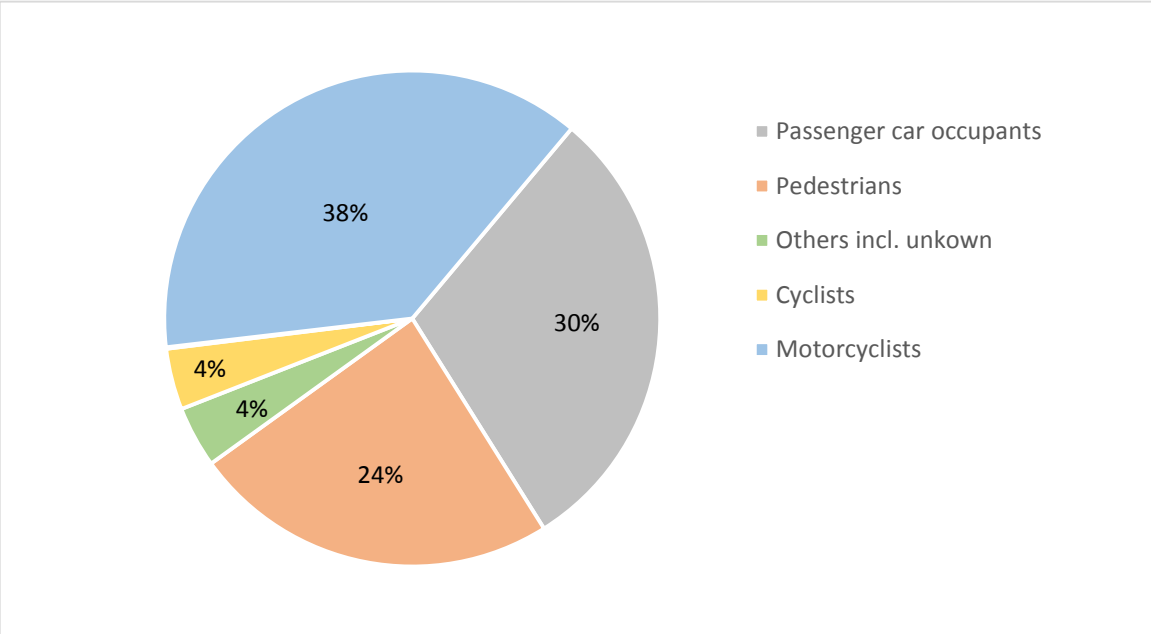
Table 3.1 - Road fatalities by road user group

	1996	2000	2010	2013	2014	2014 % change from			
						2013	2010	2000	1996
Cyclists	326	789	1513	1348	1357	0.67	-10.31	71.99	316.26
Tricycle occupants	22	27	69	57	48	-15.79	-.30.43	77.78	118.18
Motorcyclists	725	2465	10825	11983	12604	5.18	16.43	411.32	1638.48
Passenger car occupants	3778	5266	9059	9757	10084	3.35	11.31	91.49	166.91
Pedestrians	12952	8696	9944	8220	8082	-1.68	-18.72	-7.06	-37.60
Others	473	791	1282	1318	1402	6.37	9.36	77.24	196.41
Total	**17950**	**18034**	**32692**	**32683**	**33577**	**2.74**	**2.71**	**86.19**	**87.06**

Adapted from DATASUS (2018)

As a consequence of the sharp increase in motorcycle production and sales, motorcyclists have long been placed at the top of the list of fatalities by group user. As can be seen in Figure 3.5, in 2014, they were responsible for 38% of the total number of fatalities, followed by 30% of fatalities involving car occupants, 24% involving pedestrians, and 8% involving cyclists and other types of vehicles.

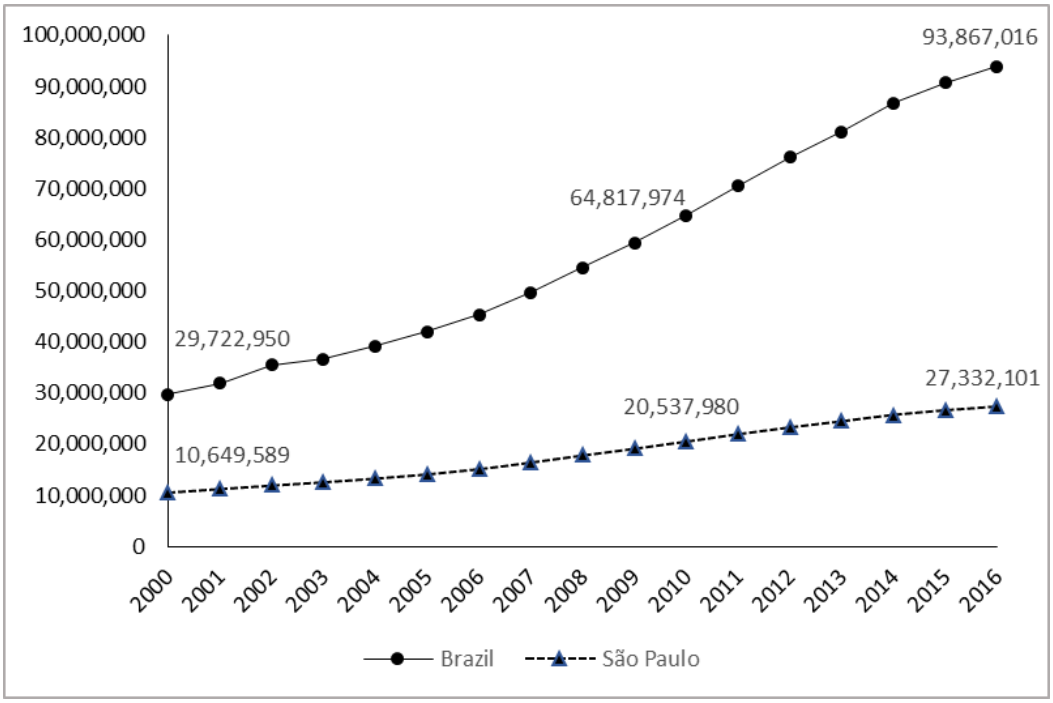
Figure 3.5 - Road fatalities in 2014 by group user in percentage



Adapted from BRSNR (2016)

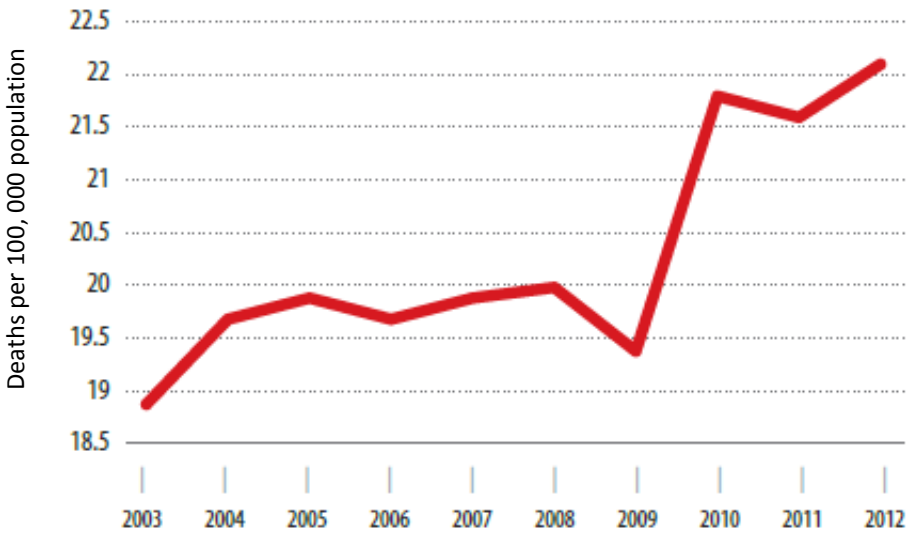
Assuming the evolution of the production of motorized vehicles in Brazil, over the last decades, this number shifted around 215%, from 2000 to 2016, and around 160% assuming the production only in the state of São Paulo, shown in Figure 3.6 (DENATRAN, 2018). In terms of fatality rate per 100, 000 population, it increased 14% assuming the period from 2004 to 2012 (WHO, 2015), shown in Figure 3.7.

Figure 3.6 - Evolution of the production of motorized vehicles in Brazil and São Paulo state



Adapted from DENATRAN (2018) – *Sistema Nacional de Registro de veículos* - RENAVAM

Figure 3.7 - Trends in reported road traffic deaths in Brazil

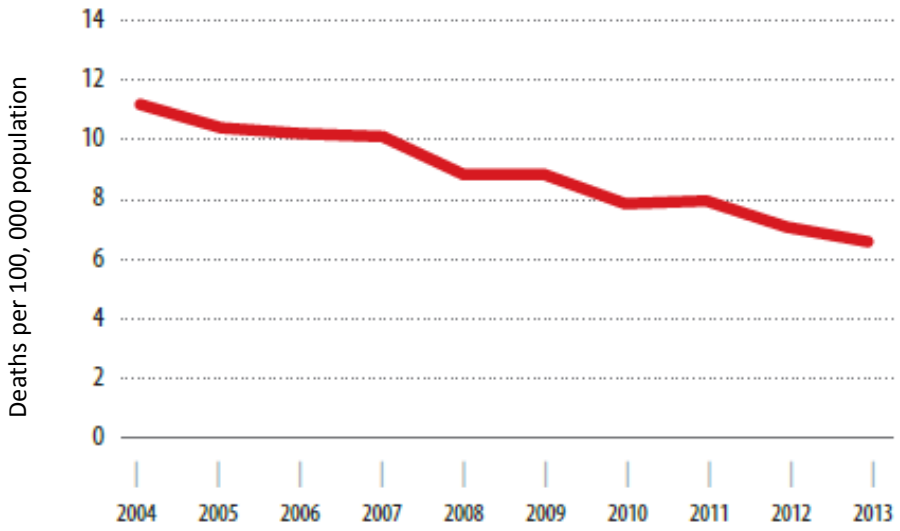


WHO (2015)

3.1.2 PROSPECT OF ROAD CRASHES AND SAFETY IN BELGIUM

Contrary to increasingly high statistics in Brazil, the number of road traffic fatalities in Belgium has significantly dropped over the last decades. In spite of a period of stagnation, in the late 1990s, after the first national assembly on road safety (*Etats Généraux de la Sécurité Routière/Staten-Generaal van de Verkeersveiligheid*) in 2001, road safety became an issue of public interest. Strategies discussed in that convention contributed to improvements in infrastructure, enforcement and education, and since then, the number of fatalities in Belgium has mostly dropped, as shown in Figure 3.8. Some small variations are attributed to changes in the economic situation (decrease in 2008) or meteorological conditions (2010 – 2011, 2014) (WHO, 2015).

Figure 3.8 - Trends in reported road fatalities in Belgium



WHO (2015)

Compared to 1990, in 2014 the number of fatalities, seriously injured and injury crashes had already decreased by more than 60%, 75% and 30%, respectively. In terms of rates, in the same period, decreases of more than 65% and 80% were found for the road traffic mortality rate (expressed by deaths per 100,000 population) and risks (expressed in deaths per number of vehicles), respectively, as presented in Table 3.2.

Table 3.2 - Reported road safety data in Belgium

	1990	2000	2010	2013	2014	2014 % change from			
						2013	2010	2000	1990
Fatalities	1976	1470	840	724	727	0.4	-13.5	-50.5	-63.2
Injury crashes	62446	49065	45927	41279	41481	0.5	-9.7	-15.5	-33.6
Injured persons hospitalized	17479	9847	5984	4947	4502	-9.0	-24.8	-54.3	-74.2
Deaths per 100, 000 inhabitants	19.9	14.4	7.7	6.5	6.5	0.0	-16.3	-54.8	-67.3
Deaths per 100, 000 registered vehicles	4.3	2.6	1.3	1.0	1.0	0.8	-22.5	-62.0	-77.4
Deaths per billion vehicle kilometers	28.1	16.3	8.5	7.1	na				

Adapted from OECD (2016)

In terms of traffic statistics, the distance traveled by motorized users and the number of vehicles has considerably increased, by more than 45% and 54%, respectively, in relation to 1990 (OECD, 2016), as shown in Table 3.3.

Table 3.3 - Traffic data in Belgium

	1990	2000	2010	2013	2014	2014 % change from			
						2013	2010	2000	1990
Registered vehicles (thousands)	4594	5735	6689	6994	7076	1.2	5.8	23.4	54
Vehicles kilometers (millions)	70276	90036	98678	102423	Na				
Registered vehicles per 1,000 inhabitants	462	560	617	627	632	0.8	2.4	12.8	36.8

Adapted from OECD (2016)

Some strategies adopted in Belgium that has contributed to the progress of the country toward road safety include (OECD, 2016):

- enforcement of lower speed limits on many rural roads and stricter control of speed limits on other roads;
- black-spot treatment;

- improvements toward infrastructure;
- improvements in safety systems in car and trucks;
- better road safety awareness through campaigns and educational measures.

Since 1990, these countermeasures have helped to reduce the number of fatalities of all road user groups (OECD, 2016). Especially among pedestrians, cyclists, moped users and passenger car occupants, fatality figures reduced by 60% to 85%, approximately, in relation to 2014 (Table 3.4). If we assume the trends over the last decade, in the period between 2010 and 2014, road fatalities reduced by 13.5%.

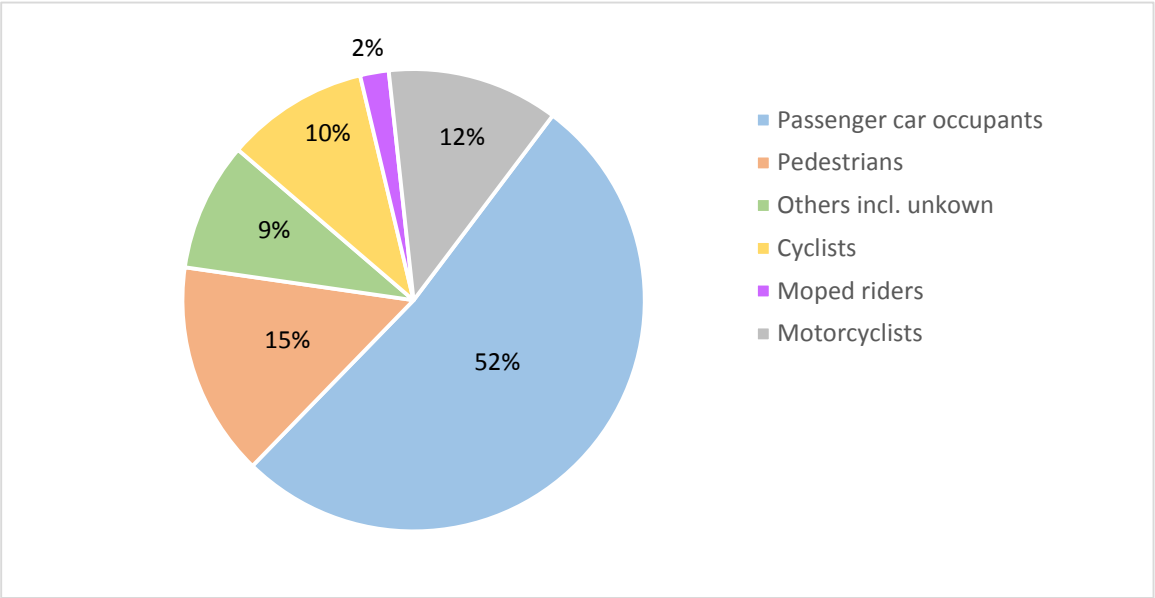
Table 3.4 - Road fatalities by road user group

	1990	2000	2010	2013	2014	2014 % change from			
						2013	2010	2000	1990
Cyclists	196	134	70	73	76	4.1	8.6	-43.3	-61.2
Moped users	110	64	22	13	17	30.8	-22.7	-73.4	-84.5
Motorcyclists	106	118	102	102	85	-16.7	-16.7	-28.0	-19.8
Passenger car occupants	1181	922	444	342	381	11.4	-14.2	-58.7	-67.7
Pedestrians	301	142	106	99	106	7.1	0.0	-25.4	-67.8
Others	82	90	96	95	62	-34.7	-35.4	-31.1	-24.4
Total	**1976**	**1470**	**840**	**724**	**727**	**0.4**	**-13.5**	**-50.5**	**-63.2**

Adapted from OECD (2016)

In 2014, passenger car occupants were responsible for 52% of the total number of fatalities (Figure 3.9). In that year, the major number of fatalities (around 37%) occurred on rural roads, 26% in urban areas and 13% on motorways (the remaining 24% of fatalities took place on unknown roads). Since 1991, the greatest reduction in fatalities has occurred on rural roads (-74%), followed by urban areas, where the reductions are around 70% (OECD, 2016).

Figure 3.9 - Road fatalities in 2014 by group user



Adapted from OECD/IT (2016)

3.1.3 ROAD SAFETY STRATEGIES, TARGETS AND LEGISLATION – BRAZIL VS. BELGIUM

In response to the rising numbers of road fatalities worldwide and aiming to guide efforts at national and local levels, the World Health Organization (WHO) and the United Nations regional commissions, in cooperation with the United Nations Road Safety Collaboration and other stakeholders developed a global plan of action, including guidelines and targets for road safety promotion in the current decade (WHO, 2015).

In March 2010, the United Nations General Assembly resolution 64/255 proclaimed that 2011-2020 would be the "Decade of Action for Road Safety". In this plan, both the Brazilian and the Belgium governments set the ambitious target of reducing the number of road deaths by half by 2020 (WHO, 2015), implying less than 420 road fatalities in Belgium and approximately 10 thousand in Brazil. In order to achieve this, the federal minister for mobility in Belgium has made additional efforts and taken measures in terms of monitoring and legislation, for instance by enforcing more severe penalties. While predictions based on past developments, show that Belgium will be able to get close to the imposed target, in Brazil, this has been questioned.

Table 3.5 compiles some traffic decrees and enforcement, which have been applied in both countries. We gathered such information from WHO (2015) and CONTRAN (2007).

Table 3.5 - Legislation according to the traffic decrees in Belgium and Brazil

	Belgium	**Brazil**
Maximum speed limits authorized	▪ Urban roads: 30-50km/h ▪ Rural roads: 70-90km/h ▪ Highways: 120km/h	▪ Urban roads: 80km/h ▪ Rural roads: 60km/h ▪ Highways: 110km/h
Maximum authorized blood alcohol content	▪ General population: 0.5 g/l ▪ Professional drivers: 0.2 g/l	Zero tolerance
Use of hand-held phones while driving	Forbidden	Forbidden
Use of hands-free devices while driving	Authorized	Authorized
Seat-belt	Mandatory for both, front and rear seats since 1975 and 1991, respectively	Mandatory for both, front and rear seats, since 1997
Child restraint system	Mandatory for passengers under 18 and smaller than 135 cm (since 2006). *They can travel either in the front or rear seat if the child restraint system conforms to the latest European standards*	Since 2008, the Brazilian traffic decree specifies that: ▪ Passengers under 10 must travel in the rear seat ▪ Passengers under 1 are obligated to travel in an adapted child restrain device specific for their age ▪ Passengers older than 1 and under 4 are obligated to travel in an adapted child restrain device specific for this age group ▪ Passengers older than 4 and under 7.5 are obligated to travel in an adapted child restrain device specific for this age group
Helmet-use law	▪ Required for all riders of motorized two-wheelers ▪ Motorcyclists (> 50 cc) also have to wear gloves, boots that protect the ankle and long sleeved/legged jacket and trousers ▪ There is no mandatory helmet-use law for cyclists	▪ Required for all riders of motorized two-wheelers ▪ There is no mandatory helmet-use law for cyclists

Other key points involve, for instance, the enforcement of more severe penalties for the non-use of seat belts. Despite the fact that seat belts are mandatory in the front and rear seats, wearing-rates in Brazil are still low. According to Household Survey National Research (PNAD), this number was 73% and 37% in 2008 for front and rear seats, respectively. In contrast to Brazil, in Belgium, the seat belt wearing rate observed in the front seats was around 80% in 2008. In 2015, this number had already exceed 92% and 85% for front and rear seats, respectively (Lequeux, 2016). Moreover, according to the Belgian Institute for Road Safety - Road Safety Knowledge Centre, the helmet-wearing rate in 2012 was 99% for drivers and passengers, while in Brazil this number was 81% in the same year (WHO, 2015).

The objective here is not to compare Brazil and Belgium or the study regions, São Paulo and Flanders. Nevertheless, we believe that developed countries, such as Belgium, should be taken as examples. Their ongoing efforts toward steady improvements, even when their performance has been better than so many other nations, should be used as a motivation for countries, such as Brazil to put more efforts into changing their trends. Moreover, such a comparison would not be suitable for the purposes of this research. This argument is valid given the differences found in the dependent variables (fatalities associated to travel mode in São Paulo, and casualties associated to travel mode in Flanders), and the aggregation levels (municipalities for models developed in São Paulo, and subzones for models developed in Flanders). As previously mentioned, the main objective of this study involving the two regions, is to assess the potential impacts of enriched information on spatial model performance, and herewith emphasize the necessity of a more diverse and complete dataset to predict road crashes. Thus, the reporting of the information that could improve the Flemish models could at most be interesting suggestions for extra data collection in Brazil.

3.2 DATA PREPARATION

Spatial macro-level CPM were developed based on geographical available information of crashes, road-networks, socioeconomic and demographic variables for both regions of study. Whereas CPM for São Paulo were developed to estimate the number of road fatalities in the

state, in Flanders, we considered road casualties as our dependent variable due to the features of road crashes in Belgium, i.e. higher number of crashes with minor injuries in relation to fatalities (see Table 3.2). Hence, both response variables were divided into two sets each based on the travel mode, called Active Transport (AT) and Motorized Transport (MT). Causalities/fatalities for active transport included pedestrians and cyclists, while for motorized transport they were associated with motorized vehicle occupants. Moreover, records from a period of three years were used to produce the dependent variables.

3.2.1 SÃO PAULO DATABASE

Information collected from São Paulo was geographically aggregated to each of 644 municipalities that are comprised by the state (São Paulo city itself was not included in the analyses given its atypical values, which are far higher than the ones for other cities).

Police and hospitals are two common sources of crash and casualty data, in Brazil. However, none of them is able to provide a full and effective data source of the accidents and fatalities in the country. Furthermore, there is no link between their databases, thus affecting the consistency of the data, and making their collection a challenge (Job et al., 2015). With respect to the coverage of the data, the best national one comes from the Health Ministry Database (DATASUS), which has the official records of road deaths, in Brazil (Job et al., 2015). Yet, the official numbers are understated by 20% (WHO, 2013). In view of this, fatality figures for São Paulo, were collected from the Mortality Information System (*Sistema de Informações de Mortalidade – SIM*), which is a public source created by DATASUS (DATASUS, 2018). Fatalities, as the response variable, was developed based on the total number of deaths for the period between 2009 and 2011. Since 1996, road fatalities have been coded using the International Classification of Diseases in its 10th revision (ICD-10) as recommended by WHO. They are classified under the V-codes in different groups, according to the mode and cause. In this research, fatalities for active transport mode, included the concerning information within the codification between V01-V19, while for motorized transport we included the information within the codification between V20-V79. Table 3.6 presents the classification per groups. This information is provided in detail in Appendix A.

Table 3.6 - Transportation accident codes according to ICD-10

Code	Category
V01 - V09	Pedestrian injured in transport accident
V10 - V19	Cyclist injured in transport accident
V20 - V29	Motorcycle rider injured in transport accident
V30 - V39	Occupant of three-wheeled motor vehicle injured in transport accident
V40 - V49	Car occupant injured in transport accident
V50 - V59	Occupant of pick-up truck or van injured in transport accident
V60 - V69	Occupant of heavy transport vehicle injured in transport accident
V70 - V79	Bus occupant injured in transport accident

DATASUS (2018)

Corresponding socioeconomic and demographic information was gathered from the last census index of 2010, made available by the Brazilian Institute for Geography and Statistics (IBGE, 2018). In spite of a great amount of available information within these categories, during the exploratory analysis, most pieces of information collected were found to be correlated with each other, therefore presenting high degrees of multicollinearity. In this context, only the information regarding the number of inhabitants, which revealed to be the most significant variable in these categories, was included in the spatial CPM.

Given the limitations to obtain road features information, we used the road network of São Paulo provided by OpenStreetMap (OSM, 2018), as it provides a user-friendly interface and a database rich in information about road characteristics. Available information of link length was included as a proxy variable of the road network. The link length for motorized transport included information of trunk, highway, primary, secondary and tertiary roads, as well as link length of residential and living streets. For active transport, we used the same road features, although highways and respective trunk length information were replaced by cycle path and link length information of other roads designed only for pedestrians. Table 3.7 and 3.8 present the list of variables that have been collected for the municipalities in the state of São Paulo, together with their definition and descriptive statistics. Variables, which were included in the final Brazilian models, are marked in bold.

Table 3.7 - Description of the variables collected for São Paulo

Variable	Description
Active transport	Total number of fatalities of active transport mode users over 3 years
Motorized transport	Total number of fatalities of motorized transport mode users observed over 3 years
Link length of active transport	Total link length of active transport in a municipality (km)
Link length of motorized transport	Total link length of motorized transport in a municipality (km)
Population	Total number of inhabitants in a municipality
Area	Total surface area in a municipality (km^2)
Male population	Total number of male inhabitants in a municipality
Female population	Total number of female inhabitants in a municipality
Population density	Total population per square kilometers in a municipality
AAGR	Average Annual Growth Rate 2000-2010 (%) in a municipality
Percentage male population	Percentage of male inhabitants in a municipality
Percentage female population	Percentage of female inhabitants in a municipality
Percentage proportion population	Rate between the number of men and woman in a municipality
Urban population	Total number of inhabitants in the urban zone of a municipality
Rural population	Total number of inhabitants in the rural zone of a municipality
HDI	Human Development Index of a municipality
GNP	Gross National Product in a municipality
Employed people	Total number of inhabitants with income in a municipality
Occupied people	Total number of inhabitants who perform some activity (with income or not) in a municipality
Motorcycle	Total fleet of motorcycles and tricycles in a municipality
Microbus	Total fleet of microbuses in a municipality
Car	Total fleet of cars in a municipality
Truck	Total fleet of trucks in a municipality
Bus	Total fleet of buses in a municipality
Total number of vehicles	Total number vehicles in a municipality
Gasoline	Total gasoline consumption in a municipality
Diesel oil	Total diesel oil consumption in a municipality
Fuel oil	Total fuel oil consumption in a municipality
GLP	Total liquefied petroleum gas consumption in a municipality
Ethanol	Total ethanol consumption in a municipality

Table 3.8 - Descriptive statistics of variables collected for São Paulo

	Variable	Average	Min	Max	SD[a]
Fatality	**Active transport**	7.34	0	295	23.428
	Motorized transport	11.88	0	317	28.716
Network	**Link length of active transport**	141.13	5.04	2626.15	210.95
	Link length of motorized transport	153.04	5.29	2831.03	227.25
	Area	383.03	5.4	1977	317.07
Socioeconomic and demographic	**Population**	46597.35	805	1221979	108465.83
	Male population	22902.55	422	595043	52538.13
	Female population	23694.81	383	626936	55938.55
	Population density	291.13	3.73	12519.10	1166.18
	AAGR	1.03	-2.15	10.92	1.25
	Percentage male population	50.52	45.76	81.09	2.52
	Percentage female population	49.48	18.91	54.24	2.52
	Percentage proportion population	102.97	84.36	428.86	17.88
	Urban population	44150.48	627	1221979	107468.51
	Rural population	2446.88	0	46284	3609.38
	HDI	0.739	0.639	0.862	0.032
	GNP	22501.11	7131.54	287646.17	18418.14
	Employed people	12678.37	155	405980	35725.41
	Occupied people	14931.77	211	471267	41144.02
Vehicle fleet	Motorcycle	4744.68	24	100831	10938.16
	Microbus	90.76	0	3544	264.87
	Car	13536.09	133	487044	38052.31
	Truck	705.21	11	18144	1544.29
	Bus	135.84	3	4445	330.34
	Total of vehicles	19212.58	220	612097	50296.09
Fuel consumption[b]	Gasoline (liters)	7961187.11	0	256246033	21723939.41
	Diesel oil (liters)	15343179.63	0	295769873	32673917.02
	Fuel oil (liters)	822438.64	0	44127640	3078410.70
	GLP (liters)	2304087.98	0	62823861	5948082.76
	Ethanol (liters)	9746540.07	0	342168947	25378940.38

[a]SD: Standard Deviation; [b]Fuel consumption in liters

The descriptive statistics for São Paulo including São Paulo city is available in Appendix B.

3.2.2 FLEMISH DATABASE

Spatial CPM for Flanders were developed at zonal level, comprising 2,198 Traffic Analysis Zones (TAZs), considered as the unit of analysis in the Flemish models. The average size of TAZs is 6.09 square kilometers with a standard deviation of 4.78 kilometers, and an average number of inhabitants equal to 2,416 persons (MOBIEL VLAANDEREN, 2018).

Casualties, as the response variable, consisted of the concerning information for the period of three years from 2010 to 2012. Information on the Flemish models were collected with the Flemish Ministry of Mobility and Public Works (MOBIEL VLAANDEREN, 2018). Likewise for São Paulo, we used the link length information, provided by OMS (2018). The same road features used in that case, were considered in the Flemish models.

The Flemish dataset, in addition to significant information related to socio-economic, socio-demographic and road networks provided foremost diverse and suitable exposure variables, i.e., Number of Trips (NOTs), vehicles flow and Vehicle Kilometers Traveled (VKT). Table 3.9 displays a list of variables that have been collected for Flanders and included in the final Flemish models (marked in bold). Table 3.10 summarizes their descriptive statistics.

Table 3.9 - Description of the variables collected for Flanders

Variable	Description
Active transport	Total number of casualties of active transport mode users in a TAZ over 3 years
Motorized transport	Total number of casualties of motorized transport mode users in a TAZ over 3 years
Capacity	Hourly average capacity of links in a TAZ (Passenger car per direction/h)
Link length of active transport	Total link length of active transport in a TAZ (km)
Link length of motorized transport	Total link length of motorized transport in a TAZ (km)
Intersection density	Number of intersections per square kilometer
NOTs of active transport	Average daily number of trips originating/destined from /to a TAZ involving active mode
NOTs of motorized transport	Average daily number of trips originating/destined from /to a TAZ involving motorized transport
VKT - Highway	Total vehicles kilometers traveled on highways in a TAZ
VKT - Other roads	Total vehicles kilometers traveled on roads other than highways in a TAZ
Car ownership	Car ownership per household in a TAZ
School children	Total number of children living in a TAZ that attend some school
Population	Total number of inhabitants in a TAZ
Speed	Average speed limit in a TAZ (km/h)
Area	Total surface area of a TAZ (km^2)
Link density	Total link length in a TAZ (km^2)
Intersection	Total number of intersection in a TAZ
Highway	Presence of a highway in a TAZ described as: “No” represented by 0, “Yes” by 1
Urban	Is the TAZ in the urban area ? “No” represented by 0, “Yes” represented by 1
Suburban	Is the TAZ in the suburban area? “No” represented by 0, “Yes” represented by 1
Households	Total number of households in a TAZ
Employees	Total number of employed people in a TAZ
Income level	Average income of residents in a TAZ described as: “Monthly salary less than 2249 euro” represented by 0, “Monthly salary more than 2250 euro” represented by 1

Table 3.10 - Descriptive statistics of variables collected for Flanders

	Variable	Average	Min	Max	SD[a]
Casualties	**Active transport**	15.04	0	298	25.06
	Motorized transport	45.36	0	500	53.83
Network	**Capacity**	1790.10	1200	7348	554.60
	Link length of active transport	14.85	0	88	10.31
	Link length of motorized transport	15.87	0.39	87.95	10.80
	Intersection density	1.75	0	50.63	3.37
	Speed	69.40	31	120	10.91
	Area	6.09	0	45	4.78
	Link density	3.37	0	20.44	2.41
	Intersection	5.80	0	40	5.90
	Highway		0	1	
	Urban		0	1	
	Suburban		0	1	
Exposure	**NOTs of active transport**	1103.40	0	8630	1316.12
	NOTs of motorized transport	2750.09	0	22650	2642.17
	VKT - Highway	27471.82	0	946153	84669.53
	VKT - Other roads	26662.85	0	303238	28133.04
Socioeconomic and demographic	**Car ownership**	1.13	0	14.00	0.47
	School children	364.09	0	92.45	772.59
	Population	2614.53	0	15803	2582.60
	Households	1091.15	0	8062	1177.90
	Employees	888.73	0	16286	1575.31
	Income level		0	1	

[a]SD: Standard Deviation

3.3 SOFTWARE PACKAGES

Spatial CPM and related tasks concerning the proposed analyses of this research were carried out using the software packages listed and described in Table 3.11 as follows.

Table 3.11 - Software packages used in the research

Software	Application
IBM SPSS 24	Non-spatial statistics (characterization of the database through descriptive measures, histograms, hypothesis test)
GWR 4.0	Construction of CPM
Geostatistical Modelling Software - geoMS – version 1.0	Construction of CPM, visualization of the point's map, adjustment of experimental variograms, kriging and cross validation
QGIS 2.18.23	Preparation of the themed and kriging maps (continuous surface)
TransCAD	Preparation of the themed maps (local maps)

Whereas the continuous surface maps resulting of the geostatistical analyses in geoMS were produced in QGIS, maps resulting from the local GWR models were produced by means of TransCAD.

3.4 METHODOLOGICAL FRAMEWORK

The proposed methodological framework was structured attempting to answer the five Research Questions (RQ) that together form the general and specific objectives of this research. Taking this into account, the core of the thesis lies in Chapters 4, 5 and 6 at which we addressed these objectives and corresponding RQ. Figure 3.10 presents an overview of this investigation. Subsequently, we briefly describe the tasks concerning the investigations in each chapter.

Figure 3.10 – General objective and research questions

General objective: To assess the potential impacts of supplementary data on spatial model performance and the suitability of different spatial modeling approaches
Research questions: **RQ1.** Based on a benchmarking exercise, what potential improvements in spatial model performance can be obtained by including additional explanatory variables? **RQ2.** What are the statistical individual contributions of variables to the developed models? **RQ3**. Are these models reliable? **RQ4.** In case of data unavailability, would the produced models be suitable to estimate unsampled unit of areas? **RQ5.** Considering geostatistics, by means of Kriging with External Drift (KED), what is the most suitable method to explore the spatial dependence of crash data and solve issues involving missing information?

In Chapter 4, analyses allowed us assess the potential impacts of enriched information on spatial model performance (RQ1 and RQ2) and verify the reliability of the produced GWPR models (RQ3). To this end, casualties were firstly associated with all available variables for São Paulo and the corresponding ones for Flanders. Models developed at this stage were called “basic models”. In the next step, prediction models were developed only for Flanders considering all the available information in the Flemish dataset. These models were called “improved models”. Due to data unavailability we could not conduct this exercise for São Paulo. Secondly, a sensitivity analysis was carried out to identify the individual statistical contribution of the input information in the casualty prediction. Lastly, model accuracy was assessed by the corresponding goodness of fit obtained with a pre-determined validation sample. This was possible by adjusting the Flemish improved models within the repeated holdout method. In this respect, prediction models were developed using only part of the data and a different part to validate them. Since GWR models are local models, each subzone has its own prediction model with its unique coefficient estimates. Unlike Generalized Linear

Models (GLM) models where you create a single model and can easily feed it with the input data to validate the constructed model, in GWR we possess several local models to be validated. Therefore, it was a challenge to validate these models. This validation exercise is of great importance especially when we would like to use the existing models to estimate the expected crash frequency of zones with missing data or even zones without any information. In this investigation, we carried out the analysis in the following order and called it "GWR holdout1":

1. GWR models are developed for 70% of the subzones.
2. A Distance Weighted Function (DWF) is used to create local coefficient estimates for each of the missing subzones (i.e. validation subzones).
3. Input data (i.e. measures of the explanatory variables) of the validation subzones are used to estimate casualties in these subzones.
4. Casualty estimates obtained in step 3 are compared with observed number of casualties to validate these models.

Aiming to investigate the spatial model prediction accuracy at unsampled subzones (RQ4), in Chapter 5, we extended the empirical evaluation in the previous chapter and took one-step further. This time we estimated casualties of the missing subzones based on the casualty estimates produced for 70% of the data. Here we tried to test the following hypothesis: for subzones without any information, would it be better to use the casualty estimates of the neighboring subzones instead of using their coefficient estimates? This implies that, we not only used weighted coefficient estimates of the surrounding subzones, but also used input information of the neighboring, inspired by the first law of geography stating that *"everything is related to everything else, but near things are more related than distant things".* We named this approach "GWR holdout2" and performed the analysis in the following order:

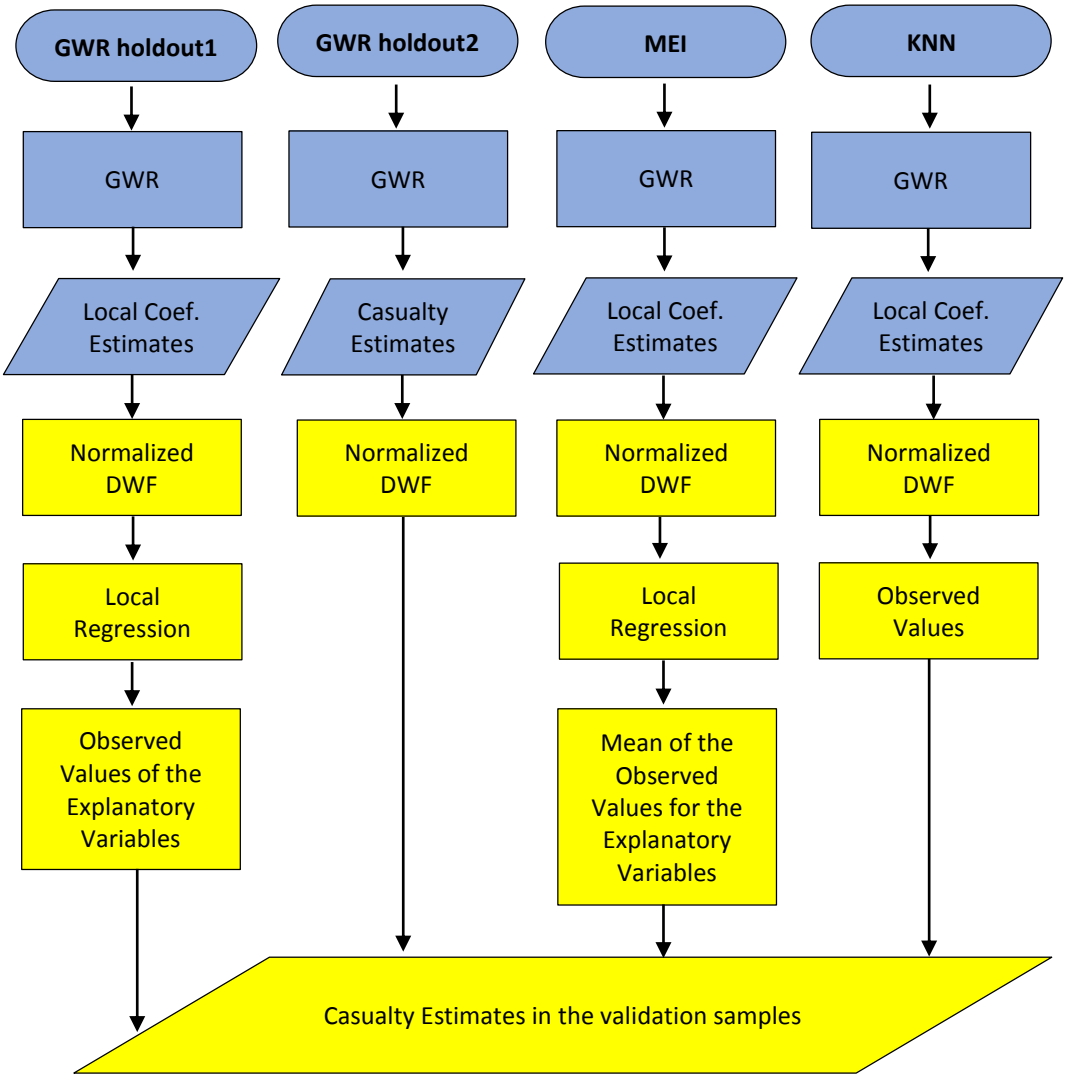
1. GWR models are developed for 70% of the subzones.
2. Expected casualties are calculated for these subzones.

3. A DWF is used to create an estimated casualty for each of the missing subzones (i.e. validation subzones).
4. Casualty estimates (from step 3) are compared with observed number of casualties to validate these models.

Additionally and , in order to verify the validity of this novel GWR validation approach, results of model performance were compared to those obtained with two missing data imputation approaches, i.e. Mean imputation (MEI) and K-nearest neighbor (KNN) imputation. Models within these approaches were developed based on the mean of the observed values (MEI), and the observed values of the explanatory variables itself (KNN). In other words, while the former was given by the imputation of one single value for each explanatory variable, the latter, was given by the imputation of the observed values for each validation sample.

Figure 3.11 summarizes the tasks and outcomes for the GWR holdout and imputation approaches. Tasks that were carried out for model estimation are marked in blue, and tasks that were carried out for model validation, thus taking into account the nearest neighbors, are market in yellow.

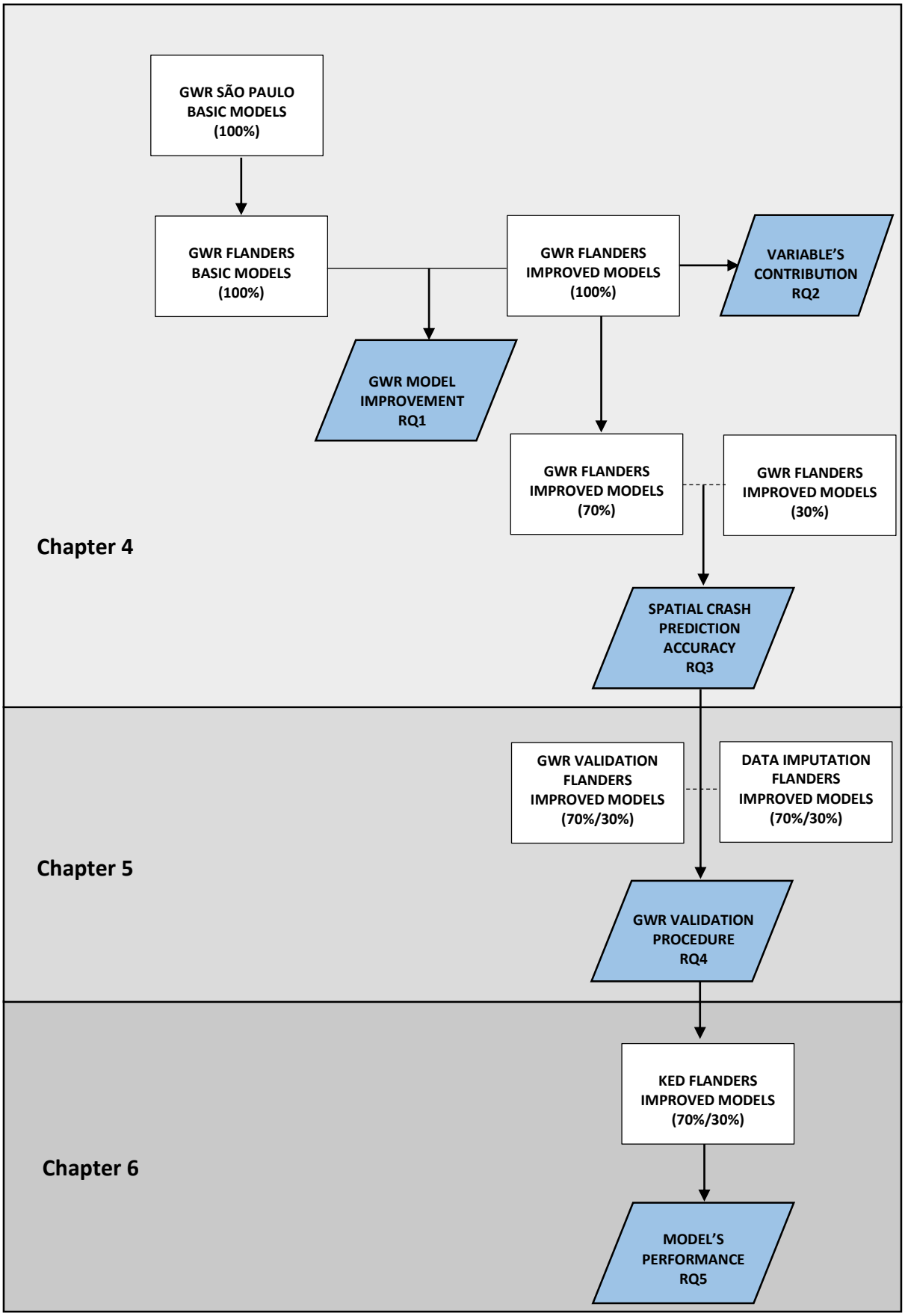
Figure 3.11 - Approaches' overview



Finally, in Chapter 6, we evaluated the performance of GWR (by means of GWR holdout 1 and GWR holdout2) in relation to Geostatistics (by means of KED), and the two missing data imputation approaches, i.e. MEI and KNN. This enabled us to answer the last research question of this thesis (RQ5).

Figure 3.12 summarizes the methodological framework and tasks concerning the investigations at analytical chapters. Detailed information concerning the methodological framework is meticulously provided in each chapter. Subsequently, in the next subsection, we describe the repeated holdout method.

Figure 3.12 - Methodology framework



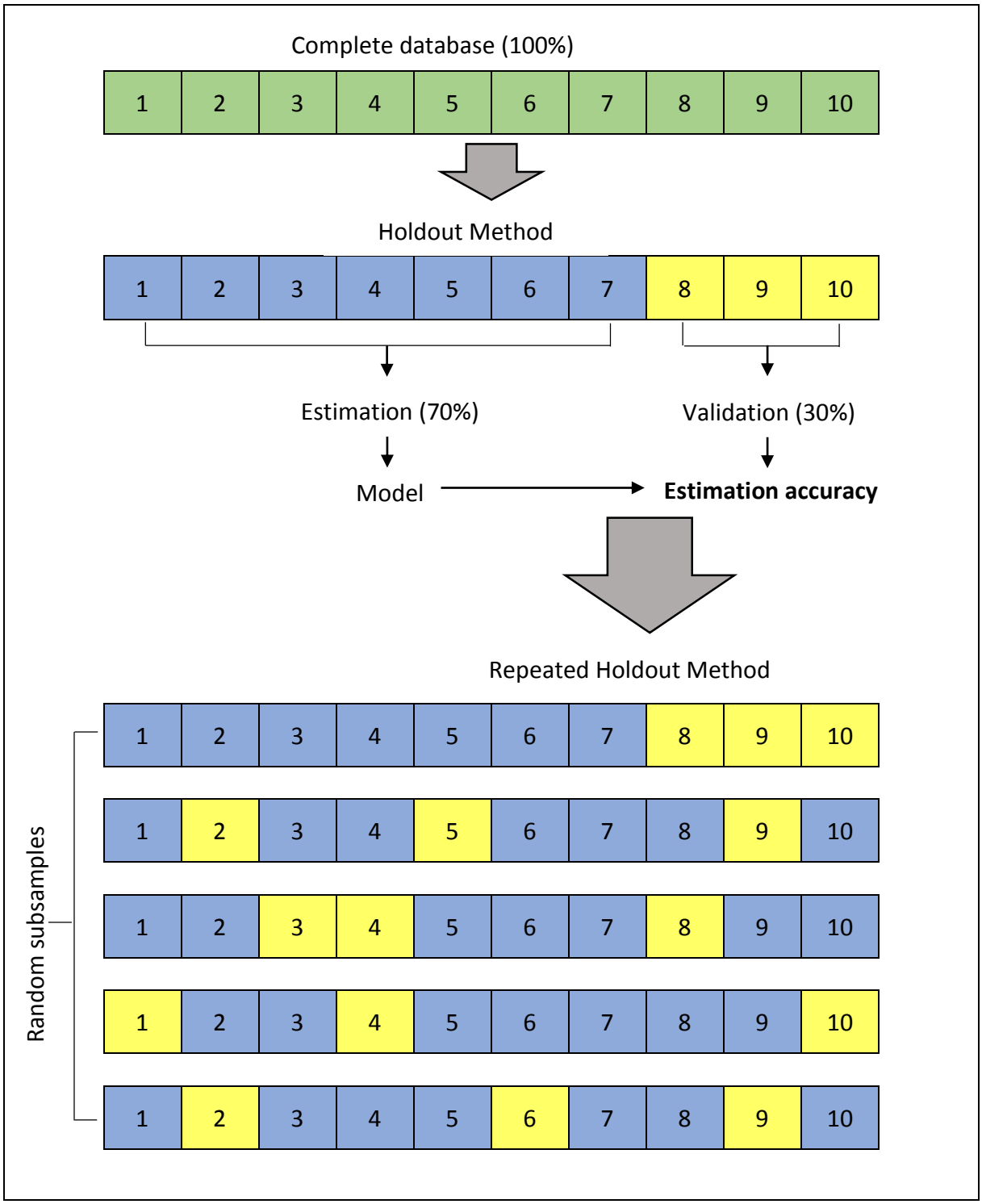
3.4.1 REPEATED HOLDOUT FRAMEWORK

In order to test the accuracy of the developed Flemish spatial models (questioned in the three core chapters in the RQ3, RQ4 and RQ5), we adjusted these models within the Holdout Method. In this respect, model accuracy was assessed by the corresponding goodness of fit obtained with a pre-determined validation sample. This means that part of the data was used for estimate the models (70%) and a different part for validating (30%). Therewith, casualty estimates were compared to the observed information, at the validation subzones, by means of Mean Squared Prediction Error (MSPE) and Pearson Correlation Coefficient (PCC).

Subsequently, aiming to improve the reliability of estimates, we repeated the holdout method five times. This practice is called Random Sub sampling or Repeated Holdout Method (by repeating the holdout method *k*-times). This means that analyses, in this study, were carried out for five subsamples with 100% of the data, however with random sets of 70/30 percent. Figure 3.13 illustrates how this process was followed.

Finally, the overall accuracy of models in the validation samples was determined by the average of the performance measures (i.e. MSPE and PCC) obtained for the five interactions with the 70/30 subsamples. This was applied for both motorized and active transport modes. Furthermore, taking into account the proposed comparison of model performance for all multivariate spatial data analysis approaches implemented in this research, we used the same five subsamples with 70/30 percent to construct the empirical casualty prediction models in Chapters 4, 5 and 6, therefore, addressing RQ3, RQ4 and RQ5.

Figure 3.13 - Segregation of the database within the Repeated Holdout Method



4 STUDY OF THE IMPACTS OF ENRICHED INFORMATION ON SPATIAL MODEL PERFORMANCE

Analyses in this chapter attempt to answer the following research questions:

RQ1. *Based on a benchmarking exercise, what potential improvements in spatial model performance can be obtained by including additional explanatory variables?*

RQ2. *What is the statistical individual contribution of variables to the produced models?*

RQ3. *Are these models reliable?*

Therefore, this chapter is subdivided into four main sections. In the first section, the improvements in the performance of Geographically Weighted Poisson Regression (GWPR) models by enhancing the explanatory variables are outlined. This investigation is carried out based on available information of Flanders and São Paulo, for 100% of the data. Data on casualties and fatalities are used as dependent variables in the models developed for Flanders and São Paulo, respectively. In the next section, a sensitivity analysis is carried out to identify the statistical individual contribution of the input information in the casualty prediction. Moreover, in order to verify the reliability of these models, in the third section, we frame them within the repeated holdout method, by randomly subsampling the entire database into two sets with 70% and 30% of the data, five times. Therewith, we suggest an approach to fit Geographically Weighted Regression (GWR) within the concept of model validation and model estimation, called GWR holdout1. Finally, the last section ends this chapter drawing the conclusions found at each stage.

4.1 IMPROVEMENTS IN SPATIAL MODEL PERFORMANCE: CASE STUDIES IN SÃO PAULO AND FLANDERS

In order to demonstrate the potential improvements in the performance of spatial prediction models by enhancing the potential explanatory variables, a case study was carried out based on available fatal and injury crash-related information from São Paulo (Brazil) and Flanders (Belgium). Given the attributes of the variables collected for both regions of study,

information concerning the number of fatalities per municipality was used as the dependent variable for models developed for São Paulo, and casualties per subzone, as the dependent variable for models developed for Flanders. For both regions, models were produced for active and motorized transport mode users, and using 100% of the data. Hence, the methodological procedure was divided into two main stages.

In the first stage, GWPR models were developed for São Paulo and Flanders, by only taking into account the same explanatory variables available in both datasets. Given the limitations of the Brazilian dataset, the results of this stage would reveal the best we could do with the available information of São Paulo, while there would be plenty of room yet to improve the Flemish models. In the second stage, in order to highlight the importance of having data which is as complete as possible, GWPR models were developed for Flanders only, by considering all available variables in the Flemish dataset.

At both stages, a multicollinearity test was conducted prior to the spatial modelling steps, enabling us to assess the suitability that variables would have in the models and select the most significant ones to compose the final models for each dependent variable. As common practice, the Variance Inflation Factor (VIF) was used to quantify how much the variance of the estimated regression coefficients increased if predictors were correlated. As a common rule of thumb, 10 was defined as a cut off value, meaning that if VIF was higher than 10, then multicollinearity was high (Kutner, Nachtsheim, & Neter, 2004) and, therefore, variables with high VIF measures should not be present in the model simultaneously. After excluding variables with high VIF, the remaining ones were used to produce the GWPR models. At the end of this stage, models, for which we used the minimum data, were developed and called basic models.

In the second stage, the affluence of available information in the Flemish dataset enabled us to produce distinct GWPR models with different combination of explanatory variables, thus selecting the ones with the best overall fit for each dependent variable. This exercise was carried out by having the intercept term as our starting point and analytically combining variables with a VIF lower than 10. Hereafter, due to the greater complexities of the GWR

estimation procedure that conceivably causes interrelationships among local coefficient estimates when there is no collinearity among the explanatory variables (Gue et al., 2008; Hadayeghi et al., 2010; Pirdavani, Bellemans, Brijs, & Wets, 2014), at this stage, evidence of multicollinearity among the produced local coefficient estimates was also verified. Thus, among all the Crash Prediction Models (CPM) developed, the most adequate for each dependent variable, was selected as it met the criteria of non-multicollinearity among variables and local coefficient estimates, and subsequently based on the lowest corrected Akaike Information Criterion (AICc) value. As a common rule-of-thumb, the difference between the models was considered significant when the difference of AICc values between two models was higher than 4 (Charlton & Fotheringham, 2009). At the end of this stage, models, for which we used the most significant information in the Flemish dataset, were developed and called improved models.

Finally, the performance of the improved models was compared with the basic models by means of AICc, Mean Squared Prediction Error (MSPE) and the Pearson Correlation Coefficient (PCC). Results of both stages, are presented in the next subsections, for both active and motorized transport.

4.1.1 GEOGRAPHICALLY WEIGHTED REGRESSION BASIC MODELS – SÃO PAULO

In spite of a great amount of available information in Brazil, most of it was limited to socio-economic and demographic variables. As result, most of the pieces of information collected (e.g., area, number of inhabitants, population by gender, urban and rural population, fuel consumption, vehicle fleet, employed population, occupied population, Gross National Product, Human Development Index, etc.), were found to be correlated with each other, therefore presenting high VIF values. Hence, produced basic models were limited to information concerning the link length and population only. VIF values among the variables of the basic models are shown in Table 4.1. At this stage, the population was used as the exposure variable in its Natural Logarithm (ln) form, for active and motorized transport. The ln was taken so that in case of having zero exposure no crash would be expected. In this

respect, the choice of using population as the exposure variable was given by its model outperformance compared to link length, and based on the assumption that this information is a better and a more meaningful proxy of exposure.

Table 4.1 - VIF values among variables – Basic Models – São Paulo

Parameters	Motorized Transport	Active Transport
Ln Population	2.241	2.217
Link length	2.241	2.217

Table 4.2 shows the local parameter estimates for both dependent variables. This information is described by five number summaries: minimum, maximum (lower, median and upper quartile), tabulated in this format and sequence. Moreover, the information concerning model performance is also presented.

Table 4.2- Local parameter estimates and model performance - Basic Models – São Paulo

Parameters	Active Transport	Motorized Transport
Intercept	-24.094, 15.716 (-13.546, -10.626, -8.081)	-16.228, 0.031 (-8.576, -6.837, -4.896)
Ln Population	-2.249, 2.463 (0.874, 1.155, 1.438)	0.019, 1.898 (0.672, 0.871, 1.053)
Link length	-0.006, 0.038 (-7.8e-04, 1.5e-04, 1.2e-03)	-0.009, 0.010 (2.2e-04, 4.4e-04, 1.4e-03)
GWPR AICc	1754.804	3095.810
Global AICc	2178.477	4940.469
MSPE	50.49	94.99
PCC	0.953	0.941

Despite the impressive results of model performance for the Brazilian basic models, this is not particularly surprising in light of the “large” aggregation level of the data (i.e. municipalities) and differences in the areal units of analysis. These effects of the scale (large aggregation) are

related to Modifiable Areal Unit Problem (MAUP) and can lead to inaccurate estimation (Openshaw & Taylor, 1979, 1981; Openshaw, 1984; Cressie, 1996).

Figure 4.1 shows maps of the local coefficient estimates as well as their significance at 0.05 level, for motorized and active transport in São Paulo. In order to determine where relationships were significant (in blue) and where they were not (in brown), we computed the t-statistics.

Figure 4.1 - Maps for motorized transport - Basic model – São Paulo

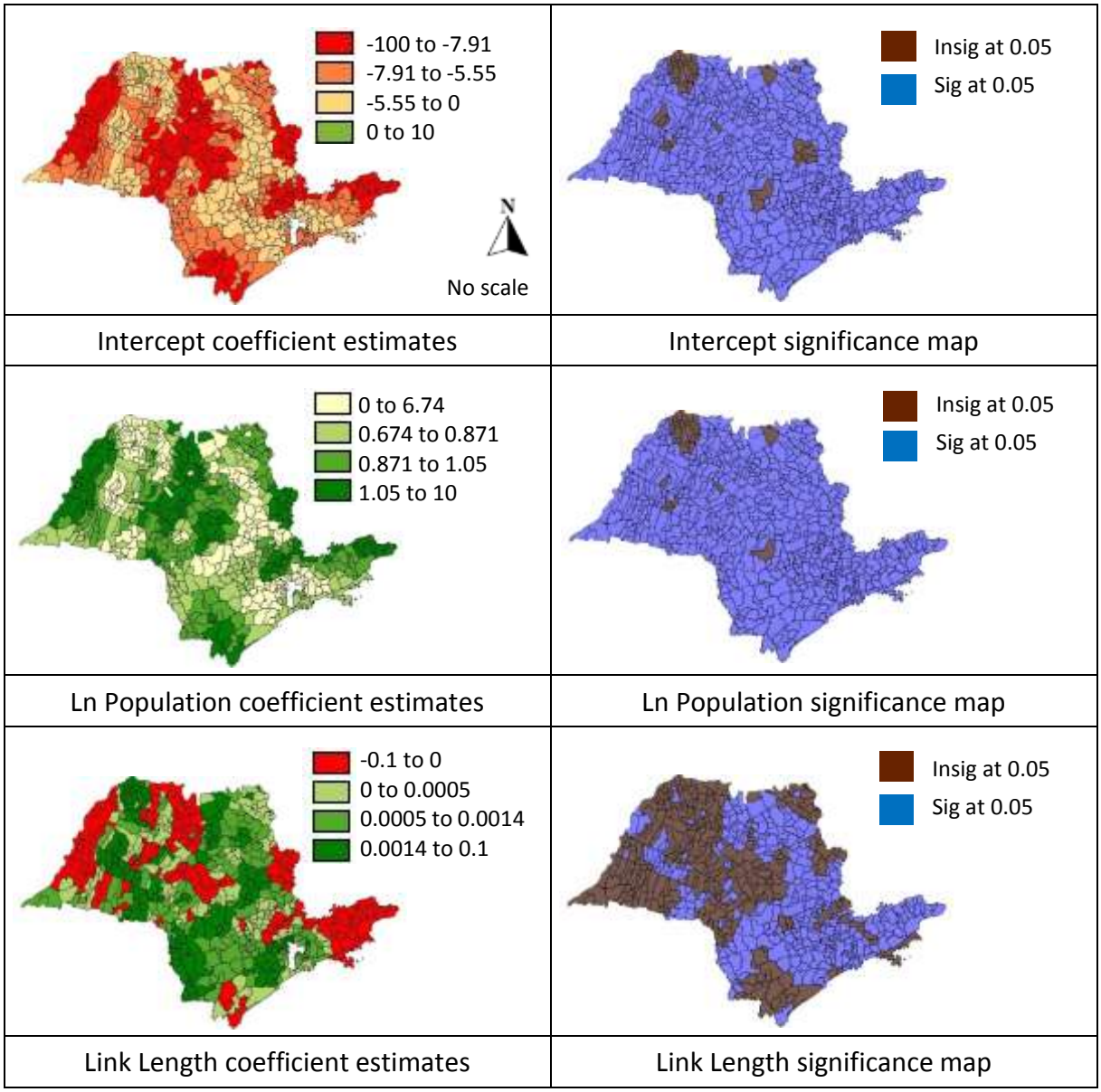


Figure 4.2 shows maps with the observed and predicted number of fatalities, for motorized transport, and a map of errors with the local differences found between these values. Subsequently, Figures 4.3 and 4.4 show the respective maps for active transport.

Figure 4.2 - Observed and predicted number of fatalities
(Motorized transport - Basic model – São Paulo)

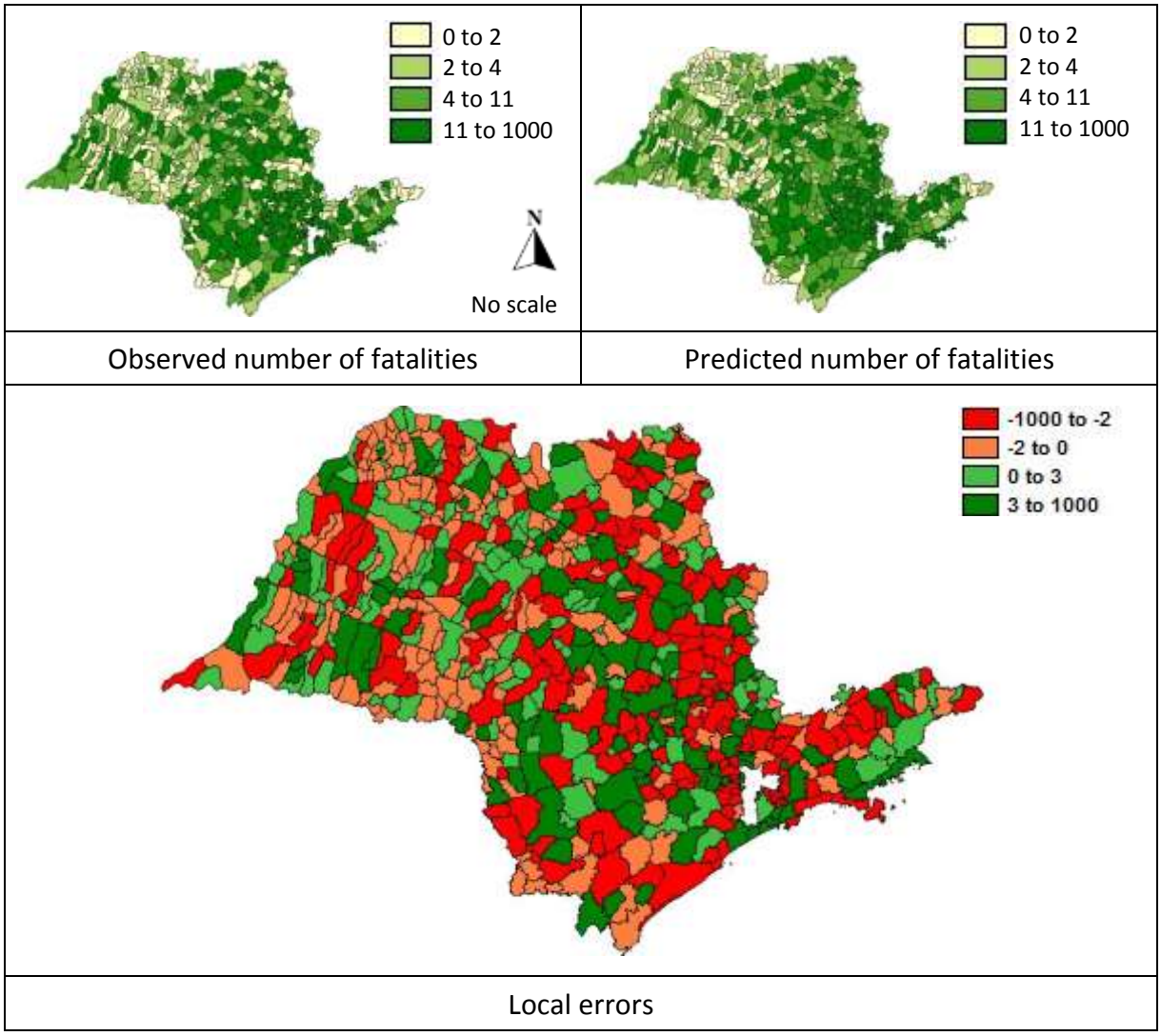


Figure 4.3 - Maps for active transport - Basic model – São Paulo

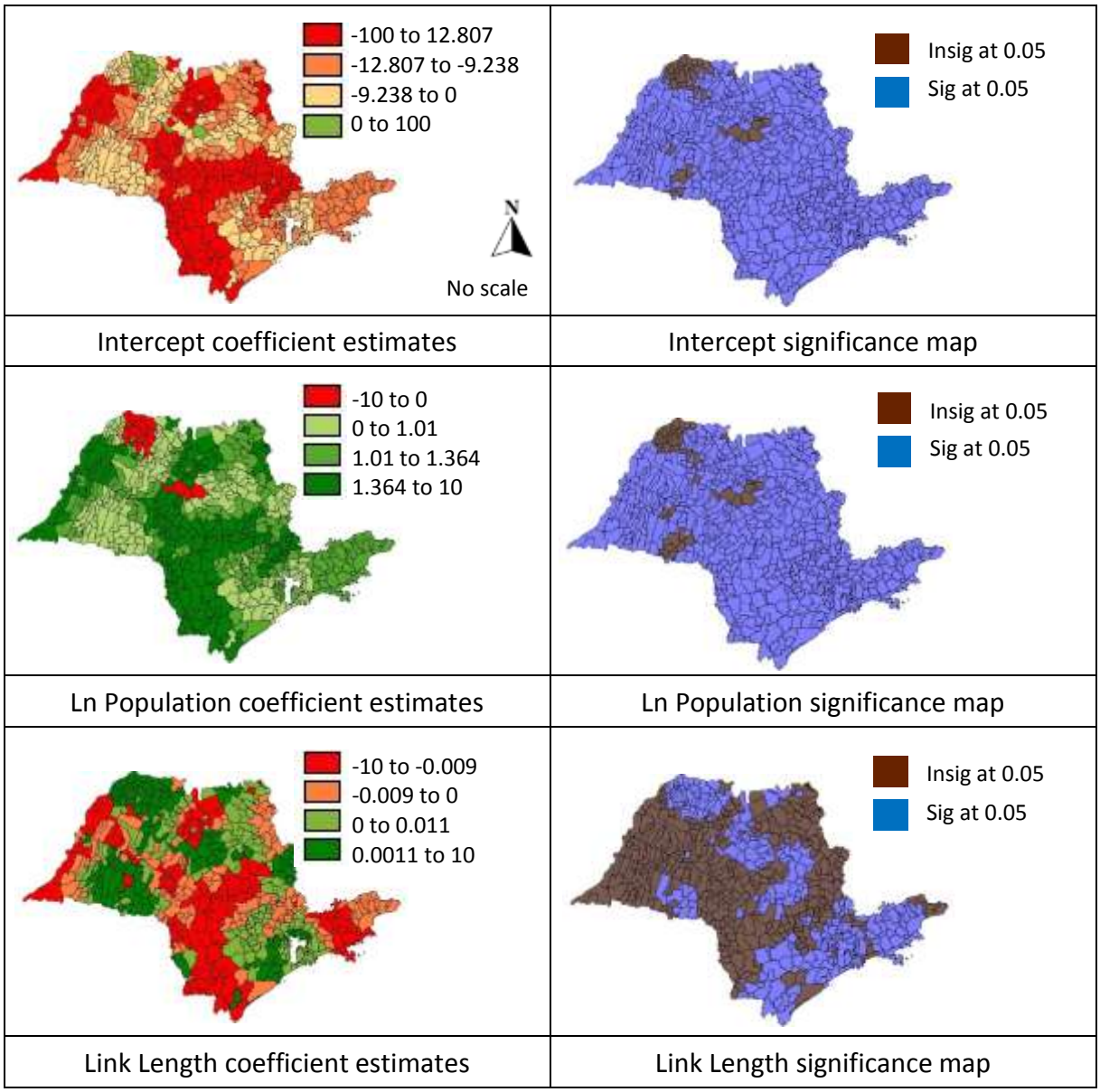
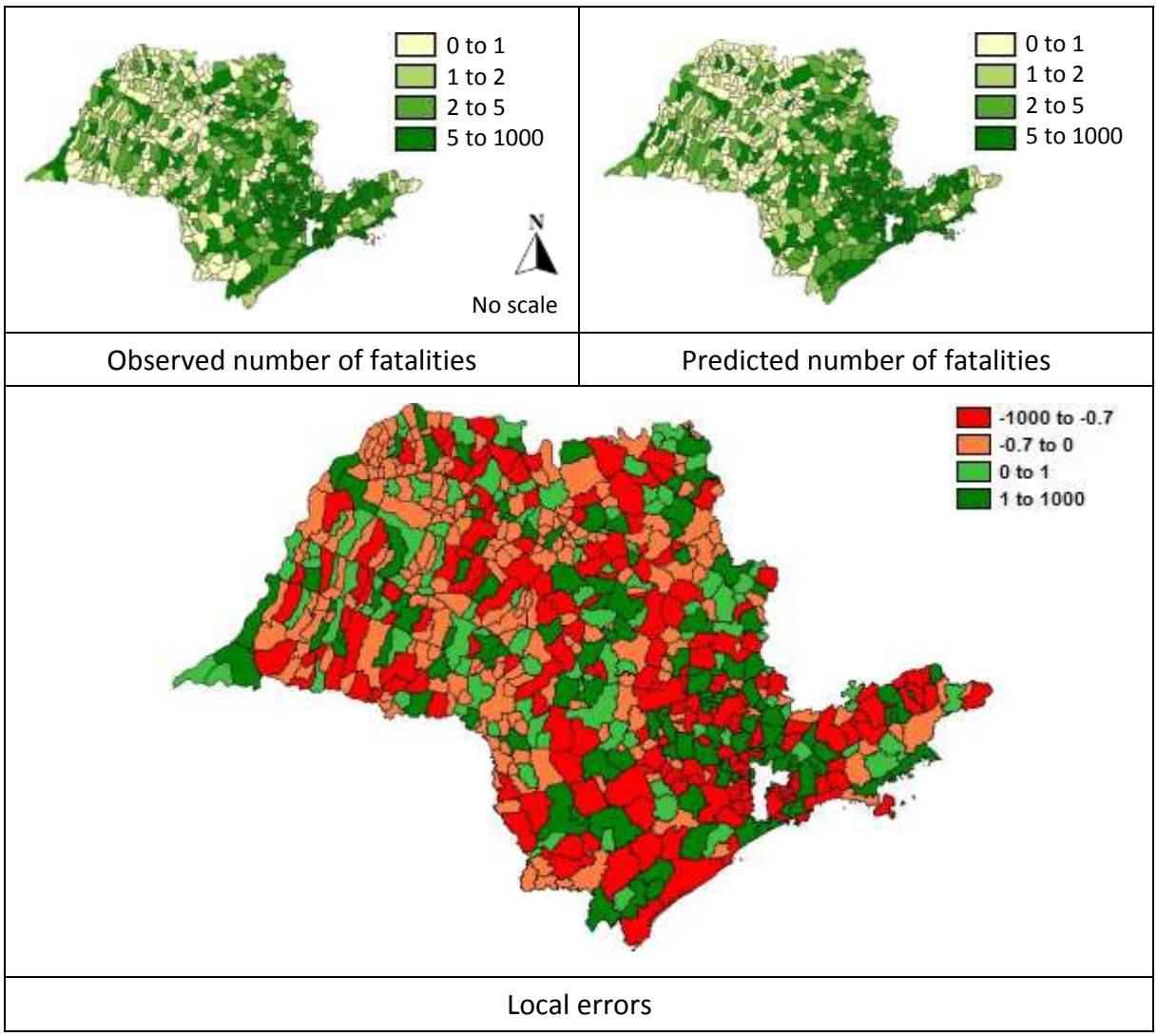


Figure 4.4 - Observed and predicted number of fatalities
(Active transport - Basic model – São Paulo)



Observed number of fatalities | Predicted number of fatalities

Local errors

4.1.2 GEOGRAPHICALLY WEIGHTED REGRESSION BASIC MODELS – FLANDERS

CPM for Flanders, at this stage, followed the same criteria and modelling composition of the basic models produced for São Paulo. Table 4.3 presents the VIF values among the variables for the final models. Subsequently Table 4.4 presents the resulting local parameters in addition to the information concerning model performance.

Table 4.3 - VIF values among variables – Basic Models – Flanders

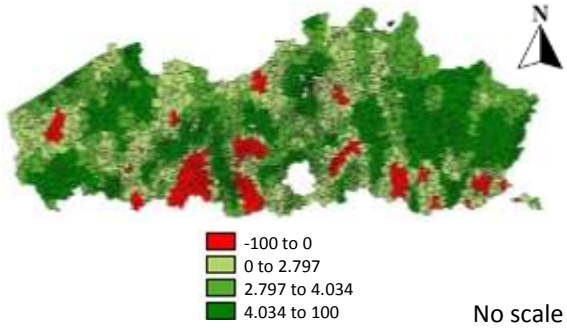
Parameters	Motorized Transport	Active Transport
Ln Population	1.262	1.224
Link length	1.262	1.224

Table 4.4 - Local parameter estimates and model performance - Basic Models - Flanders

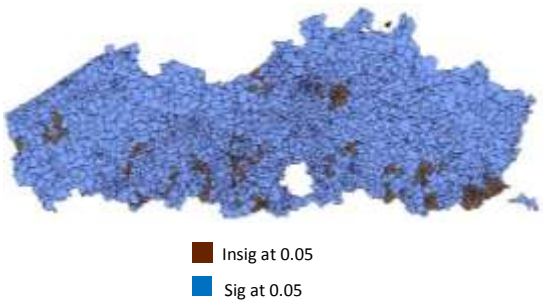
Parameters	Active Transport	Motorized Transport
Intercept	-8.890, 11.749 (0.153, 1.813, 2.966)	-5.981, 8.582 (1.894, 3.336, 4.246)
Ln Population	-1.646, 1.668 (-0.046, 0.130, 0.365)	-1.252, 1.307 (-0.063, 0.077, 0.267)
Link length	-0.264, 0.130 (-0.038, -0.012, 0.010)	-0.149, 0.139 (-0.031, 0-0.008, 0.009)
GWPR AICc	29345.571	60993.481
Global AICc	50570.229	102673.889
MSPE	384.87	1786.66
PCC	0.629	0.626

Figure 4.5 presents the local coefficient estimates, as well as their significance maps at 0.05 level. Subsequently, Figure 4.6 shows maps of observed and predicted number of casualties, and the local errors obtained for motorized transport, in Flanders. The corresponding maps for active transport are presented in Figures 4.7 and 4.8.

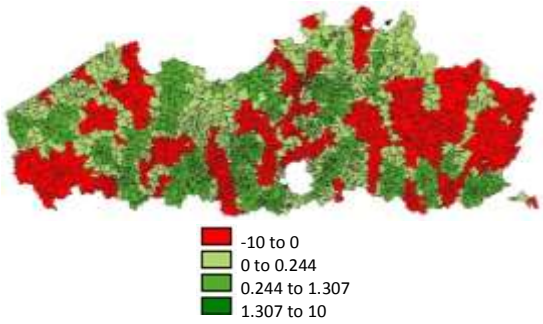
Figure 4.5 - Maps for motorized transport - Basic model – Flanders



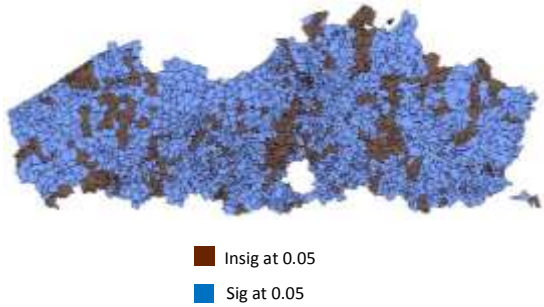
Intercept coefficient estimates



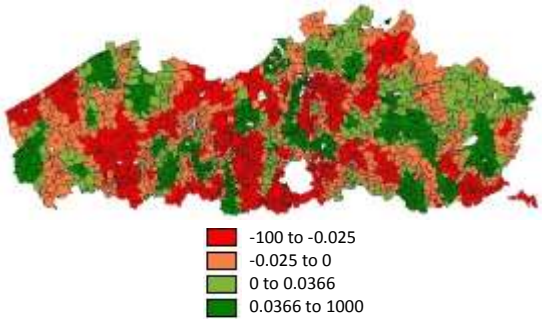
Intercept significance map



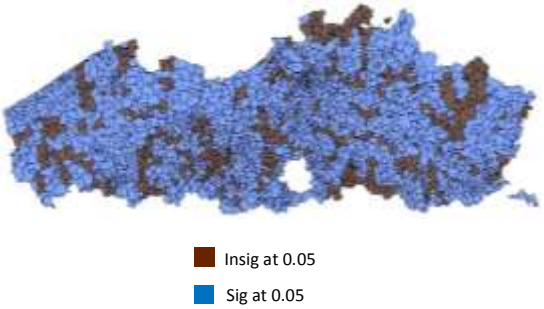
Ln Population coefficient estimates



Ln Population significance map

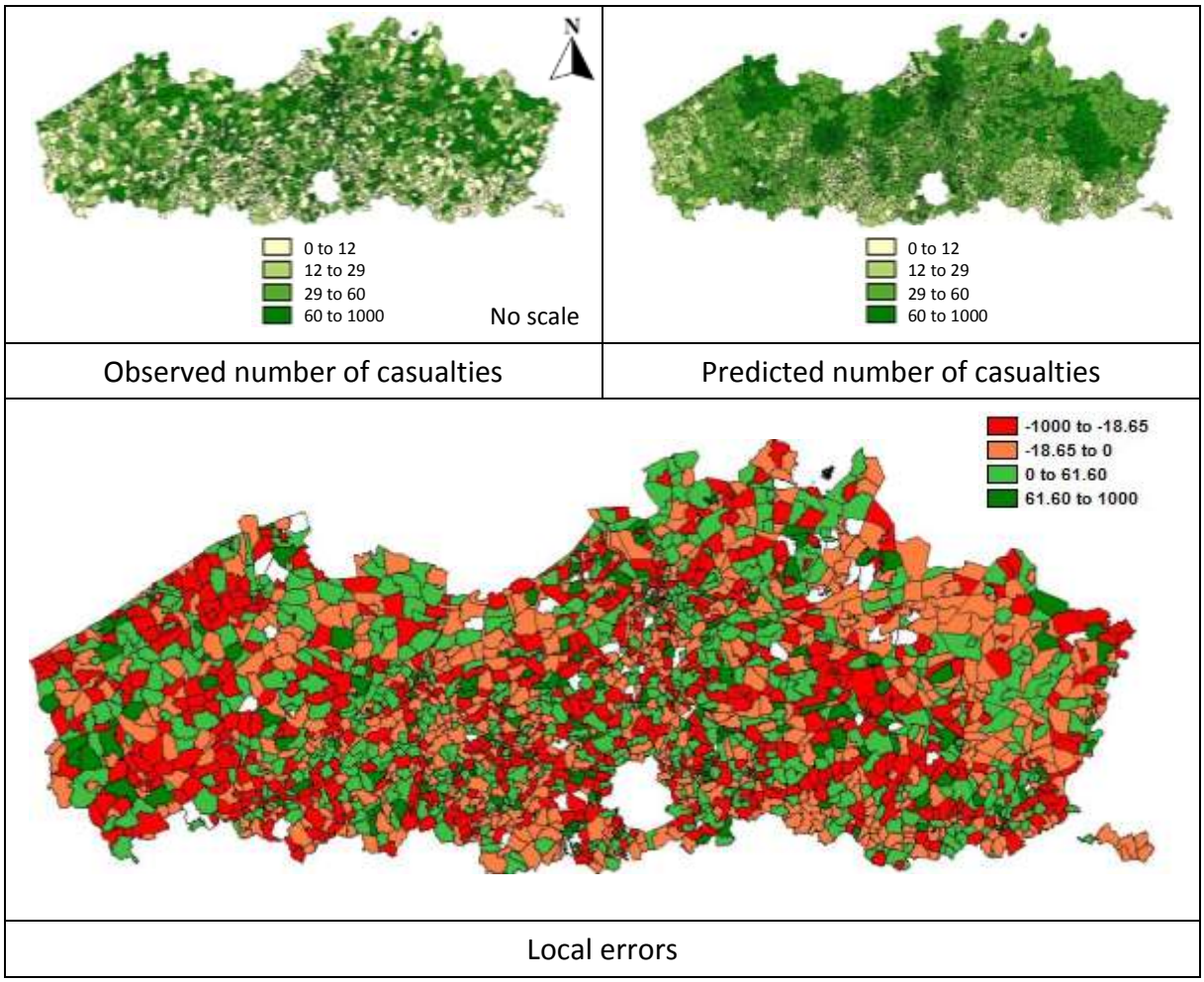


Link Length coefficient estimates



Link Length significance map

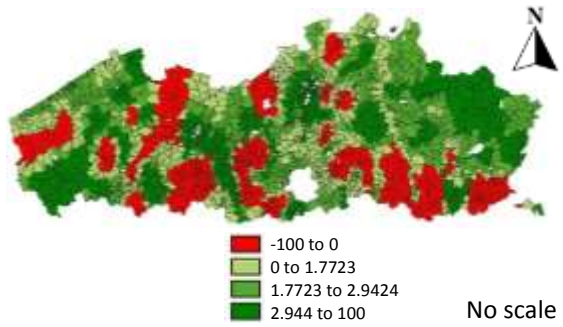
Figure 4.6 - Observed and predicted number of casualties
(Motorized transport - Basic model – Flanders)



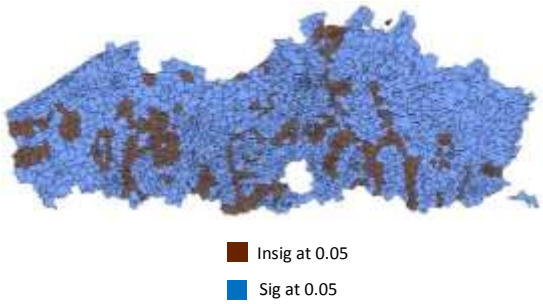
Observed number of casualties	Predicted number of casualties

Local errors

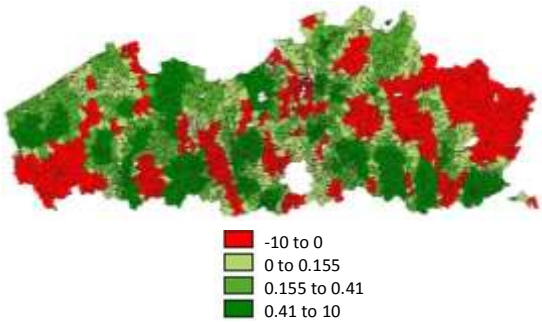
Figure 4.7 - Maps for active transport - Basic model – Flanders



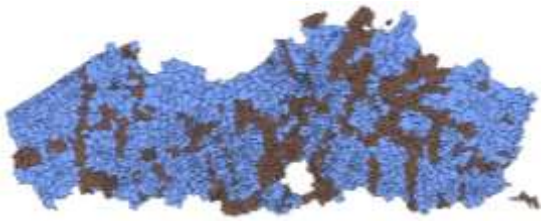
Intercept coefficient estimates



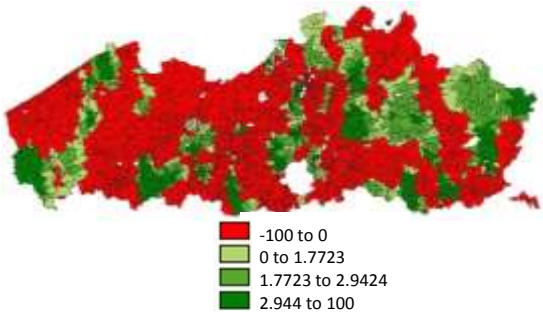
Intercept significance map



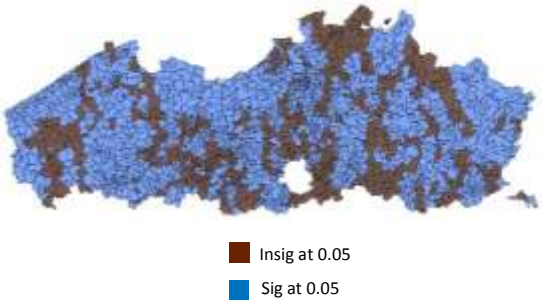
Ln Population coefficient estimates



Ln Population significance map

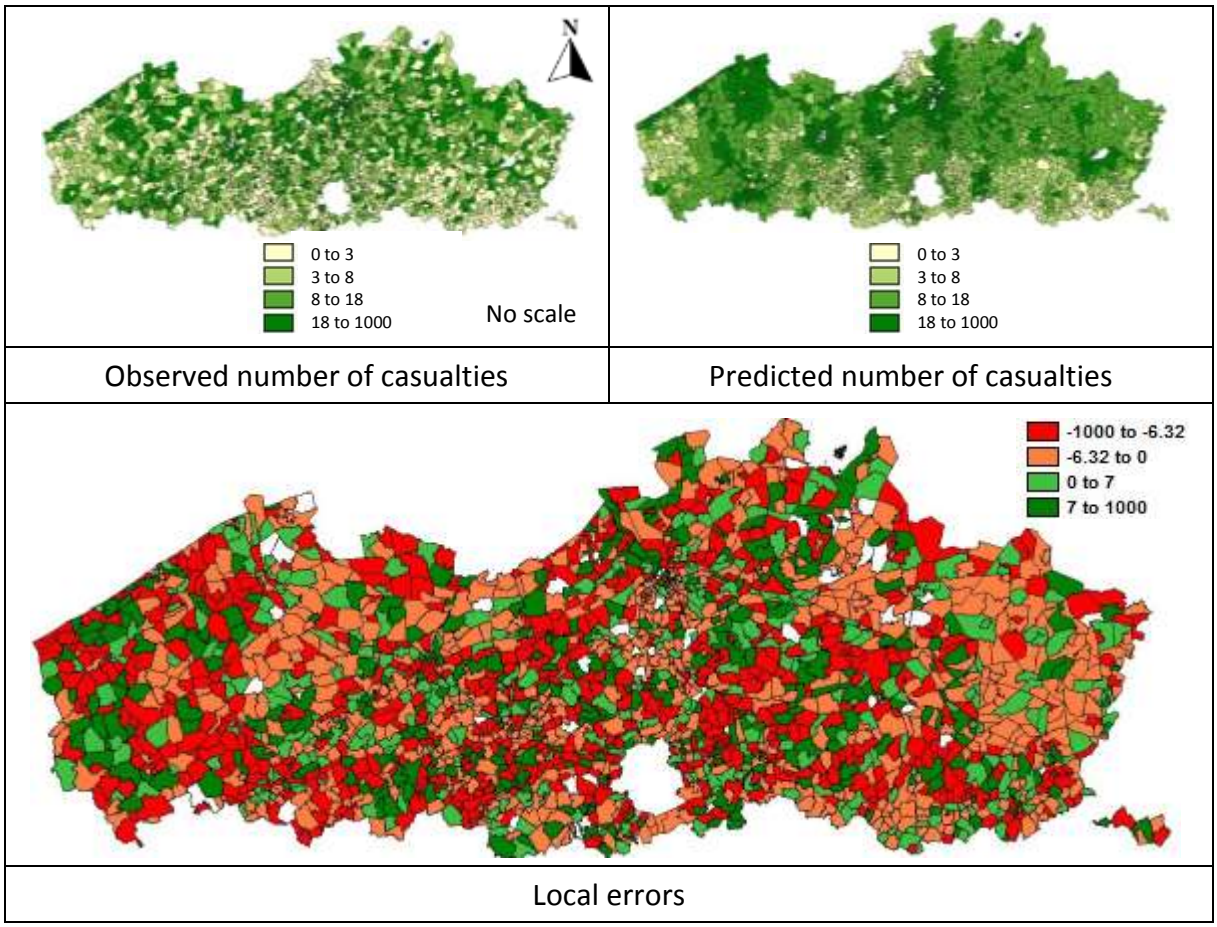


Link Length coefficient estimates



Link Length significance map

Figure 4.8 - Observed and predicted number of casualties
(Active transport - Basic model – Flanders)



Observed number of casualties | Predicted number of casualties

Local errors

4.1.3 GEOGRAPHICALLY WEIGHTED REGRESSION IMPROVED MODELS – FLANDERS

Models at this stage were developed for Flanders only and by considering all available information in the Flemish dataset. Contrary to the Brazilian case, the Flemish dataset, in addition to significant information related to socioeconomic, sociodemographic and road network (i.e. income level, speed, capacity, number of links, links and intersection density, presence of highways, urbanization degree, to name a few) provided foremost diverse and suitable exposure variables, i.e. Number of Trips (NOTs), vehicles flow and Vehicle Kilometers traveled (VKT). This enabled us to produce several distinct models and choose the best fitted one.

After carrying out the VIF test among variables and produced coefficient estimates, the final improved models for active and motorized transport modes comprised the following information: NOTs, number of children attending school (school children), road capacity (capacity), intersection density, car ownership, and VKT. At this stage, we used VKT as the exposure variable in its logarithm form. VIF values among the variables and local coefficient estimates for the final set of variables are shown in Tables 4.5 and 4.6. Subsequently, Table 4.7 presents the respective local coefficient estimates found for each explanatory variable in the five number summary format, together with the information concerning model performance.

Table 4.5 - VIF values – motorized transport – Flanders

Parameters	Among variables	Among local coefficient estimates
NOTs	2.945	3.077
School children	2.226	3.276
Capacity	1.613	1.370
Intersection Density	1.305	1.542
Car ownership	1.405	1.076
VKT	1.997	1.651

Table 4.6 -VIF values – active transport – Flanders

Parameters	Among variables	Among Coefficient Estimates
NOTs	3.003	3.098
School children	2.274	2.914
Capacity	1.047	1.131
Intersection Density	1.389	1.306
Car ownership	1.050	1.035
VKT	1.280	1.232

Table 4.7 - Local parameter estimates and model performance - Improved Models - Flanders

Parameters	Active Transport	Motorized Transport
Intercept	-14.835, 13.933 (-0.928, 1.580, 4.068)	-7.604, 11.041 (1.585, 3.312, 4.973)
NOTs	-0.002, 0.005 (-2.78-04, 4.8e-05, 4.4-04)	-9.4e-04, 0.001 (-1.04e-04, 2e-05, 1.23e-04)
School children	-0.007, 0.003 (-9.16-04, -2.33e-04, 3.4e-04)	-0.004, 0.003 (-5.74e-04, -1.43e-04, 2.33e-04)
Capacity	-0.008, 0.007 (-3.46e-04, 7.3e-05, 5.47e-04)	-0.007, 0.005 (-3.89-04, 1.5e-05, 4.5-04)
Intersection density	-4.183, 1.686 (-0.139, 0.023, 0.152)	-1.441, 1.171 (-0.086, 0.025, 0.131)
Ln VKT	-0.653, 1.486 (-0.054, 0.128, 0.306)	-0.666, 0.888 (-0.092, 0.076, 0.246)
Car Ownership	-7.378, 6.584 (-1.649, -0.375, 0.445)	-6.763, 4.722 (-1.084, -0.128, 0.456)
GWPR AICc	22553.607	49525.611
Global AICc	45486.734	94634.071
MSPE	261.17	1354.12
PCC	0.771	0.738

Figure 4.9 presents the respective local coefficient estimates and significance maps, for motorized transport.

Figure 4.9 - Maps for motorized transport - Improved model – Flanders

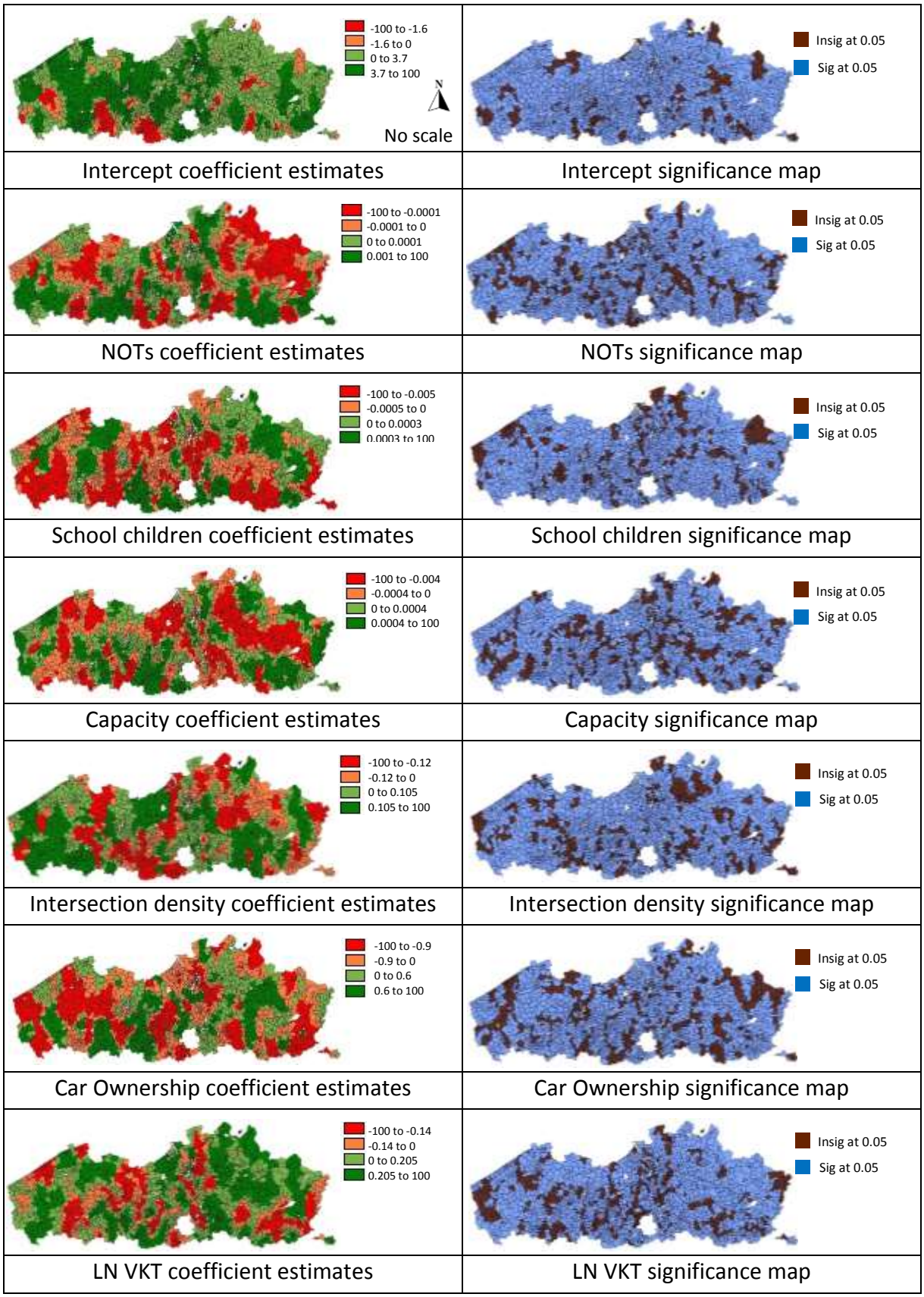


Figure 4.10 shows the produced maps of observed and predicted number of casualties, and the corresponding local errors. Figures 4.11 and 4.12 present the corresponding maps for active transport.

Figure 4.10 - Observed and predicted number of casualties
(Motorized transport - Improved model – Flanders)

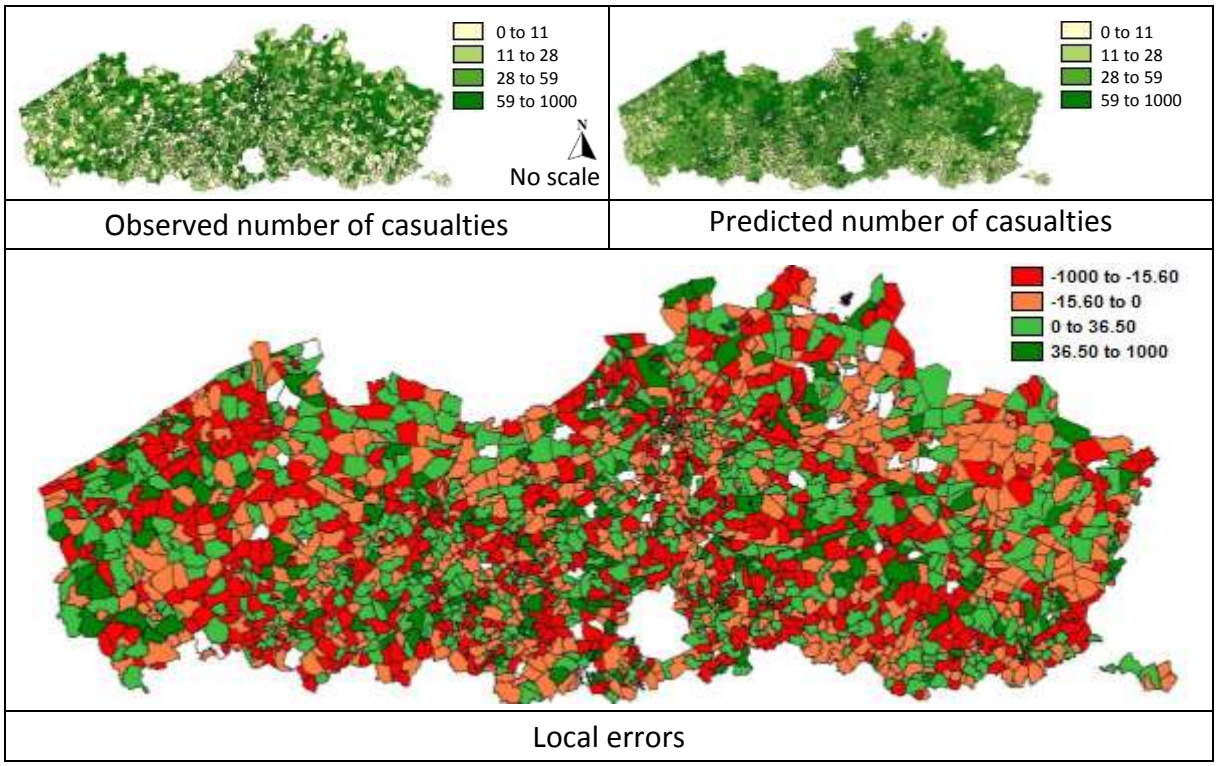


Figure 4.11 - Local coefficient estimates and significance maps
(Active transport - Improved model – Flanders)

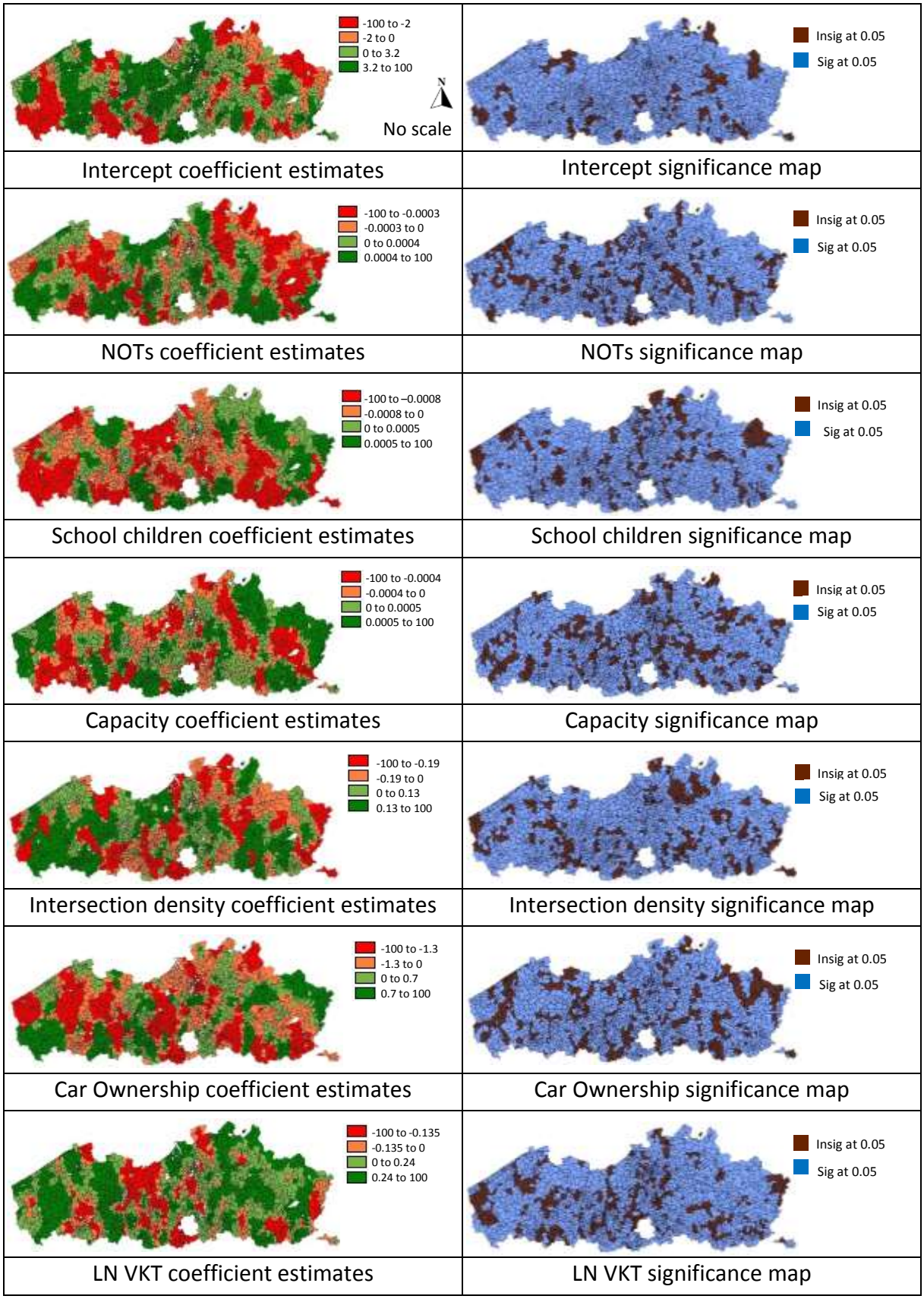
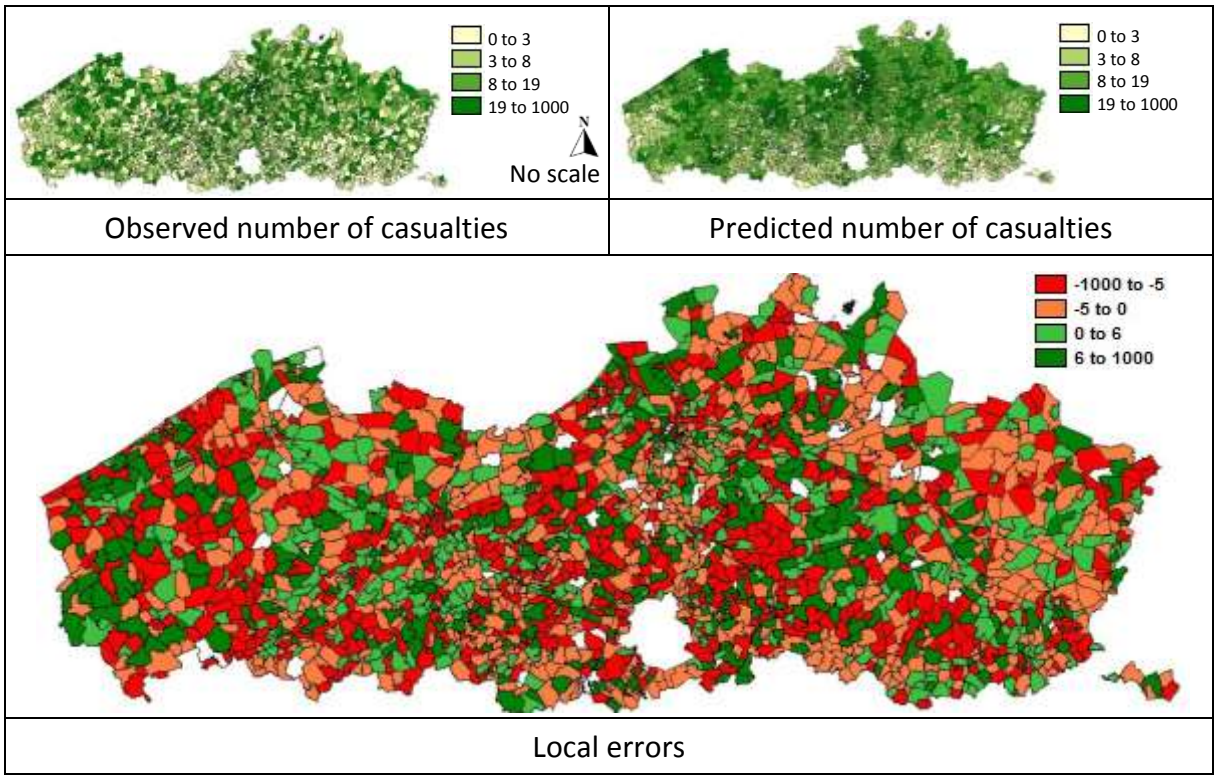


Figure 4.12 - Observed and predicted number of casualties
(Active transport - Improved model – Flanders)



It can been seen that there were no real geographical clusters in positive or negative values of errors, for models developed for both active and motorize transport. This enables us to assume that there were no regions where the models led to over or under predictions.

4.1.4 REMARKS: BASIC MODELS VS. IMPROVED MODELS

In general, results of model performance for the basic models, revealed a negative correlation between casualties and both explanatory variables in a large number of TAZs, which in turn were significant at the 0.05 confidence level. This was observed for both active and motorized transport results. Given the rather direct association of population and link length with exposure, we would expect that both explanatory variables were likely to have a positive correlation with casualties, in most TAZs of this study, as also found in Wang, Quddus and Ison (2009) and Pirdavani, Brijs, Bellemans and Wets (2013), for example.

One explanation for these counterintuitive signs could be that some variables, at some locations, are locally correlated while no global multicollinearity is observed among them (Wheeler & Calder, 2007; Guo et al., 2008; Hadayeghi et al., 2010; Pirdavani et al., 2014). Such local correlation has been attributed as the major reason for problems with counterintuitive signs. In order to address this issue, at the second stage, we excluded variables at which high VIF values were found among the produced coefficient estimates. The limitation of variables within the basic models, restrained us from conducting this exercise at the first stage, therefore corroborating with the importance of a more diverse set of explanatory variables.

Such counterintuitive signs could also be a result of the omission of important variables, which leads to omitted variable bias. Although not thoroughly explored in this study, one could assume the correlation of link length and other road features or exposure variables that were omitted, therefore producing bias. The exclusion of such essential variables, specially an exposure variable, could systematically invalidate further conclusions that could be derived from the results (Washington et al., 2010; Mitra & Washington, 2012). A more in-depth investigation concerning this problem could help to provide a better insight of the direction of these effects, which in this study, remains speculative.

From a statistical point of view, the improved Flemish models (not surprisingly) outperformed the basic models for both dependent variables. In the model developed for motorized transport, reductions of proximately 20% and 25% were observed compared to the basic model, for AICc and MSPE respectively (Table 4.8). Likewise for the active mode, 25% and 35% reductions were obtained for AICc and MSPE, respectively (Table 4.9).

Table 4.8 - Model parameters for motorized transport in Flanders

	Basic Model	**Improved Model**
GWPR AICc	60993.481	49525.611
Global AICc	102673.889	94634.071
MSPE	1786.66	1354.12
PCC	0.626	0.738

Table 4.9 - Model parameters for active transport in Flanders

	Basic Model	Improved Model
GWPR AICc	29345.571	22553.607
Global AICc	50570.229	45486.734
MSPE	384.87	261.17
PCC	0.629	0.771

Moreover, results of a more diverse dataset, especially including an exposure variable enabled the development of coherent coefficient estimates and signs. Results of this study revealed positive associations with casualties, in most TAZs, for intersection density, capacity, NOTs and VKT. These results are consistent with findings in previous studies (such as Hadayeghi et al., 2003; Aguero-Valverde & Jovanis, 2006; Pirdavani et al., 2013a, 2013b; Shariat-Mohaymany et al., 2015; Xu et al., 2017).

Casualties were found to have a negative correlation with the number of children attending school and car ownership, in a large number of TAZs. This negative association with other social standing variables, e.g., income level, has been found in other previous studies (such as in Li, Wang, Liu, Bigham, & Ragland, 2013; Pirdavani et al., 2013a, 2013b; Pirdavani et al., 2014; Pirdavani et al., 2016). One could assume that, less casualties are expected to occur in more affluent areas (in this study, where car ownership is higher). Yet, particularly concerning the negative association of casualties and car ownership, it could be that, land-use might have an influence on the speed limits, and therefore on the event. Especially in Flanders, car ownership is generally larger in suburban and rural zones and lower in urban zones and city centers, where more alternative means of transportation are available, for example. In suburban and rural zones, speed limits on the road are typically higher compared to those allowed in the urban areas and city centers. Hence, such negative association, might be explained by the fact that car ownership is a proxy variable of the average speed limit in the TAZ. Given the negative association of casualties and children attending school, we would assume that school children might be a proxy variable of the presence of schools in the TAZ, as well as speed limit and human factors associated to the driver's behavior, might have an influence on the event. Since

the introduction of lower speed limits, i.e. 30km/h, near schools, speed limits in such TAZs are lower than in those without a school.

In order to corroborate these assumptions, and enable us to better understand the interactions between variables, such a macro-level analysis could be used as a basis for local investigations. This could help to enforce appropriate countermeasures, especially in areas at which higher estimates were found. For instance, at subzones where casualties were found to have a positive association with school children, micro-level analysis could suggest changes in the speed limits or signalizing intersections. This could be identified as the major contribution of having a more complete and diverse dataset. In spite of more reliable models, they would allow policy makers to prioritize hot zones, and depending on the targets, specific TAZs could be used to investigate the interaction between variables, both within and outside of the models.

4.2 SENSITIVITY ANALYSIS

In spite of the large body of research, which have enabled researchers and policy makers to associate crashes to other variables (discussed in Chapter 2), they are limited in their scope to assess the influence of the predictor variables on crash occurrence. Especially for policy makers, an insight into the statistical contribution of these variables concerning the prediction models could also contribute in the decision-making process, by helping policy makers outline data collection priorities. Besides the financial and practical considerations it might have, such an investigation is justifiable as to the best of our knowledge, this practice has only been explored in terms of microscopic-level analysis, focusing on the influence of the Highway Safety Manual (HSM) data variables (AASHTO, 2010) on safety predictions. Some approached found in previous studies include the fractional factorial method (Akgüngör & Yıldız, 2007), Boosted Regression Trees (BRT) (Saha, Alluri, & Gan, 2015) and the “change one-factor-at-a-time” approach, which is the most commonly used sensitivity method in the literature (Alluri & Ogle, 2012; Findley, Zegeer, Sundstrom, Hummer, & Rasdorf, 2012; Jalayer & Zhou, 2013, Williamson, Jalayer, Zhou, & Pour-Rouholamin, 2015).

In the present study, we conducted such investigation by a step-wise approach, and accounting for variables used in the improved Flemish models. To this end, we analytically added each variable to the prediction models, altering the other ones and evaluating their statistical contribution of each variable one at a time and their interactions, thus accounting for simultaneous variations of the input variables. To this end, the intercept term was used as a starting point. Hence, explanatory variables were analytically added to the prediction models and ranked according to their contribution in the model's performance. This contribution was measured by means of the AICc variations (%), where the larger reduction in AICc by the inclusion of a variable, the greater its contribution to the model performance. Subsequently, this process was repeated with the remaining variables, but taking into account their interactions. Thereupon, variables were tabulated according to their relative percentage of influence on the models in relation to the intercept term, namely Relative I, and in relation to its previously best fitted model composition, namely Relative II. Tables 4.10 and 4.11 show these improvements in model performance by means of the percentage reductions found on AICc, for motorized and active transport, respectively.

The results show that road capacity has the highest statistical contribution in the performance of CPM, for both active and motorized transport modes, suggesting that this information has priority over others. Secondly, VKT statistically contributes more to motorized transport models, while car ownership contributes more to active transport models, and so on. This practice could be useful as it would help policy makers prioritize data collection, for instance by targeting variables that add higher statistical contributions to one specific travel mode or both, thus reducing costs with data collection. This statement is supported in this study, for instance by the fact that road capacity has shown to bring major statistical contributions within the models, especially in relation to NOTs, which often have priority in data collection. However, the results of this investigation revealed that this information would not bring such significant contributions to the models, neither for active, nor for motorized transport.

Table 4.10 - Sensitivity analysis for motorized transport casualties

Variables	AICc	Relative I	Relative II
Intercept	69996.705	-	-
Intercept + **Capacity**	65149.59	**-6.92**	**-6.92**
Intercept + LN VKT	65345.234	- 6.65	
Intercept + Inters. density	66067.506	-5.61	
Intercept + Car ownership	66100.094	-5.57	
Intercept + School children	66106.673	-5.56	
Intercept + NOTs	67022.431	-4.25	
Intercept + Capacity + **LN VKT**	61120.219	**-12.68**	**-6.18**
Intercept + Capacity + School children	61573.189	-12.03	
Intercept + Capacity + Inters. density	61696.957	-11.86	
Intercept + Capacity + Car ownership	61675.121	-11.89	
Intercept + Capacity + NOTs	62728.751	-10.38	
Intercept + Capacity + LN VKT + **School children**	57616.189	**-17.69**	**-5.73**
Intercept + Capacity + LN VKT + Inters. density	58050.529	-17.07	
Intercept + Capacity + LN VKT + Car ownership	58261.013	-16.77	
Intercept + Capacity + LN VKT + NOTs	58963.854	-15.76	
Intercept + Capacity + LN VKT + Schoolchildren + **Inters. density**	53909.947	**-22.98**	**-6.43**
Intercept + Capacity + LN VKT + School children + Car ownership	54620.406	-21.97	
Intercept + Capacity + LN VKT + School children + NOTs	54711.502	-21.84	
Intercept + Capacity + LN VKT + School children + Inters. Density + **Car ownership**	51521.459	**-26.39**	**-4.43**
Intercept + Capacity + LN VKT + Schoolchildren + Inters. Density + NOTs	51676.283	-26.17	
Intercept + Capacity + LN VKT + Schoolchildren + Inters. Density + Car ownership + **NOTs**	49525.611	**-29.25**	**-3.87**

Table 4.11 - Sensitivity analysis for active transport casualties

Variables	AICc	Relative I	Relative II
Intercept	33593.505	-	-
Intercept + **Capacity**	33593.505	**-6.65**	**-6.65**
Intercept + + Inters. density	31409.097	-6.50	
Intercept + LN VKT	31428.144	-6.45	
Intercept + Car ownership	31518.198	-6.18	
Intercept + School children	31746.727	-5.50	
Intercept + NOTs	32294.599	-3.87	
Intercept + Capacity + **Car ownership**	29132.765	**-13.28**	**-7.10**
Intercept + Capacity + Inters. density	29266.078	-12.88	
Intercept + Capacity + School children	29512.963	-12.15	
Intercept + Capacity + LN VKT	29658.326	-11.71	
Intercept + Capacity + NOTs	30345.477	-9.67	
Intercept + Capacity + Car ownership + **Inters. density**	27310.697	**-18.70**	**-6.25**
Intercept + Capacity + Car ownership + NOTs	28226.467	-15.98	
Intercept + Capacity + Car ownership + School children	31255.628	-6.96	
Intercept + Capacity + Car ownership + LN VKT	31696.386	-5.65	
Intercept + Capacity + Car ownership + Inters. Density + **School children**	25276.683	**-24.76**	**-7.45**
Intercept + Capacity + Car ownership + Inters. Density + LN VKT	25981.055	-22.66	
Intercept + Capacity + Car ownership + Inters. Density + NOTs	26308.823	-21.68	
Intercept + Capacity + Car ownership + Inters. Density + School children + **NOTs**	23876.451	**-28.93**	**-5.54**
Intercept + Capacity + Car ownership + Inters. Density + School children + LN VKT	31556.101	-6.06	
Intercept + Capacity + Car ownership + Inters. Density + School children + **LN VKT**	22553.608	**-32.86**	**-5.54**

4.3 MODEL VALIDATION

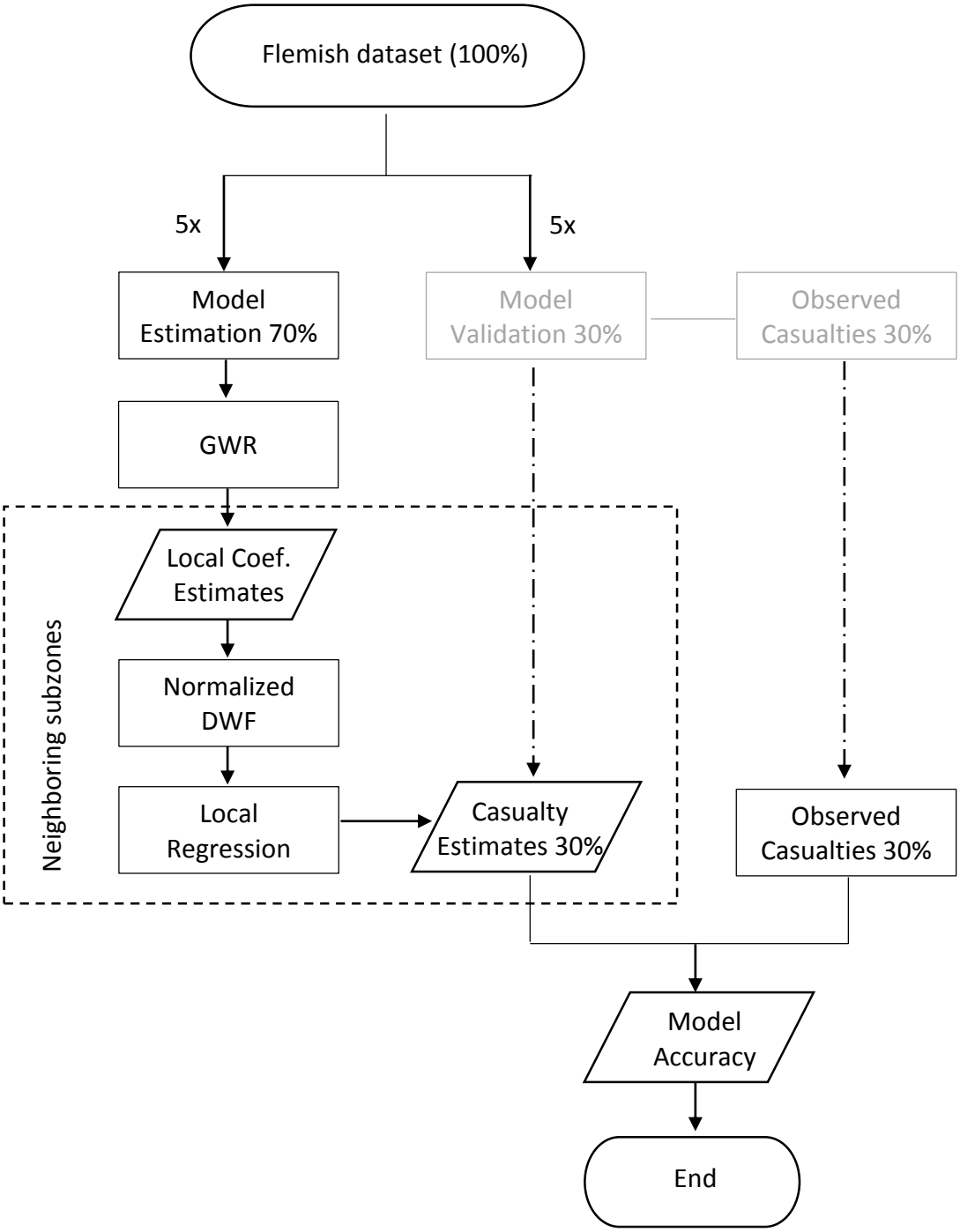
Analyses at this subsection were carried out aiming to check the validity of the produced Flemish GWPR models. This included empirical work where model accuracy was assessed by the corresponding goodness of fit obtained with a pre-determined validation sample. Taking this into account, we implemented the concept within the holdout framework, meaning that

part of the available dataset was used for model estimation (70%) and a different part for validating (30%) the model. However, the basis of the GWR approach is to generate local specific models, which implies that validation with GWR cannot be accomplished by conventional methods. To overcome this limitation, we developed, and propose in the next subsection a GWR validation framework. Yet, in order to improve the reliability of estimates within the validation samples, we repeated the process within the holdout method for five random subsamples containing each 70% and 30% of the data. This enabled us to average the performance measures of these different interactions, thus yielding an overall goodness of fit. This concept, also known as the repeated holdout, was detailed in Chapter 3, and is discussed in the next subsection along with the proposed GWR validation framework.

4.3.1 PROPOSAL OF A HOLDOUT METHOD WITH GWR

In order to improve the reliability of casualty estimates and enable appropriate inferences regarding the proposed methodology, we adopted the concept of repeated holdout. Analyses were conducted on five different Sub-Samples (SS), each with 100% of the data used in the Flemish improved models. Each SS was randomly split into two disjoint parts, i.e. 70% was used for model estimation and 30% was used for validating the models on unseen data. Such approach is illustrate in Figure 4.13, followed by a description of the procedure and results of model performance.

Figure 4.13 - GWR holdout method (GWR holdout1)



Given the basis of GWR and attributed features on previous analyses in this chapter, the geographical weighting scheme was defined as follows:

- Bandwidth size: defined by the optimal number of nearest neighbors chosen by the software used, for the primary analysis with the whole dataset;
- Bandwidth selection criteria: AICc;
- Kernel type: bi-square, given by the Equation 4.1.

$$W_{ij} = \begin{cases} (1 - {d_{ij}}^2/b^2)^2 & \text{if } d_{ij} < b \\ 0 & otherwise \end{cases} \tag{4.1}$$

Where W_{ij} is the measure of contribution of location j when calibrating the model for location i. d_{ij} represents the distance between locations i and j and b is the bandwidth.

Hence, produced local weights were normalized and assigned in the validation samples to the nearby coefficient estimates produced with GWR with 70% of the data. This exercise was carried out separately for each explanatory variable. Then, produced weighted coefficient estimates for the neighboring subzones were consolidated and used to compose the local regressions. Hereafter, these regressions were adjusted with the observed information in the validation samples, and upon that yielding the final estimates. At the end of this process, GWR models, for which we used the weighted coefficient estimates and observed values of the explanatory variables, were developed and called "GWR holdout1". Finally, casualty estimates with GWR holdout1 were compared to the observed information, at the validation areas, by means of MSPE and PCC. Table 4.12 shows the results of the model performance for both dependent variables.

Table 4.12 - Model performance (GWR holdout1)

		SS1	SS2	SS3	SS4	SS5	Average
Motorized Transport	**PCC**	0.371	0.389	0.446	0.399	0.333	**0.39**
	MSPE	2281.07	1975.39	2028.31	2115.43	2228.11	**2125.66**
Active Transport	**PCC**	0.353	0.342	0.288	0.263	0.445	**0.34**
	MSPE	417.07	471.95	799.87	682.8	508.12	**575.96**

Results of model performance enables us to confirm the suitability of the GWPR improved models and the proposed GWR holdout method. Such an approach would enable future

studies to adjust the local features of GWR within the concept of model estimation and model validation datasets, and explore other potentialities of such a powerful tool.

4.4 CHAPTER DISCUSSIONS AND CONCLUSIONS

The difficulty in obtaining crash-related information in Brazil, and its consequences in terms of model performance and development of potential studies that could help to understand the crash phenomena and enforcement of appropriate countermeasures were the major motivations to carry out this investigation. Although some data can be found in different road departments, police, health and census online sources (i.e. DATASUS, IBGE and DENATRAN), there is no link between their databases, and none of them is able to provide a full and effective data source with regard to road casualties. Therefore, the absence of a comprehensive and complete database hampered the evaluation and follow-up of national road safety programs, as well as the development of studies that could contribute to national goals toward road safety.

Particularly concerning the explanatory variables, most available information is restricted to socioeconomic and demographic variables. In spite of their merits, socioeconomic and demographic variables are often highly correlated with each other, and are not appropriate for safety-planning purposes, e.g. implementation of safety countermeasures. As Brazil, other developing countries have faced these challenges, and this drawback has unfortunately led to the use of poor and unreliable CPM to promote road safety in these countries. Whereas the limitation of prediction models to socioeconomic variables leads to failures in terms of countermeasures, unreliable and omitted information are disadvantages as this lack has a negative impact on the models in statistical terms.

In view of the foregoing and answering RQ1, this investigation aimed to demonstrate the potential improvements in the performance of spatial crash prediction models by means of a more diverse dataset, including some additional potential explanatory variables. To this end, benchmarking was carried out based on macro-level CPM developed with available fatality/crash-related information from São Paulo and Flanders.

In contrast to developing countries, European countries such as Belgium have invested a great deal of time and money in obtaining crash-related information and make them available through public channels and to academia. This practice has led to outcomes such as new strategies and studies, and this trade-off has brought improvements in traffic safety by reducing number of casualties.

Contrary to the Brazilian case, the Flemish dataset, in addition to significant information related to socioeconomic, sociodemographic and road network provided foremost diverse and suitable exposure variables. From a statistical perspective, this convenience contributed to the outperformance of the improved models, for both dependent variables (active and motorized transport). In the model developed for motorized transport, reductions of approximately 20% and 25% were observed compared with the basic model, for AICc and MSPE respectively. Likewise for the active mode, 25% and 35% reductions were found for AICc and MSPE. Moreover, Flemish models at the second stage, presented a powerful set of coefficient estimates together with suitable coefficient signs. One potential outcome of the resulting macro-level CPMs could be the identification of hot zones together with their major influence factors. Such results could be used as a reference to microscopic investigations, and the implementation of suitable safety countermeasures' enforcement in a long-term transportation planning processes, therefore helping road safety planners to prioritize data collection, besides its financial consideration.

Above all, this investigation highlighted the strong dependency of spatial CPMs on suitable and more diverse input information, enabling these models to perform as a "powerful tool", as is usually found in the literature and properly explore the spatial dependence of crash data. Crashes are caused by multiple factors that vary locally, and this complexity implies that ideally, casualties are best predicted through a set of appropriate predictive variables, including, at least, one potential exposure variable.

By modelling casualties in Flanders, based on the equivalent available data in São Paulo, results were found not to be suitable. Apart from producing unreliable coefficient estimates, they would also not be useful for safety planning and practical aspects. In other words, despite

the efforts to improve the statistical fit of the crash prediction models and associate crashes with the available explanatory variables in such circumstances, these models fail to comprehensively explain road casualties and, therefore, diminish the ability of applying suitable safety countermeasures. On the contrary, by modelling casualties based on the entire available data in Flanders, a better model overall fit for active and motorize transport modes was obtained. This suggests that a more diverse set of appropriate explanatory variables, including a relevant exposure variable, helps perhaps to address problems with counterintuitive signs and omitted variable bias. Moreover, in the ideal scenario, it could help policy makers to determine local appropriate countermeasures toward safety promotion (e.g. by altering the speed limits and intensifying local speed enforcement, improving intersections, installing traffic management and control systems, implementing crosswalks, etc.).

It is also worth mentioning that variables included in the Flemish models, besides having been found to be significant in previous studies as well as in the present one, are more accessible than those used in microscopic analysis (e.g., driving data, braking and steering information or variables related to weather conditions). Moreover, they are just examples of other potential information that could be used to develop those predictive/descriptive models. Variables that are used in the models developed for Flanders, could be interesting suggestions for extra data collection in Brazil, as other local variables could also play a significant role, other than those included in the Flemish models.

Subsequently, a sensitivity analysis was carried out allowing us to assess the statistical contribution of each variable in the prediction model performance, thus answering RQ2. Especially for countries where data is limited, either because of the lack of financial resources or other imposed conditions, this practice could also empower policy makers and responsible offices to prioritize data collection. For instance, results revealed that, the information concerning road capacity would signify a major statistical contribution for models, for both dependent variables. This is different from NOTs, for instance, which often have priority in data collection, but as revealed in this study, would not bring such a significant contribution to the Flemish models, neither for active, nor for motorized transport.

This investigation could have more value if similar analyses were carried out in different regions, based on their available information. The consolidation of the produced results would enable, for instance, the development of a solid benchmark and, therefore, validating the priorities outlined in this study and helping to determine the importance of different variables to model performance in different areas. In addition to this, for future studies, we suggest a more in-depth investigation addressing problems, such as omitted variable bias and endogeneity, as they could help to verify the validity of the assumptions of this study. One possible investigation could be for instance, to perform micro-level analysis in the identified hot zones, therefore assessing model performance by adding and removing variables to the models.

Finally, results obtained within a GWR holdout framework confirmed the suitability of the GWPR models, thus answering RQ3 and validating the proposed method.

5 INVESTIGATION OF SPATIAL MODEL PREDICTIONS ACCURACY AT UNSAMPLED SUBZONES

Data unavailability is a challenge often faced by researchers and policy makers in Brazil and many other developing countries. Unfortunately, this drawback has increasingly discouraged academia, and thus potential studies that could contribute toward road safety promotion. Taking this into account, in this chapter, we aim to test the accuracy of Geographically Weighted Regression (GWR) to predict casualties in locations where data is incomplete (e.g. due to lack of resources, procedures, political will, etc.). This helps us to answer "RQ4: *In case of data unavailability, would the produced models be suitable to estimate unsampled unit of areas*?". To this end, we extend the empirical evaluation discussed in Chapter 4 and estimate casualties of the missing subzones based on the casualty estimates produced for 70% of the data. This implies that, we not only use weighted coefficient estimates of the surrounding subzones, but also the input information of the neighboring. At the end of this process, a novel GWR validation approach within the framework of repeated holdout is proposed and called GWR holdout2. Yet, in order to verify the validity of the suggested procedure, two missing-data imputation approaches are carried out, enabling us to assess models' performance and draw the conclusive inferences.

5.1 GWR INTERPOLATION APPROACH INTO THE HOLDOUT FRAMEWORK

Likewise for the validation procedure proposed in Chapter 4 (See subsection 4.3.1), the concept of repeated holdout was used at this stage. Considering this, Sub-Samples (SS) used in that investigation for model estimation (70%) and model validation (30%) were reproduced at this stage with equal geographical weighting scheme. However, in this investigation, we modeled casualties based on the casualty estimates produced for 70% of the data with GWR, and omitted the available information at the validation samples. This implies that analyses were performed based on the information of the nearest neighboring, only. Models developed at this stage were called GWR holdout2.

In GWR holdout2, distance weights derived from the normalized kernel function, i.e. Distance Weighted Function (DWF) were assigned to the produced GWR local estimates with the validation dataset. Thereafter, the consolidation of the produced local weighted estimates, for the neighboring subzones, resulted in our final estimates at the validation samples. Hence, casualty estimates with GWR holdout2 were compared to the observed number of casualties at the validation subzones, by means of Mean Squared Prediction Error (MSPE) and Pearson Correlation Coefficient (PCC). Figure 5.1 presents the employed approach followed by the results of model performance in Table 5.1.

Figure 5.1 - GWR validation approach (GWR holdout2)

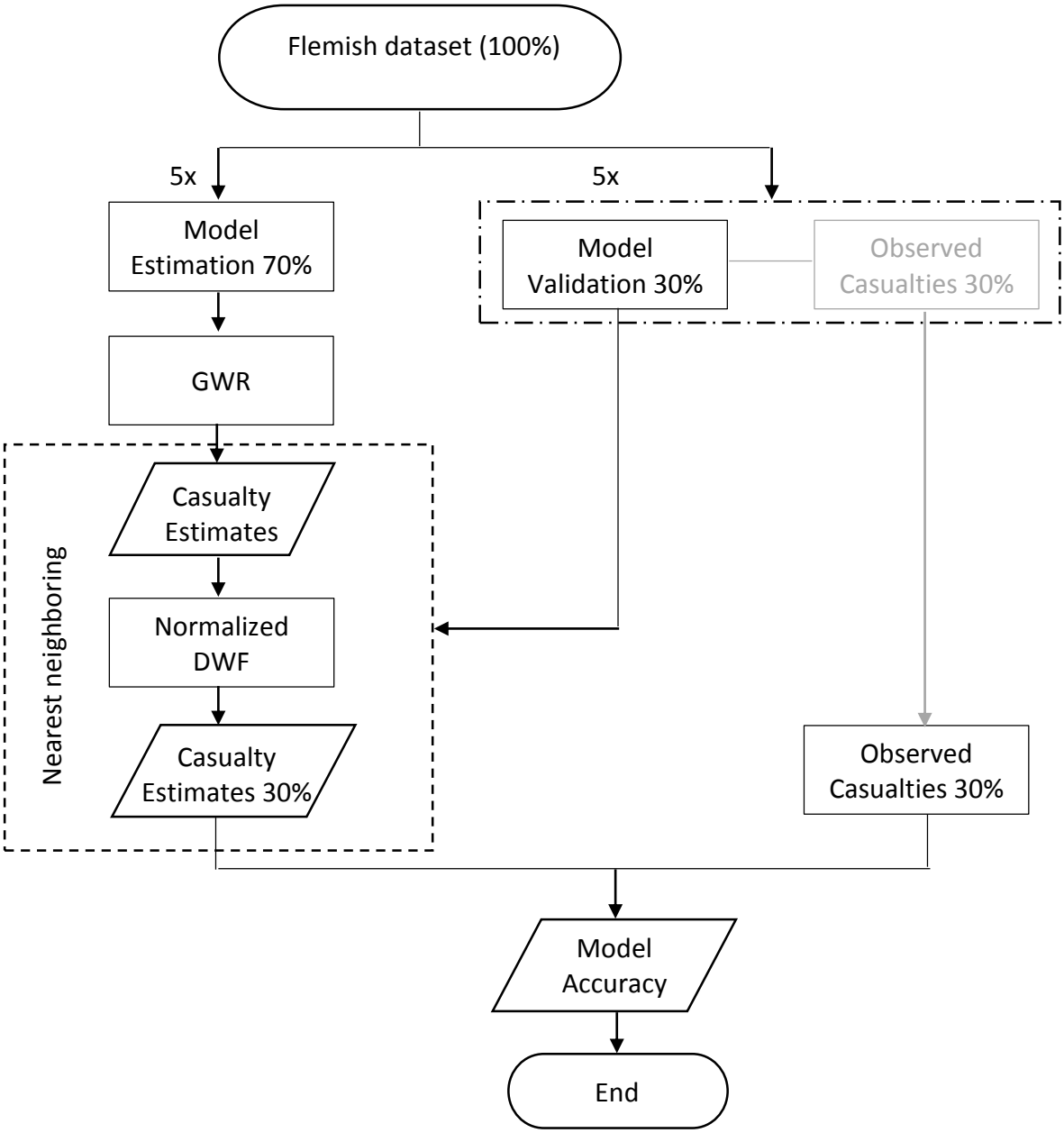


Table 5.1 - Model performance (GWR holdout2)

		SS1	SS2	SS3	SS4	SS5	Average
Motorized Transport	**PCC**	0.413	0.39	0.493	0.445	0.431	**0.43**
	MSPE	2079.39	1893.99	1816.74	188.76	1877.31	**1909.64**
Active Transport	**PCC**	0.374	0.296	0.491	0.411	0.458	**0.41**
	MSPE	390.14	484.61	606.13	551.76	504.51	**507.43**

Results of model performance confirmed the suitability of the proposed procedure fitting GWR into a cross-validation framework. Compared to GWR holdout1, GWR holdout2 outperformed, presenting improvements between 10% and 20% for both explanatory variables in terms of PCC and MSPE. In order to corroborate with these inferences and test the validity of the proposed method, two missing-data imputation approaches were carried out. Subsequently, we compared their performance to the results obtained here. These methods and their outcomes are discussed in the next subsection.

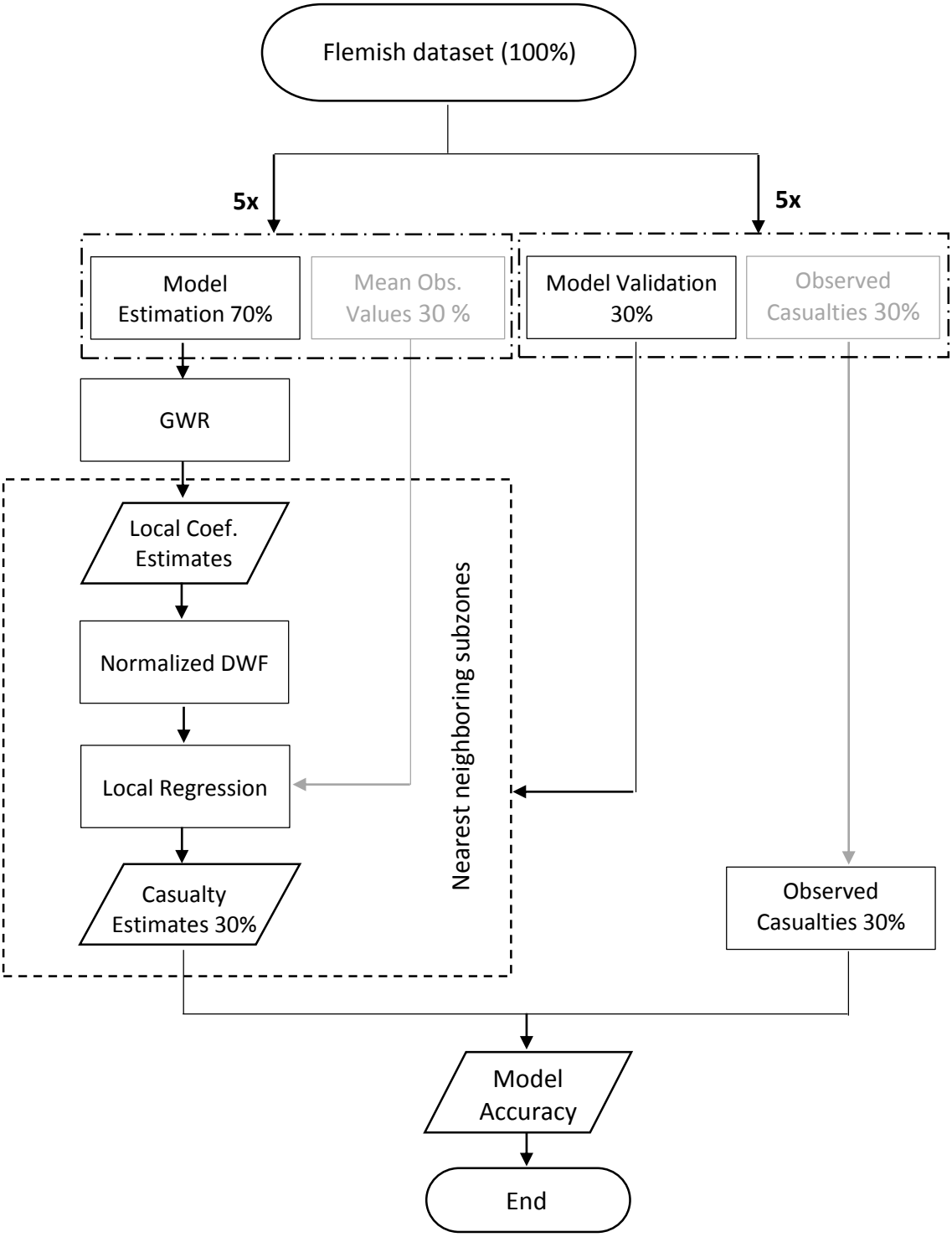
5.2 MISSING-DATA IMPUTATION

In this investigation, prediction models were framed into two common imputation methods, i.e. Mean imputation (MEI) and K-nearest neighbor (KNN) imputation. In order to carry out analyses as close as possible to the previous GWR holdout approaches, casualties at unsampled subzones were estimated based on the information of the nearest neighboring and geographical weighting scheme, previously employed. This allowed us to compare the produced models fairly. Thereafter, these approaches were handled under different systematics as follows.

In the models developed with MEI, local regressions at validation samples were adjusted with the mean of the observed values from the database with 70% of the data. This means that, one single value for each explanatory variable was imputed for all cases. Hence, weights, derived from the normalized kernel function, were assigned to the local coefficient estimates produced with GWR for 70% of the data. This process was held separately, for each

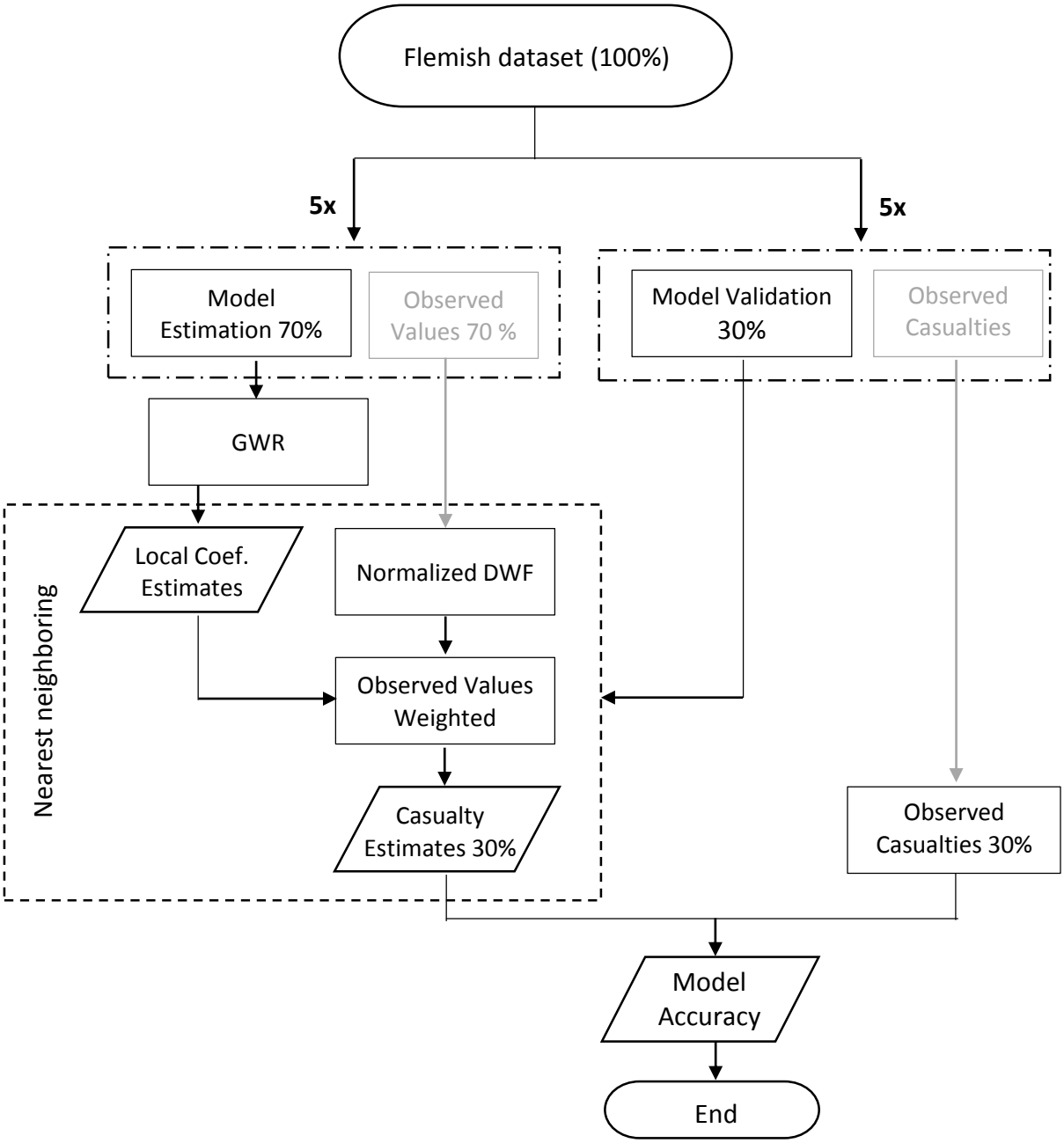
explanatory variable. Thereupon, weighted outcomes were averaged and the produced mean values were used to compose the local regressions. Finally, casualty estimates were compared to the observed number of casualties at the validation subzones, by means of MSPE and PCC. Figure 5.2 presents a schema of the methodology adopted.

Figure 5.2 - Holdout method based MEI



Relying on the k nearest neighbors for a given missing value, in KNN, distance weights were assigned to the observed values of the explanatory variables for the nearest neighboring. Then, produced values were assigned to the local GWR coefficient estimates. Subsequently, the consolidation of these local weighted estimates for the neighboring subzones and explanatory variables resulted in our final estimates at the validation samples. Hence, casualty estimates were compared to the observed number of casualties, by means of MSPE and PCC. Figure 5.3 illustrates how this process was followed.

Figure 5.3 - Holdout method based – KNN



Tables 5.2 and 5.3 show the results of model performance, for motorized and active transport.

Table 5.2 - Data imputation model performance for motorized transport

		SS1	SS2	SS3	SS4	SS5	Average
MEI	**PCC**	0.381	0.363	0.397	0.365	0.354	**0.37**
	MSPE	2172.81	1935.77	2016.83	2031.87	2003.41	**2032.14**
KNN	**PCC**	0.416	0.386	0.497	0.434	0.427	**0.43**
	MSPE	2108.71	1902.86	1839.89	1920.76	1875.51	**1929.55**

Table 5.3 - Data imputation model performance for active transport

		SS1	SS2	SS3	SS4	SS5	Average
MEI	**PCC**	0.371	0.301	0.42	0.291	0.359	**0.35**
	MSPE	392.67	408.3	662.27	614.17	567.57	**543.60**
KNN	**PCC**	0.383	0.307	0.485	0.387	0.425	**0.4**
	MSPE	385.15	481.78	634.29	578.7	535.84	**523.15**

Results of the data imputation approaches, revealed an outperformance of the KNN method for both dependent variables. Improvements approximately to 15% and 5% were found respectively for PCC and MSPE, suggesting that preserving the original data and its structure led to a lower distortion of the distribution of the imputed data.

As it can be seen in Table 5.4, our GWR cross validation approach (GWR holdout2) outperformed the imputation models, therefore confirming the validity of our approach and the inferences drawn from its application. For both dependent variables, higher PCC and lower MSPE were found by accounting for the produced estimates of the nearest neighboring.

Table 5.4 - General view of model performance

			SS1	SS2	SS3	SS4	SS5	Average
Motorized Transport	**MEI**	**PCC**	0.381	0.363	0.397	0.365	0.354	**0.37**
		MSPE	2172.81	1935.77	2016.83	2031.87	2003.41	**2032.14**
	KNN	**PCC**	0.416	0.386	0.497	0.434	0.427	**0.43**
		MSPE	2108.71	1902.86	1839.89	1920.76	1875.51	**1929.55**
	GWR holdout2	**PCC**	0.413	0.39	0.493	0.445	0.431	**0.43**
		MSPE	2079.39	1893.99	1816.74	188.76	1877.31	**1909.64**
Active Transport	**MEI**	**PCC**	0.371	0.301	0.42	0.291	0.359	**0.35**
		MSPE	392.67	408.3	662.27	614.17	567.57	**543.60**
	KNN	**PCC**	0.383	0.307	0.485	0.387	0.425	**0.4**
		MSPE	385.15	481.78	634.29	578.7	535.84	**523.15**
	GWR holdout2	**PCC**	0.374	0.296	0.491	0.411	0.458	**0.41**
		MSPE	390.14	484.61	606.13	551.76	504.51	**507.43**

5.3 CHAPTER DISCUSSION AND CONCLUSIONS

Analyses in this chapter were carried out aiming to test the accuracy of GWR to predict road casualties in locations where data is incomplete. To enable this investigation, we used the models produced in Chapter 4, and omitted part of the available information in the modelling process. In order to assess the accuracy of the models, analyses were held in the repeated holdout framework, taking into account the nearest neighboring information only. In this context, a new GWR cross-validation approach was developed, and validated, by comparing its overall fit in relation to two common imputation methods. We chose using MEI and K-nearest neighbor, given their friendly interface and attributes, which could be easily adapted to both GWR validation procedures suggested in this manuscript, thus enabling a fair comparison between them. While in MEI, the mean of the observed values of the explanatory variables was the value which was imputed, in KNN, this process relied on the observed values of the explanatory variables itself.

Results showed that our novel GWR validation approach – GWR holdout2 (slightly) outperformed other competing ones in its overall accuracy. Moreover, this finding enabled us to confirm the effectiveness of the approach in relation to our previously suggested one (GWR holdout1), thus answering the RQ4 (*In case of data unavailability, would the produced models be suitable to estimate unsampled unit of areas*?). In particular, four findings are noteworthy from this investigation:

(1) Results of model performance give an indication of the overall reliability of GWR in the holdout framework to estimate unsampled subzones, thus answering the main question of this investigation. Moreover, results showed GWR as an effective tool, and confirmed the suitability of the modelling attributes selected by the software used to foresee other domain;

(2) Our proposed GWR validation approach is suitable to estimate road casualties at unsampled areas (i.e. zones with missing information), based on the nearest neighboring information only. GWR holdout2 outperformed common imputation methods to fill out missing information. Although advantageous by their simplicity, and acceptable if the variation of the data is low, these methods are just approximations, thus yielding different kind of bias, especially for MEI. Nonetheless, it could be the case that these models result in apparent satisfactory results;

(3) Findings with GWR holdout2 corroborated the assumption of Tobler (1970), i.e. first law of geography, suggesting that directly using casualty estimates of the neighboring subzones to estimate outcomes for the missing subzones is less sensitive to model inaccuracy compared to GWR holdout1.

(4) As we would expect, KNN gives the best trade-off between imputation errors and data structure, when compared to MEI, meaning that preserving the original data is a better option.

6 EVALUATION OF SPATIAL DATA ANALYSIS APPROACHES ON CRASH PREDICTION

Investigations in this chapter are carried out aiming to evaluate the performance of multivariate spatial data analysis approaches, based on Geostatistics and Geographically Weighted Regression (GWR) to estimate road casualties. To this end, Flemish improved models are adjusted in the repeated holdout framework by means of Kriging with External Drift (KED). Thereafter, possible linkages between GWR and KED are discussed, as well as their advantages and disadvantages toward road safety. This enables us to answer the last Research Question (RQ5) of this thesis: *"Considering geostatistics, by means of KED, what is the most suitable method to explore the spatial dependence of crash data and solve issues involving missing information?"*

6.1 KRIGING WITH EXTERNAL DRIFT (KED)

As previously discussed in Chapter 2, Geostatistics can be very powerful in crash modelling processes, therefore assisting road safety-planning studies. Despite the fact that in the past, Geostatistics was mostly linked to spatially continuous problems, such as geology and earth sciences, the technological advances and the availability of geocoded information has enabled its adoption in different fields. As a result, geostatistics is now commonly applied to natural and social sciences (Goovaerts, 1997). Specifically for transportation planning studies, Geostatistics by means of its intrinsic characteristics, has enabled advantageous outcomes, therefore supporting its application on crash data, which we believe to be a potential line of research.

In this study, Geostatistics is performed by means of KED as it uses secondary information to assist in the interpolation. In relation to univariate geostatistical and non-geostatistical methods, KED is advantageous as it incorporates the local trend within the neighborhood search window as a linear function of some explanatory variables, rather than the coordinates

(Goovaerts, 1997). Hence, the primary variable is estimated based on the secondary one, as both are highly correlated.

Taking this into account, in this study, the primary variable was defined as the observed number of casualties in a TAZ. Given the inherent characteristics of the KED estimator and aiming to compare its model performance to GWR, the secondary variable was defined as the values estimated through a Generalized Linear Model (GLM). GLMs, in turn, were constructed based on the same explanatory variables used in the previous Flemish improved models (in Chapters 4 and 5), i.e. Number of Trips (NOTs), number of children attending school (school children), road capacity (capacity), intersection density, car ownership, and Vehicle Kilometers Traveled (VKT), as the exposure variable.

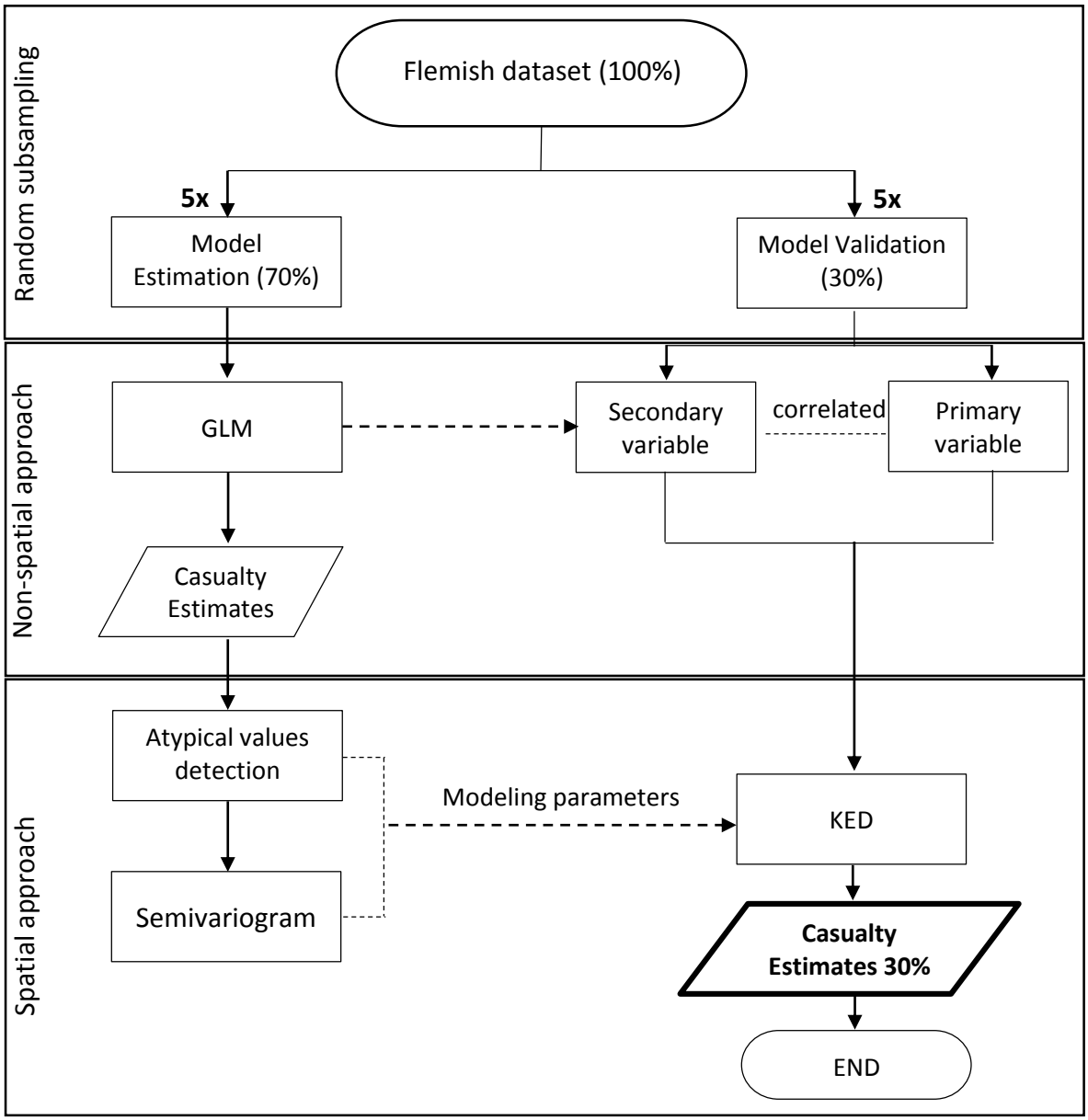
Considering the above, the method framework adopted to estimate casualties in Flanders, can be summarized in a two-step method, comprising: (1) non-spatial approach, by means of GLM; and (2) spatial approach, through the application of the KED taking into account casualty estimates obtained from a GLM. Analyses with empirical models were carried out within the repeated holdout framework. This means that, in the first step, the equation resulting from the GLM calibrated with 70% of the data, for each subsample (SS), was used to estimate the secondary variable for the remaining 30% (for each SS). Likewise, in the second step, experimental and theoretical semivariograms were adjusted for 70% of the data (for each SS) and reproduced in the validation subsamples. Figure 6.1 illustrates this systematic, followed by a detailed overview of the tasks and outcomes at each stage. In order to provide a better insight of these tasks, we subdivided them in seven main procedures, which are featured in the schema and discussed as follows:

1. Random segregation of the complete dataset (100%) into two sets used for model estimation (70%) and validation (30%).
2. Random subsampling of the holdout method (k=5, meaning 5 complete subsamples with 100% of the information, randomly divided in 70/30).

Thereafter, for each of these subsamples:

3. Estimation of the GLM model for the 70% dataset;
   3A. Estimation of the secondary variable (70%).
4. Validation of the GLM model (process 3) for the 30% dataset;
   4A. Estimation of the secondary variable (30%).
5. Verification of correlation between primary variable, i.e. observed casualties, and secondary variable, i.e. estimates produced in process 4A (30%).
6. Experimental semivariograms calculation and adjustment of theoretical semivariograms for the secondary variable, (70% outcome from process 3A).
7. KED (30%).

Figure 6.1 - Two-step procedure within the repeated holdout framework



Firstly, the complete database of the improved Flemish models (discussed in Chapter 4 and namely here as SS1), was randomly segregated into two sets with 70% and 30% of the data, respectively for model estimation and validation. Subsequently, a random subsampling was carried out five times (each time with 100% of the data), resulting in five random subsets, each with 70% of data, and five random subsets, each with 30% of data. Aiming to compare and evaluate the performance of all multivariate spatial data analysis approaches, which were used in this research, we performed the analyses in this chapter with the same five subsamples (70/30) used to produce the empirical models in Chapters 4 and 5. Hence, the two-step procedure was carried out as follows.

6.1.1 STEP 1: GENERALIZED LINEAR MODELS

The first procedure within the non-spatial approach, concerns the construction of GLMs to estimate the secondary variable. GLMs were estimated for the model estimation subsamples, based on the set of independent variables used to construct the previous GWR Flemish models, in Chapters 4 and 5. Two outcomes are produced as results of this process: (a) casualty estimates for 70% of the data, which later was used to calculate the experimental semivariograms, and (b) the calibrated equation used to estimate the secondary variable for the set of 30% of the data, expressed in Equation 6.1. The global parameters estimates (β), which were used to calibrate the equations for active and motorized transports, at each subsample, are presented in Table 6.1.

$$ln[E(C)] = \beta_0 + \beta_1 \ln(x_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 \quad (6.1)$$

Where:

$E(C)$: expected casualties;

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and β_6: model parameters;

x_1: exposure variable (i.e. VKT);

x_2, x_3, x_4, x_5 and x_6: other explanatory variables (i.e. NOTs, school children, capacity, intersection density and car ownership, respectively).

Table 6.1 - Global model parameter estimates

	Explanatory variables	SS1*	SS2	SS3	SS4	SS5
Motorized Transport	Intercept	2.828	2.846	2.800	2.816	2.864
	NOTs**	0.000057	0.000069	0.000077	0.000068	0.000077
	School children	-0.00011	-0.00011	-0.00012	-0.00011	-0.00013
	Capacity	0.000096	0.000068	0.000066	0.000096	0.000068
	Intersection density	0.034	0.032	0.031	0.030	0.028
	Car ownership	-0.105	-0.106	-0.077	-0.138	-0.093
	Ln VKT	0.071	0.073	0.072	0.073	0.069
Active Transport	Intercept	1.200	1.384	1.021	1.413	1.178
	NOTs	0.00015	0.00016	0.00016	0.00015	0.00019
	School children	-0.00015	-0.0017	-0.000089	-0.00012	-0.00019
	Capacity	0.00023	0.00024	0.00021	0.00024	0.00020
	Intersection density	0.036	0.034	0.028	0.034	0.032
	Car ownership	-0342	-0.522	-0.296	-0.456	-0.296
	Ln VKT***	0.128	0.127	0.143	0.113	0.128

SS*: Subsample; NOTs**: Number of Trips; VKT***: Vehicles Kilometers Traveled

The second procedure concerns the verification of correlation between the primary and secondary variables. In addition to this, the performance of the GLMs was verified. Table 6.2 presents the results of model performance, by means of Pearson Correlation Coefficient (PCC) and Mean Squared Prediction Error (MSPE). These statistics metrics were calculated based on the observed and estimated casualties through GLMs models and the 30% validation samples.

Table 6.2 - Model performance by means of GLMs

		SS1	SS2	SS3	SS4	SS5	Average
Motorized Transport	**PCC**	0.375	0.339	0.368	0.301	0.316	**0.34**
	MSPE	2179.43	1974.26	2065.48	2128.63	2067.49	**2083.06**
Active Transport	**PCC**	0.373	0.285	0.381	0.304	0.359	**0.34**
	MSPE	386.90	482.03	657.39	599.93	551.41	**535.53**

6.1.2 STEP 2: GEOSTATISTICS BY MEANS OF KED

Casualty estimates, which were the product of the previous stage (estimated casualties through GLMs models), were used as the input data for the geostatistical modeling (variographic analysis, validation and kriging). Therefore, the first procedure within the spatial approach concerns the experimental semivariograms calculation and adjustment of model parameters of the theoretical semivariograms within the secondary variable.

Variographic Analysis

Firstly, atypical values were detected and omitted from the variographic analysis. This exercise was carried out aiming a better representation of the spatial behavior of the Regionalized Variable (RV). This process was adopted based on the Inter-Quartile Range (IQR) given by the difference between the third and first quartile (IQR=3Q-1Q), meaning that any observation that was more than 1.5xIQR above the third quartile or below the first quartile, was considered an outlier (Turkey, 1977), and was thus omitted from the variographic analysis. Table 6.3 shows the obtained values of the quartiles and IQR found for each subsample, for the primary (observed number of causalities) and secondary variables (estimated casualties through GLMs models).

Hence, experimental semivariograms for the five model estimation subsamples were calculated and adjusted by the theoretical curves. Subsequently, experimental semivariograms of the primary variables were developed, for the validation subsamples, taking into account the:

- IQR values;
- attributes of calculation of the experimental semivariograms (e.g. lag distance, tolerance, cut distance, direction);
- modeling parameters for adjustment of the experimental semivariograms to the theoretical ones, e.g. nugget (C_0), sill (C), range and model structure.

Table 6.3 - Percentiles and interquartile range

		Primary variable			Secondary variable		
		1Q	**3Q**	**IQR**	**1Q**	**3Q**	**IQR**
Motorized Transport	**SS1**	11	58	**18.5**	36	49	**68.5**
	SS2	11	60	**133.5**	36	50	**71**
	SS3	11	59	**131**	36	50	**71**
	SS4	11	58	**128.5**	36	50	**71**
	SS5	11	58	**128.5**	36	50	**71**
Active Transport	**SS1**	2	17	**39.5**	11	17	**26**
	SS2	2	17	**39.5**	11	17	**26**
	SS3	2	17	**39.5**	11	17	**26**
	SS4	2	17	**39.5**	11	16	**23.5**
	SS5	2	17	**39.5**	11	16	**23.5**

Initially, the development of the experimental semivariograms and their directions proceeded from the angle 0° (North to South) to 90° (East to West), according to the standardization of axes of geoMS. Experimental semivariograms were generated with test angles ranging from 15° to 15°, and angular tolerance of 1°. We used 100 lags, which is the maximum allowed by the employed software. For all cases (primary and secondary variables for motorized and active transport of all subsamples), the size of the lag (h) adopted was 1.100 meters, and the cutting distance, 110 kilometers. This distance was adopted based on half of the maximum length that covers the region from East to West. Hence, the spatial structure of the RVs for active and motorized transport was found to be the same in all directions, thus showing an isotropic behavior, and depending only of the magnitude of the lag vector, h. As a result, semivariograms for all cases, and secondary variables were described as omnidirectional. Based on this, semivariograms were calculated with angular tolerance of 180^0 (maximum angular aperture). Table 6.4 shows the parameters that best described the spatial structure of the samples for the primary and secondary variables, and both travel modes.

Table 6.4 - Parameters of the experimental semivariograms (for all cases)

Direction	Omnidirectional
Lag (m)	1.100
Tolerance	180^0
Number of lags	100
Cut distance (km)	110

Hence, the adjustment of the experimental semivariograms to a general function was carried out by visual inspection. For all cases, the Spherical theoretical model was the one that gave the best fit for the points of the experimental semivariograms. Table 6.5 summarizes the graphical parameters obtained by the adjusted experimental semivariograms. These parameters were the ones that best described the spatial structure of the RV, for each case.

Table 6.5 - Graphical parameters of the theoretical semivariograms

		Primary variable			**Secondary variable**		
		C_0*	**C****	**Range**	**C_0**	**C**	**Range**
Motorized Transport	**SS1**	96.717	738.765	43780.4	4.194	70.777	61283.74
	SS2	8.383	910.995	48155.295	15.136	72.392	48149.032
	SS3	10.932	852.728	39812.17	12.894	76.055	64844.778
	SS4	8.823	840.367	37966.501	3.646	89.337	65661.394
	SS5	6.136	822.199	43784.431	8.098	79.059	91935.014
Active Transport	**SS1**	3.251	80.101	30664.344	1.38	14.395	56908.739
	SS2	1.075	81.419	52530.326	3.339	13.898	48149.674
	SS3	2.948	79.92	52532.52	1.034	15.414	39398.802
	SS4	9.386	73.254	65669.934	2.126	11.133	39396.369
	SS5	1.273	84.658	48157.752	3.87	8.979	61289.821

*nugget effect; **partial sill

Figures 6.2 and 6.3 present the final theoretical semivariograms obtained after adjusting them for both, primary and secondary variables and travel modes (active and motorized). The semivariograms produced for the validation subsamples are available in Appendix C.

Figure 6.2 - Theoretical semivariograms for motorized transport (model estimation)

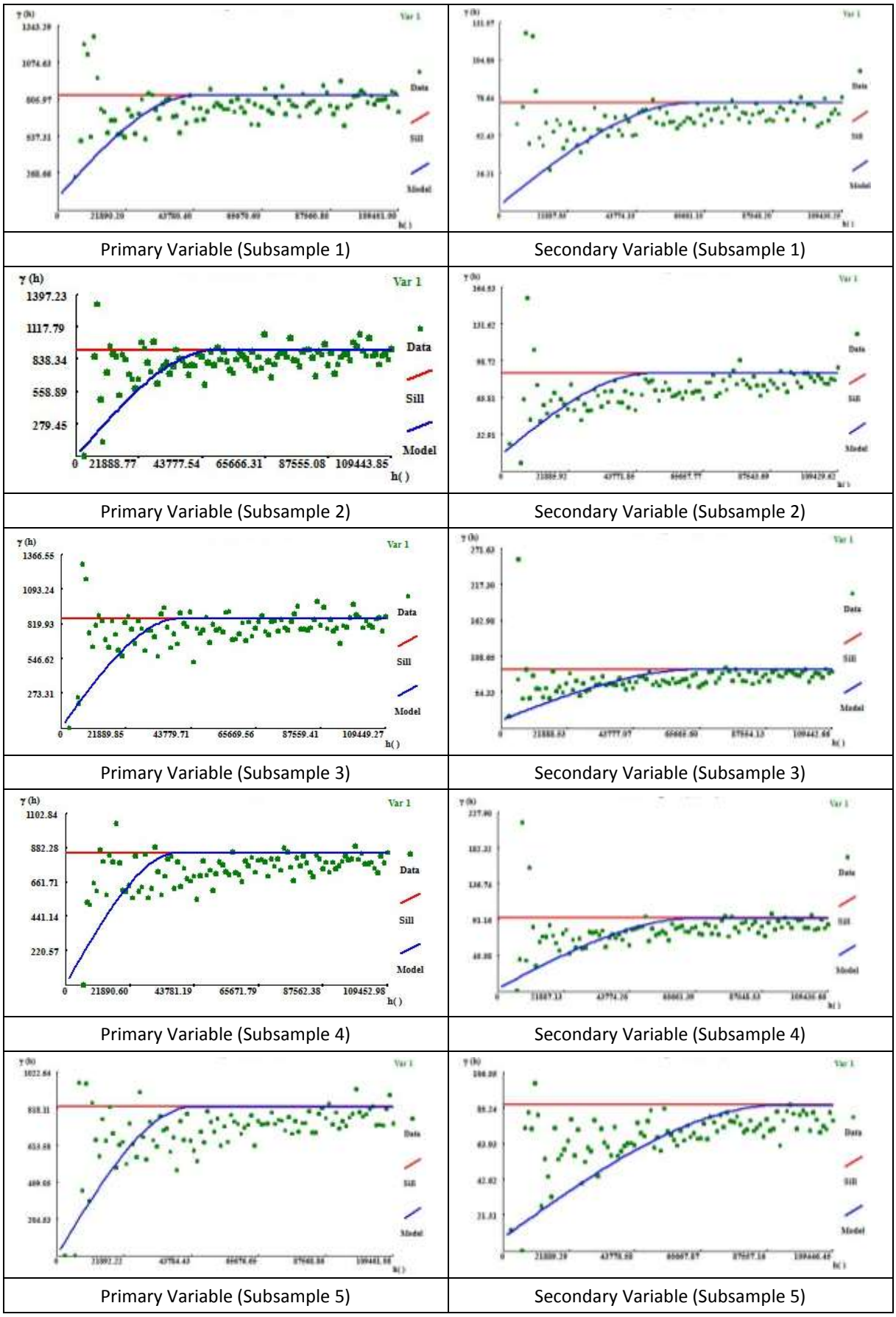
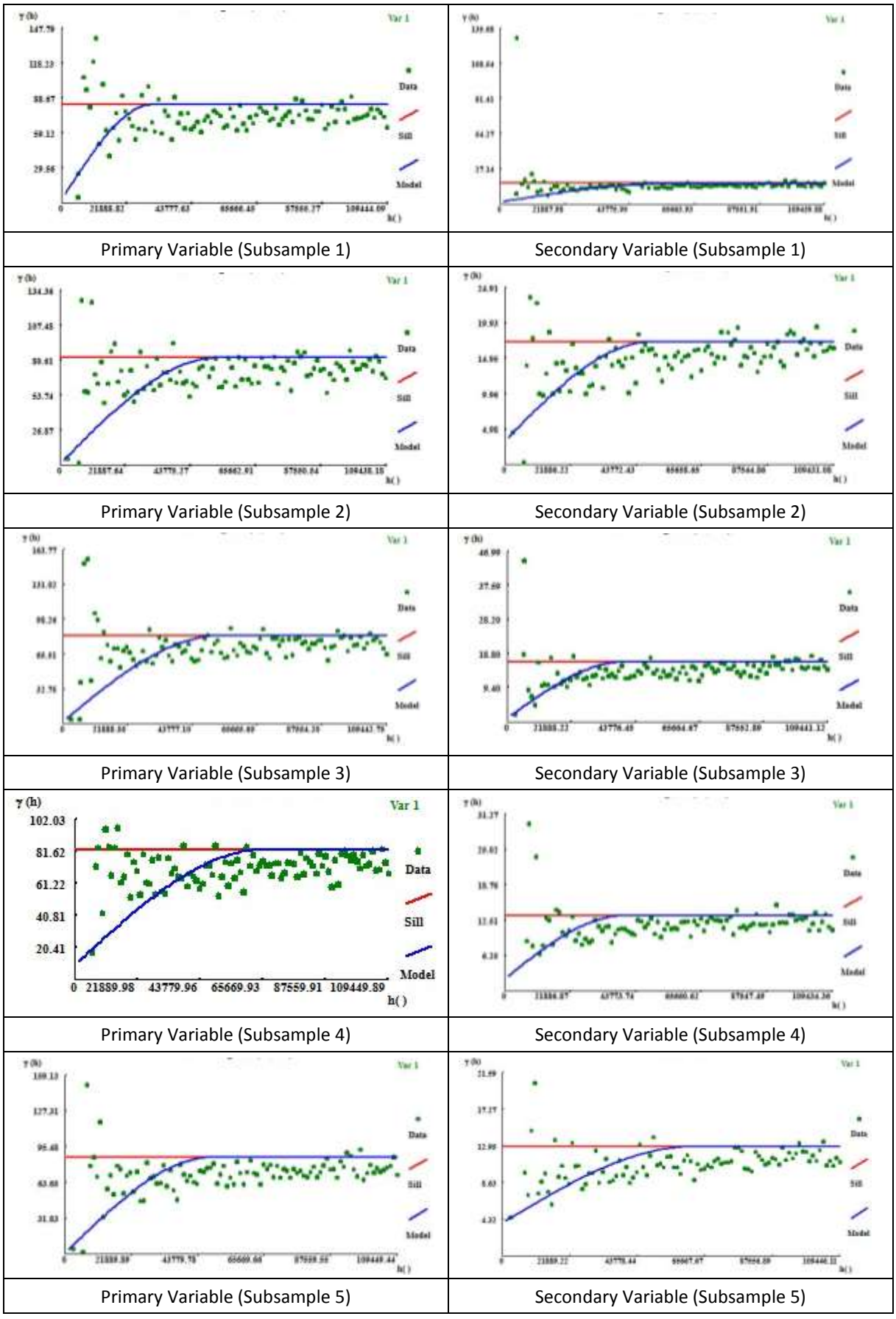


Figure 6.3 - Theoretical semivariograms for active transport (model estimation)



Primary Variable (Subsample 1)	Secondary Variable (Subsample 1)
Primary Variable (Subsample 2)	Secondary Variable (Subsample 2)
Primary Variable (Subsample 3)	Secondary Variable (Subsample 3)
Primary Variable (Subsample 4)	Secondary Variable (Subsample 4)
Primary Variable (Subsample 5)	Secondary Variable (Subsample 5)

Cross validation

In order to obtain estimated values for the samples, we proceeded with cross validation. Subsequently, estimated values were compared with the observed ones in the validation subsamples, by means of Pearson Correlation Coefficient (PCC) and Mean Squared Prediction Errors (MSPE). Table 6.6 presents the results of the adjusted semivariograms performance together with the goodness of fit of the approaches described in the previous chapters of this thesis.

Table 6.6 - General view of spatial model performance

			SS1	SS2	SS3	SS4	SS5	Average
Motorized Transport	GWR holdout1	PCC	0.371	0.389	0.446	0.399	0.333	**0.39**
		MSPE	2281.07	1975.39	2028.31	2115.43	2228.11	**2125.66**
	GWR holdout2	PCC	0.413	0.39	0.493	0.445	0.431	**0.43**
		MSPE	2079.39	1893.99	1816.74	1880.76	1877.31	**1909.64**
	MEI	PCC	0.381	0.363	0.397	0.365	0.354	**0.37**
		MSPE	2172.81	1935.77	2016.83	2031.87	2003.41	**2032.14**
	KNN	PCC	0.416	0.386	0.497	0.434	0.427	**0.43**
		MSPE	2108.71	1902.86	1839.89	1920.76	1875.51	**1929.55**
	KED	PCC	0.317	0.326	0.373	0.297	0.283	**0.32**
		MSPE	2339.29	2055.2	2127.45	2233.21	2388.02	**2228.63**
Active Transport	GWR holdout1	PCC	0.353	0.342	0.288	0.263	0.445	**0.34**
		MSPE	417.07	471.95	799.87	682.8	508.12	**575.96**
	GWR holdout2	PCC	0.374	0.296	0.491	0.411	0.458	**0.41**
		MSPE	390.14	484.61	606.13	551.76	504.51	**507.43**
	MEI	PCC	0.371	0.301	0.42	0.291	0.359	**0.35**
		MSPE	392.67	408.3	662.27	614.17	567.57	**543.60**
	KNN	PCC	0.383	0.307	0.485	0.387	0.425	**0.4**
		MSPE	385.15	481.78	634.29	578.7	535.84	**523.15**
	KED	PCC	0.38	0.338	0.419	0.374	0.301	**0.36**
		MSPE	386.11	480.74	652.24	603.79	637.43	**552.06**

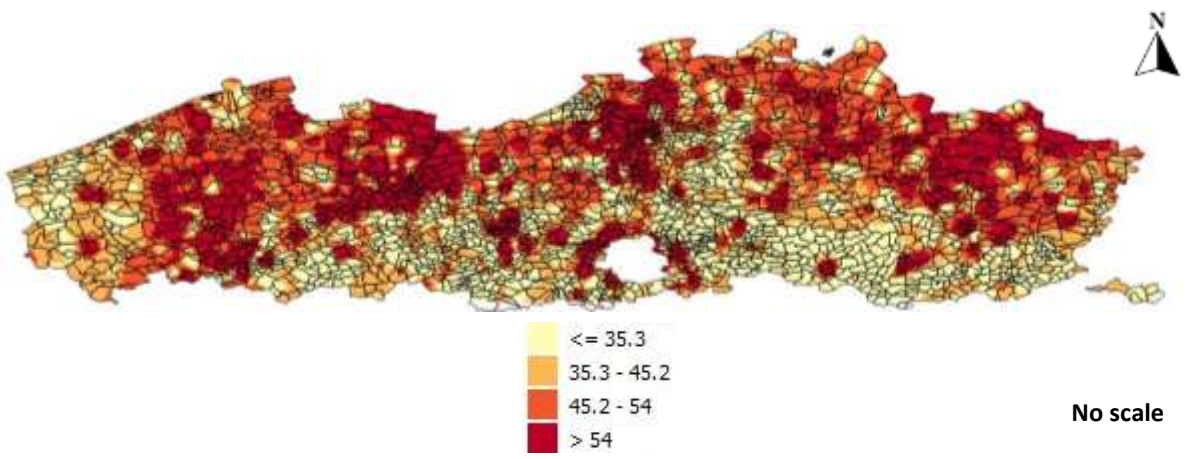
From a statistical point of view, results revealed that GWR holdout2 (proposed in Chapter 5) outperformed all other spatial approaches, for active and motorized transport mode users.

Kriging

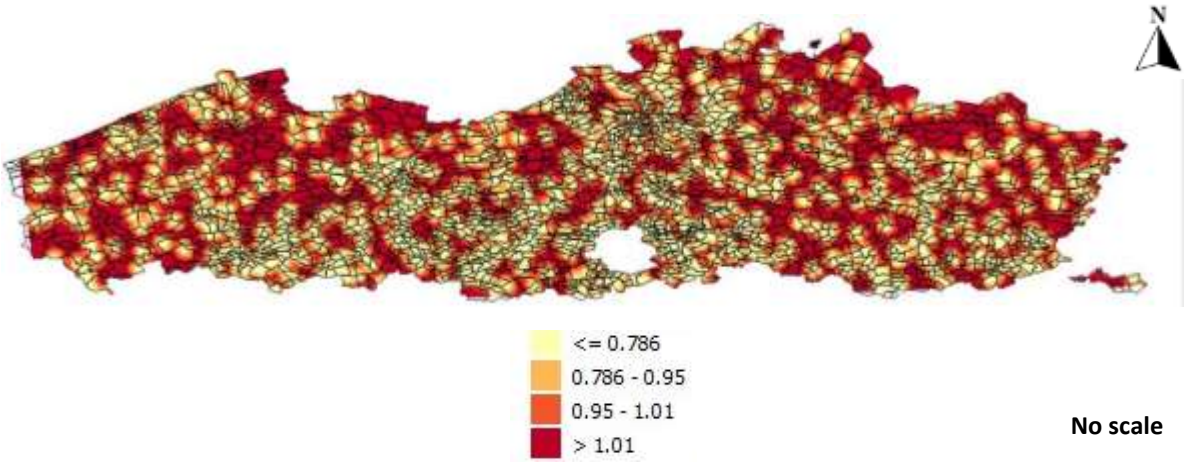
Finally, the last spatial modeling procedure concerns the kriging. The theoretical semivariogram parameters were used at the intrinsic weighting scheme of the interpolation method. The product of this process were estimates for the primary variable (casualties), which is represented by continuous surface maps in Figures 6.4 and 6.5 (for motorized and active transport, respectively), together with its corresponding variance estimation maps. The maps presented here, correspond to SS1, which was the one used as basis for the preliminary analyses in Chapter 4.

Although it is not the focus of this investigation, results of the continuous surface maps enabled us to identify some spatial patterns in the number of casualties, in Flanders. For instance, the majority of casualties for motorize transport were estimated in the regions surrounding Brussels, which is Belgium's capital. In addition to the great amount of people employed in Brussels, that live outside the city limits, in its surrounding is located the "Brussels ring", which crosses the three regions of Belgium: Flanders, Wallonia, and Brussels, thus resulting on high flow of vehicles on the highways. High estimates were also observed in the North of the Flemish provinces (i.e. West-Flanders, East-Flanders, Antwerp and Limburg), where the daily flow of vehicles is also high, especially in the highways that make their connection.

Figure 6.4 - Kriging and variance of estimation maps for motorized transport



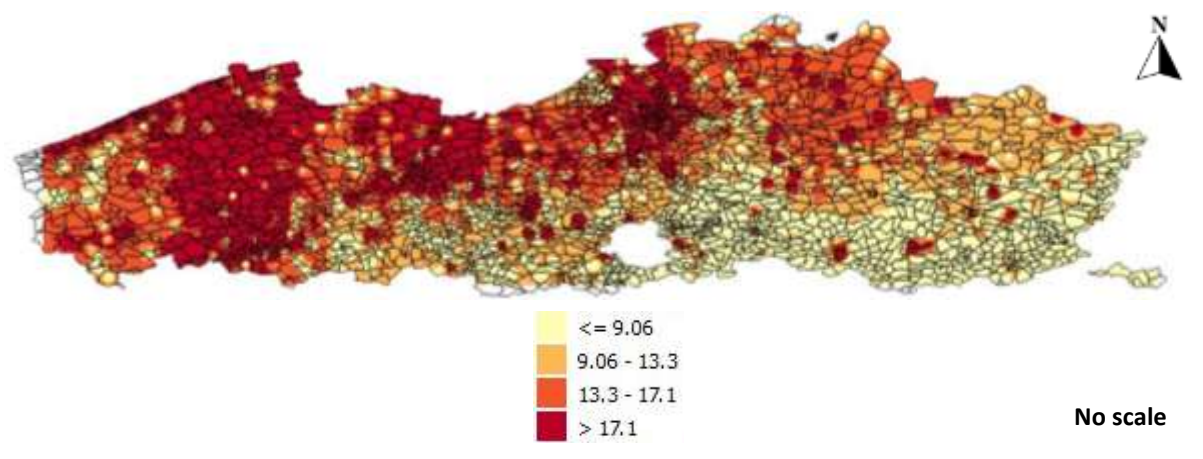
Kriging map of casualties for motorized transport – Subsample 1



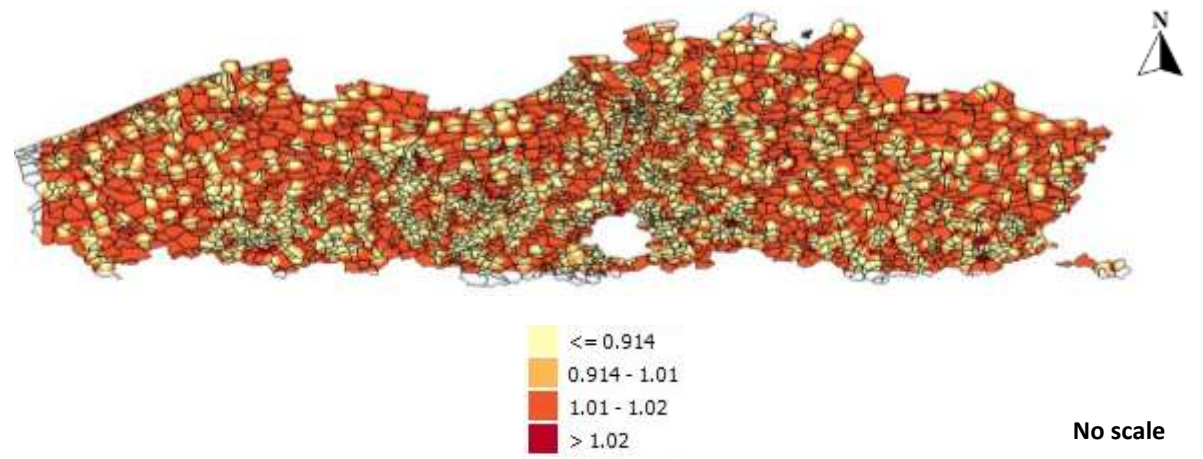
Variance of estimation map for motorized transport – Subsample 1

Concerning the spatial distribution of casualty estimates for active transport, results revealed different patterns, moreover emphasizing the differences between the northern and the southern part of Flanders. In the North, cycling is more popular than in the rest of the region, especially for utilitarian purposes, e.g. work and studies. Besides the cultural aspects, that might be involved, this could be explained by the amount of facilities available for this group, and companies/universities, which are greater in the Flemish area, therefore leading a greater number of users and their risk of being involved in a road crash.

Figure 6.5 - Kriging and variance of estimation maps for active transport



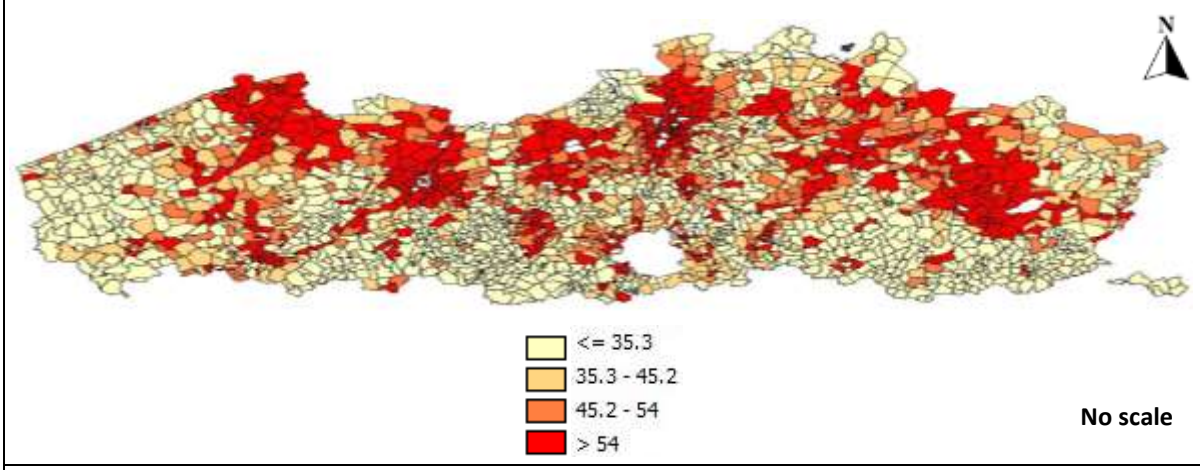
Kriging map of casualties for active transport – Subsample 1



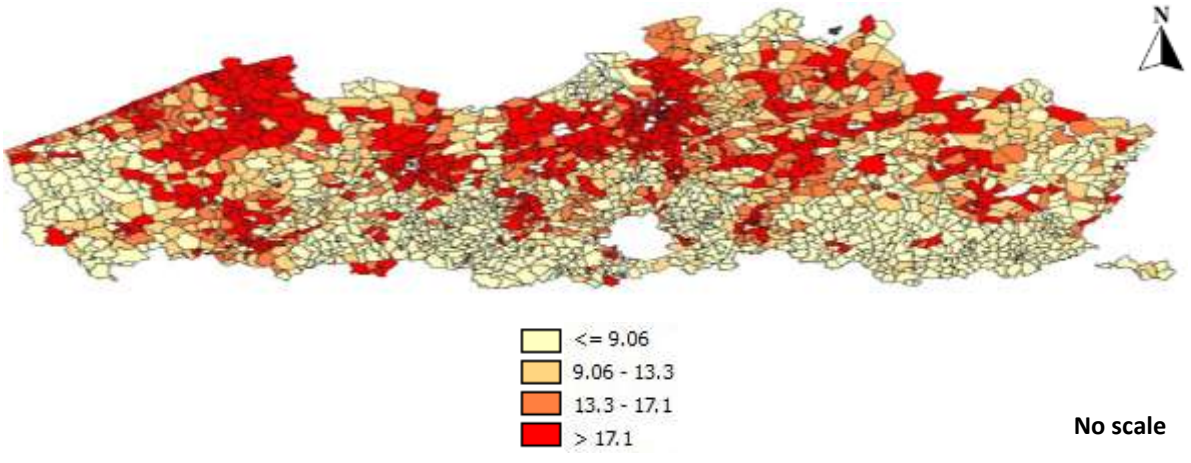
Variance of estimation map for active transport – Subsample 1

Moreover, results with the continuous surface maps of casualty estimates were consistent with the ones obtained for the improved Flemish models, through GWR, in Chapter 4 (subsection 4.1.3). Figure 6.6 presents the local maps obtained in that investigation, adjusted in the same classes of values of the continuous surface maps displayed in Figures 6.4 and 6.5.

Figure 6.6 - GWR local maps of casualty estimates



Local casualty estimates for motorized transport – Subsample 1



Local casualty estimates for active transport – Subsample 1

For both active and motorized transport, the variance of estimation presented a pattern trend distributed in the space. Therefore, lower accuracy of inference was seen in areas where casualty estimates were also low, thus corroborating with the assumptions of areas identified as hot zones. Other remarks of the spatial estimation performance together with the conclusions of this chapter are presented in the next subsection.

6.2 CHAPTER DISCUSSION AND CONCLUSIONS

Investigations in this chapter aimed to evaluate the performance of different multivariate spatial interpolation tools, based on geostatistics and GWR to estimate casualties. The choice

of working with spatial models only, was given by the clear spatial dependence of the data. This was observed in the global and local corrected Akaike Information Criterion (AICc) values, obtained with the empirical models developed in Chapter 4, thereafter confirmed by other statistical tests.

From the statistical point of view, results with KED were in line with those from GWR, MEI and KED. KED outperformed GWR (GWR holdout1), by means of PCC and MSPE for active transport. Concerning the results for motorized transport, GWR outperformed, but without compelling differences. Additionally, KED, by means of its intrinsic processes (e.g. spatial prediction, smoothing of maps, identification of hotspots, assessment of uncertainty by means of the kriging variance, and the comprehension of the spatially structure of the RV by means of the semivariogram) helped us to demonstrate the potential of geostatistical tools to be applied to spatial problems, other than interpolation. Furthermore, this investigation enabled us to observe that such complexity of kriging methods requires more attention from the analyst, and not appropriate choices of spatial structure of the data (isotropy or anisotropy), data transformation, semivariogram and model parameters, kriging estimator, interpolation grid, etc. can lead to model bias, therefore skewing prediction results. This means that, cross validation results are directly influenced by the fitting parameters defined for the theoretical semivariograms.

In this study, we used a package at which theoretical semivariograms were manually fitted by trial-and-error of models, and the best ones were selected by the visual appearance of the experimental semivariograms. Therefore, in spite of our efforts to fit the optimal semivariograms, it does not necessarily mean that those were the best-fitted ones. As outlined in the literature review, alternatively, there are tools at which process are automated, and automatic fit theoretical semivariograms to experimental ones, e.g. Maximum likelihood methods and weighted least squares (Cressie, 1985, 1993; Wackernagel, 2003). However, it is also worth mentioning that an automatic fit does not necessarily will provide the optimal semivariogram, neither that they are more advantageous in relation to manual fit. Both approaches are useful under different aspects. Particularly concerning the manual fit, if on the one hand it requires more attention and experience from the analyst concerning the spatial

structure of the data, on the other hand, it enables a better understanding of the data and its spatial distribution.

In contrast to KED, GWR is advantageous as it provides a set of local parameters, i.e. coefficient estimates, standard errors and pseudo t-values, which can be mapped in the geographic space to represent the non-stationarity of the data. Taking this into account, GWR outcomes enabled us a better understanding of the nature of varying relationships between variables across space. Particularly for the modelling process, a more detailed review of the linkages, advantages and disadvantages with both approaches can be found in Harris, Brunsdon and Fotheringham (2011). In Table 6.7, we summarize some key points highlighted in this investigation that meet the ones found in that study.

In particular, the following findings are noteworthy from this investigation:

(1) Geostatistical tools enable the estimation of a value at any point in the space. Therefore, in case of data unavailability, it can be more advantageous in relation to other spatial tools, and an alternative approach to GWR to estimate values at a specific road or site. In addition to this, resulting surface maps could help identify regional hotspots or local sites where casualty incidences are higher, for instance, at a specific highway.

(2) GWR is a more appropriate spatial technique considering both crash prediction and impact analysis simultaneously. GWR provides a set of local parameters, which enables a full understanding of the nature of varying relationships between variables across space. In addition to hot zones, it can help by identifying the most influential factors and direction of their effects. Especially for long-term safety planning, this can help implement and enforce appropriate safety countermeasures.

Table 6.7 - Comparison between GWR and KED

	GWR	KED
Secondary information	▪ Possibility of more than one variable to be entered in the model ▪ Explanatory variables are accounted for in the same process	▪ Limited to one secondary variable (which is expected to be highly correlated with the primary one) ▪ Particularly in this study, we used a two-step method for the generation and modeling of the secondary variable.
Model calibration	▪ Kernel function (Directly weights data)	▪ Semivariogram function
Irregular unit areas	-	▪ Assumption of equal areas (homogeneous geometric supports)
Modelling process	▪ Friendly interface In GWR spatial patterns are modelled through variations in regression coefficients, thus using more parameters to describe the mean structure than KED does (Harris et al., 2011)	▪ Complex As a more complex model, KED requires more attention, especially for the graphical parameters
Modelling outcomes	▪ Local coefficient estimates, standard errors, pseudo t-values and estimated values	▪ Estimated values and variances of estimation
Maps of estimates	▪ Associated to the pre-defined geographic coordinate	▪ Continuous surface

Therefore, answering the last RQ of this thesis, we conclude that both GWR and KED have their merits and are suitable to explore the spatial dependence of crash data. However, the choice of one or another approach implies in gain or losses from one or another aspect (e.g. toward prediction or impact analysis). This could be solved if KED and GWR were combined, for example. One suggestion could be starting the analyses with KED (e.g. by estimating values at any point in the space, having the BLUP aspect) and extending to GWR.

7 CONCLUSIONS AND RECOMMENDATIONS

This research aimed to assess the potential improvements of supplementary data on spatial model performance and the suitability of different spatial modelling approaches on crash prediction. In this investigation, we set two specific objectives and five Research Questions (RQ). Responses to these questions were given in the corresponding chapters and are summarized herein. Subsequently, we present our conclusions and suggestions for future research.

The first set of analysis, in which RQ1 was addressed, revealed that a more diverse and comprehensive dataset led to reductions of corrected Akaike Information Criterion (AICc) and Mean Squared Prediction Error (MSPE) by 20% and 25% for motorized transport, and by 25% and 35% for active transport, respectively. In addition, this trade-off led to a set of reliable coefficient estimates, and minimized possible problems due to the omission of important variables. Considering the practical aspects, these coefficients would have important implications for long-term strategies (e.g. identifying, designing and executing appropriate road safety countermeasures). In particular, the following possibilities are noteworthy:

- identifying local influential factors and their impacts, in terms of effect size and direction. This could be helpful by suggesting changes in traffic regulations and signalization, e.g. by altering the speed limits, intensifying local speed enforcement, improving junctions, implementing crosswalks to make it safer for pedestrians, implementing signalized intersections, installing traffic management and control systems, etc.;

- identifying hotspots together with their major influencing factors, and extend the macro to microscopic-level analysis. This could possibly help road safety planners better define and pursue data collection. For instance, in a region in the study area in which high values of casualties were estimated, intersection density has found to be a great contributing factor. Hence, further local investigations could include information

of signalized intersections, existence or nonexistence of appropriate crossing lanes, to name a few. Alternatively, at Traffic Analysis Zones (TAZs) where casualties were found to have a positive association with the number of children attending a school in that TAZ, micro-level analysis could suggest changes in the speed limits or specific signalization around the school environment.

Subsequently, we carried out a sensitivity analysis. This enabled us to identify the statistical contribution of each variable in the prediction models and, therefore, find an answer to the RQ2. Especially in countries where data availability has been an issue (e.g. due to the lack of financial resources or because of other imposed conditions), this practice could be a useful to prioritize data collection strategies. Furthermore, results of this investigation could serve as a foundation for future investigation (e.g. by investigating the interaction between variables in the identified hot zones, both within and outside the models).

Together, results for RQ1 and RQ2 has further strengthened our confidence that the quality and attributes of the data play a significant role in the tradeoff between input information and modelling outcomes. Especially for the impact analysis, considerable attention must be paid to aspects involving multicollinearity, measurement error, omitted variables, etc. Other aspects, such as the statistical contribution of variables in the models, , parsimony between variables (i.e. models that are not overfitted and provide the best model performance), their implications to one or more travel modes, etc., could also help policy makers prioritize data collection. Variables included in the Flemish models are just an example of a variety of other potential information that could be used to develop more efficient models under both, predictive and explorative aspects. Nonetheless, these variables could be interesting suggestions for extra data collection in Brazil, alongside any other interesting variable.

Further investigations based on the GWR modeling attributes, enabled us to verify the accuracy of the Flemish models in casualty estimation at unsampled subzones and their performance over other interpolation methods, i.e. Mean Imputation (MEI), k-nearest neighbor (KNN) and Kriging with External Drift (KED). To this end, we applied the repeated holdout method in the Flemish models, introducing two GWR validation approaches. While, GWR holdout1 was based on the local coefficient estimates derived from the neighboring

subzones and measures of the explanatory variables for the validation subzones, GWR holdout2 used the casualty estimates of the neighboring subzones directly to estimate outcomes for the missing subzones. This original scheme would enable future studies to adjust the local features of GWR within the concept of model estimation and validation datasets and, therefore, exploring other advantages of such powerful tool. Besides confirming the suitability of the GWPR models (queried in RQ3), results of this investigation accentuated the suitability of GWR technique as an advantageous predictive tool for unsampled areas (i.e. zones with missing information) (RQ4). Especially in countries where data availability is an issue, such a GWR validation framework would allow casualties or crash frequencies to be estimated while effectively capturing the spatial variation of the data. Furthermore, results with GWR holdout2 corroborated the assumption of Tobler (1970), which states that "everything is related to everything else, but near things are more related than distant things" (i.e. first law of geography). This suggests that directly using casualty estimates of the neighboring subzones to estimate outcomes for the missing subzones is less sensitive to model inaccuracy compared to any other tested approach in this thesis.

The last step of this research was the application of KED to estimate casualties. Our intention was to present geostatistics as an alternative approach to crash prediction analysis with GWR. Results of this investigation enabled us to answer the last research question - RQ5, and confirm the suitability of KED in crash prediction. Additionally, our study provided further evidence for selecting GWR over KED when impact analysis is a criterion. Particularly for safety-planning purposes, it is essential to understand the spatially varying relationship between the input and output variables, and that was not possible with KED. In contrast, the ability of KED to estimate values at any point in the space, suggests that this tool could be an interesting alternative to GWR when data availability is an issue. Especially when we do not have explanatory variables (e.g. when only the values of crashes are known), the univariate methods of kriging could be an alternative for crash prediction. In view of these results, some consideration are noteworthy, and could help planners and researchers choose between these available tools as follows:

Compared with KED, GWR:

- is more appropriate considering both prediction and impact analysis;
- provides a set of local parameters for each variable included in the models;
- is more suitable for data within different spatial unit of analysis;
- is less computationally demanding (considering the software used) and, therefore, a less time consuming approach; and

In comparison to GWR, KED:

- enables the estimation of a value at any point in the space;
- provides the Best Linear Unbiased Predictor (BLUP), meaning estimates with minimum error and variance;
- provides continuous surface maps; and
- on the one hand is more complex/time consuming and, therefore, requires more attention for the graphical parameters. On the word hand, it enables the analyst to better understand the data and their spatial distribution.

This investigation has led us to conclude that KED could be a potential alternative to GWR, for prediction purposes, when data availability is an issue. Furthermore, KED could complement GWR prediction and impact analysis by estimating the missing values for a variable or subzone. This would enable planners and researchers to explore the potential and intrinsic characteristics of both tools and, therefore, overcoming the gaps found in one or another.

7.1 CONCLUSIONS

Although the awareness about the importance of spatial crash prediction modeling exists, the implementation and follow-up of road safety strategies has been hampered by the lack of essential information. Especially in developing countries where data availability has been an issue, this has discouraged researchers and policy makers, as they often find themselves in situations where they have to choose between doing nothing or restricting models to the

existing data. In particular, this drawback has suppressed the development of potential studies that could contribute to national goals that are related to road safety.

Our research underlined the importance of appropriate data for crash prediction. Results of a benchmarking provided us evidences to conclude that in spite of the merits of socio-economic and demographic variables, attention must be paid when models are developed for safety-planning purposes. In our case, this information fails to provide a full and effective understanding of the crash phenomena. In addition to the modelling inaccuracy, and thus unreliable predictions, those inappropriate models are prone for issues such as multicollinearity, omitted variable bias and endogeneity. Furthermore, most of the available variables in the Brazilian database are not appropriate for exploratory purposes. Hence, models are limited in their scope to identify, implement and enforce appropriate countermeasures.

Another important implication of this research is the introduction of the two GWR validation approaches (GWR holdout1 and GWR holdout1). These novel concepts can be interesting solutions to optimize GWR analyses and make use of the existing modeling framework to estimate crash frequency of the zones with missing or even without any information. Findings with this investigation has further strengthened our conviction that in addition to the data quality, the choice and soundness of the spatial tools, play an important role in spatial data analysis process and model performance.

Our research has led us to conclude that in order to improve the road safety evaluation process, efforts must be made in effective strategies. Following the example of developed countries, this could start by facilitating and encouraging the development of potential research by means of some straightforward changes, such as those identified along this research:

- promoting room for cooperative work between authorities and academia;
- collecting a broader range of potential data for both prediction and exploratory analysis;
- making the information available through public channels and to the academia;

- creating solid data basis;
- addressing the lack of data on exposure;
- prioritizing and suggesting data collection strategies based on their significance in improving prediction model performance;
- expanding the spatial methods used and opting for the most suitable one;
- harmonizing and integrating the different data sources and moreover, solving problems with underreporting crash data.

We believe that there should be no room neither for thousands of other unsuccessful road safety programs and campaigns, nor to keep “sugar-coating things” (expression commonly used in Brazil). In other words, hundreds of thousands of people will be injured or become victims of traffic crashes in Brazil, if proper safety countermeasures and investments are not implemented and realized. Finally, we hope that our research will be valuable in notifying the authorities and other stakeholders in developing counties about the importance of collecting and making appropriate data available, as well as processing this information within pertinent tools. Furthermore, our results are encouraging and make us believe that further work could contribute to enhance the quality of future analysis in road safety.

7.2 SUGGESTIONS FOR FUTURE RESEARCH

During the development of this research, some gaps and challenges within methodological aspects were identified, creating room for future studies. Hence, for future investigations, we suggest the following considerations:

- investigating the impact caused by the omission of variables, e.g. endogeneity and omitted variable bias. In this respect, one possible investigation could be micro-level analysis in the identified hot zones so as to assess the model performance by adding and removing variables into the models;
- problems involving spatial change of support for regional data taking into account that TAZs are irregular and geostatistics assumes homogeneity of supports. Techniques as

the semivariogram deconvolution (see Goovaerts, 2008) may be of interest to the application of geostatistics in crash modeling with aggregated data;

- considering the automatic fit of the theoretical semivariograms by means of maximum likelihood or weighted least squares methods;
- investigating other geostatistical tools that incorporate secondary information;
- extending the sensitivity analysis to other regions;
- investigating financial aspects, for instance identifying costs with data collection, indicating those which are more cost efficient and, therefore, providing the pros and cons while accounting for their costs;
- in line with results found in Chapter 6, we recommend future studies to explore the combination of GWR and KED on casualty prediction and, therefore, making the most of both approaches together.

REFERENCES

Aarts, L., & van Schagen, I. (2006). Driving speed and the risk of road crashes: a review. *Accident Analysis & Prevention*, 38 (2), 215-224. doi: 10.1016/j.aap.2005.07.004

AASHTO. (2010). *Highway Safety Manual (HSM).* (American Association of State Highway and Transportation Officials, Ed.) (1ed.). Washington, DC.

Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal injury crashes in pennsylvania. *Accident Analysis & Prevention*, 38 (3), 618-625. doi: 10.1016/j.aap.2005.12.006

Ahmed, M., Abdel-Aty, M., & Yu, R. (2012). Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data. *Transportation Research Record: Journal of the Transportation Research Board,* 2280, 51–59. doi: 10.3141/2280-06

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International symposium on information theory*, Tsahkadsor, Armenian SSR. Budapest: Akademiai Kiado, 267–281.

Akgüngör, A. P., & Yıldız, O. (2006). Sensitivity analysis of an accident prediction model by the fractional factorial method. *Accident Analysis & Prevention,* 39 (1), 63-68. doi: 10.1016/j.aap.2006.06.013

Allluri, P., & Ogle, J. (2012). Effects of state-specific SPFs, AADT estimations, and overdispersion parameters on crash predictions using SafetyAnalyst. In *91st Annual Meeting of the Transportation Research Board, Washington, DC.*

AMBEV. (2017). *Retrato da Segurança Viária*. Retrieved from https://www.ambev.com.br/conteudo/uploads

Amundsen, F. H., & Ranes, G. (2000). Studies on traffic accidents in Norwegian road tunnels. *Tunnelling and Underground Space Technology*, 15 (1), 3-11. doi.org/10.1016/S0886-7798(00)00024-9

Andrey, J., & Knapper, C. K. (2003).*Weather and Transportation in Canada*. Department of Geography Publication Series, no. 55. ISBN: 0-921083-65-3.

Armstrong, M. (1989). *Basic linear Geostatistics*. New York: Springer-Verlag Berlin Heidelberg.

Bédard, M., Guyatt, G.H., Stones, M. J., & Hirdes, J. P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention,* 34 (6), 717-727. doi: 10.1016/S0001-4575(01)00072-0

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71 (2), 353–360. doi: 10.1093/biomet/71.2.353

Brijs, T., Karlis, D., & Wets, G.(2008). Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis & Prevention,* 40 (3), 1180 – 1190. doi: 10.1016/j.aap.2008.01.001

Brookhuis, K. A., De Waard, D., Marchau, V. A. W. J., Wiethoff, M., Walta, L., & Bekiaris, E. (2006). *Self-explaining and forgiving roads to improve traffic safety*. In: De Waard D, Brookhuis KA, Tofetti A (eds) Developments in human factors in transportation, design, and evaluation. Shaker Publishing, Maastricht, 51–63.

Caliendo, C., Guida, M. & Parisi, A. (2007). A crash-prediction model for multilane roads. *Accident Analysis & Prevention*, 39 (4), 657-670. doi: 10.1016/j.aap.2006.10.012

Carroll, P. S. (1971). *Techniques for the use of driving exposure information in highway safety research*. HSRI, University of Michigan.

Chapman, R. (1973). The concept of exposure. *Accident Analysis & Prevention*, 5 (2), 95–110. doi: 10.1016/0001-4575(73)90018-3

Ciuffo, B. F., Punzo, V., & Quaglietta, E. (2011). Kriging meta-modelling to verify traffic micro-simulation calibration methods. In *90th Annual Meeting of Transportation Research Board, Washington, DC.*

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74 (368), 829–836. doi: 10.2307/2286407

Coburn, T. (2012). Geostatistics for Natural Resources Evaluation. *Technometrics*, 42, 437-438. doi: 10.1080/00401706.2000.10485733

Cressie, N. A. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17 (5), 563–586.

Cressie, N. A. (1993). *Statistics for spatial data*. John Wiley & Sons.

Cressie, N.A. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems*, 3:159–180.

DATASUS. (2018). Departamento De Informática do SUS. Sistema de Informações sobre Mortalidade - Estatísticas Vitais. Retrieved April, 2018, from http://www2.datasus.gov.br

de Guevara, F. L., Washington, S.P., & Oh, J. (2004). Forecasting crashes at the planninglevel: simultaneous negative binomial crash model applied in Tucson, Arizona. *Transportation Research Record: Journal of the Transportation Research Board,* 1897 (1), 191–199. doi: 10.3141/1897-25

Delen, D., Sharada, R. & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention,* 38 (3), 434-444. doi: 10.1016/j.aap.2005.06.024

DENATRAN. (2018). Vehicle fleet. Departamento Nacional de transito. Retrieved from http://www.denatran.gov.br/estatistica/237-frota-veiculos.

Doherty, S. T., Andrey, J. C., & MacGregor, C. (1998). The situational risks of young drivers: the influence of passenger, time of day and day of week on accident rates. *Accident Analysis & Prevention,* 30 (1), 45–52. doi: 10.1016/S0001-4575(97)00060-2

Duarte, F., Calvo, M. V., Borges, A., & Scatoni, I. B. (2015). Geostatistics applied to the study of the spatial distribution of insects and its use in integrated pest management. *Revista Agronomica del Noroeste Argentino*, 35 (2), 9-20.

Elvik, R. (2007). *State of the art approaches to road accident black spot management and safety analysis of road networks*. Institute of Transport Economics Norwegian Centre for Transport Research – rapport 883, Oslo.

Findley, D., Zegeer, C., Sundstrom, C., Hummer, J., & Rasdorf, W. (2012). Applying the Highway Safety Manual to two lane road curves. *Journal of the Transportation Research Forum,* 51 (3), 25-38.

Fotheringham, A. S., Brunsdon, C. & Charlton, M.E. (2002). G*eographically Weighted Regression: the analysis of spatially varying relationship*. New York: Wiley.

Fristrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., & Thomsen, L. K. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis & Prevention,* 27 (1), 1–20. doi: 10.1016/0001-4575(94)E0023-E

Golob, T. F., & Recker,W. W. (2003). Relationship among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of Transportation Engineering*, 129, 342–353. doi: 10.1061/(ASCE)0733-947X(2003)!129:4(342)

Gomes, V. A., Pitombo, C. S., Rocha, S. S., & Salgueiro, A. R. (2016). Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting. *Open Journal of Statistics*, 6 (03), 514-527. doi: 10.4236/ojs.2016.63044

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. New York: Oxford University Press.

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228, 113-129. doi: 10.1016/S0022-1694(00)00144-X

Goovaerts, P. (2005). Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *International Journal of Health Geographics*, 4 (1), 31. doi: 10.1186/1476-072X-4-31

Goovaerts, P. (2006). Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *International Journal of Health Geographics*, 5 (1), 52. doi: 10.1186/1476-072X-5-52

Goovaerts, P. (2008). Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences*, 40 (1), 101-128. doi: 10.1007/s11004-007-9129-1

Goovaerts, P. (2009). Medical geography: A promising field of application for geostatistics. *Mathematical Geosciences,* 41 (3), 243-264. doi: 10.1007/s11004-008-9211-3

Goovaerts, P., & Jacquez G. M. (2004). Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics,* 3 (1), 14. doi: 10.1186/1476-072X-3-14

Gundogdu, I. B. (2014). Risk governance for traffic accidents by Geostatistical Analyst methods. *International Journal of Research in Engineering and Science*, 2 (9), 35-40.

Guo, L., Ma, Z., & Zhang, L. (2008). Comparison of bandwidth selection in application of geographically weighted regression: a case study. *Canadian Journal of Forest Research,* 38 (9), 2526 – 2534. https://doi.org/10.1139/X08-091

Guven, O., & Kitanidis, P. (1988). Geostatistics in hydrology. *Eos Trans AGU*, 69 (34), 802–802, doi:10.1029/88EO01074

Hadayeghi, A., Shalaby, A. S., & Persaud, B. N. (2010). Development of planning level transportation safety tools using geographically weighted poisson regression. *Accident Analysis & Prevention,* 42 (2), 676 – 688. doi: 10.1016/j.aap.2009.10.016

Hadayeghi, A., Shalaby, A. S., & Persaud, B. N. (2003). Macrolevel accident prediction models for evaluating safety of urban transportation systems. *Transportation Research Record: Journal of the Transportation Research Board,* 1840 (1), 87– 95. doi: 10.3141/1840-10

Hao, W., Kamga, C., & Wan, D. (2016). The effect of time of day on driver's injury severity at highway-rail grade crossings in the United States. *Journal of Traffic and Transportation Engineering*, 3 (1), 37-50. doi: 10.1016/j.jtte.2015.10.006

Harris, P., Brunsdon, C., & Fotheringham, A. (2011). Links, comparisons and extensions of the geographically weighted regression model when used as a spatial predictor. *Stochastic Environmental Research and Risk Assessment*, 25 (2), 123-138. doi: 10.1007/s00477-010-0444-6

Harvey, A. C., & Durbin, J. (1986). The effects of seat belt legislation on british road casualties: a case study in structural time series modelling. *Journal of The Royal Statistical Society*, 149 (3), 187-227. doi: 10.2307/2981553

Hauer, E. (1982). Traffic conflicts and exposure. *Accident Analysis & Prevention*, 14 (5), 359–364. doi: 10.1016/0001-4575(82)90014-8

Hauer, E. (1995). On exposure and accident rate. *Traffic Engineering and Control,* 36 (3), 134–138.

Hauer, E., Ng, J. C. N., & Lovell, J. (1996). Estimation of safety at signalized intersection. *Transportation Research Board*, 1285, 42–51.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., & Abdel-Aty, M. (2016). Macro and micro models for zonal crash prediction with application in hot zones identification. *Journal of Transport Geography,* 54, 248-256. doi: 10.1016/j.jtrangeo.2016.06.012

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B (Statistical Methodology),* 60 (2), 271–293. doi:10.1111/1467-9868.00125

IBGE. (2018). Instituto Brasileiro de Geografia e Estatística. Retrieved April, 2018, from https://ww2.ibge.gov.br/estadosat/perfil.php?sigla=sp

Jalayer, M., & Zhou, H. (2013). A sensitivity analysis of crash prediction models input in the Highway Safety Manual. In *The 2013 ITE Midwestern District Meeting, Milwaukee, WI.*

Job, R. F. S., Lancelot, E. R., Gauthier, G. F., De Melo E Silva, F., Howard, E. W., Ledesma, R., … Castro Lancharro, B. (2015). *Federative Republic of Brazil - National Road Safety Capacity Review (English)*. Washington, DC: World Bank. Retrieved from http://documents.worldbank.org/curated/en/904921468232158448/pdf/Brazil-Road-Safety-review-English-official.pdf

Journel, A. G., & Huijbrebts, C. J. (1978). *Mining geostatistics*. London, UK: Academic Press.

Jovanis, P., & Chang, H. L. (1986). Modeling the relationship of accidents to miles traveled. *Transportation Research Record*, 1068, 42–51.

Karlaftis, M. G., & Golias, I. (2002). Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident Analysis & Prevention,* 34 (3), 357-365. doi: 10.1016/S0001-4575(01)00033-1

Kassteele J. Van De., & Stein, A. (2006). A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output. *Environmetrics*, 17 (4) 309–322. doi: 10.1002/env.771

Kassteele, J. Van De., & Velders, G. J. M. (2006). Uncertainty assessment of local NO2 concentrations derived from error-in-variable external drift kriging and its relationship to the 2010 air quality standard. *Atmospheric Environment*, 40 (14), 2583–2595. doi: 10.1016/j.atmosenv.2005.12.023

Kloeden, C. N., Ponte, G., & McLean A. J. (2001). *Travelling speed and the risk of crash involvement on rural roads*. Department of Transport and Regional Services Australian Transport Safety Bureau, Report no. cr 204. Retrieved from http://casr.adelaide.edu.au/ruralspeed/RURALSPEED.PDF

Kononov, J., & Janson, B. (2002). Diagnostic Methodology for the detection of safety problems at intersections.*Transportation Research Record: Journal of the Transportation Research Board*, 1784, 51-56. doi: 10.3141/1784-07

Kutner, M. H., Nachtsheim, C.J., & Neter, J. (2004). *Applied Linear Regression Models*. Irwin, Chicago: McGraw-Hill.

Langley, J., Mullin, B., Jackson, R. & Norton, R. (2000). Motorcycle engine size and risk of moderate to fatal injury from a motorcycle crash. *Accident Analysis & Prevention*, 32 (5), 659-663. doi: 10.1016/S0001-4575(99)00101-3

Lascala, E. A., Johnson, F. W., & Gruenewald, P. J. N. (2001). Neighborhood characteristics of alcohol-related pedestrian injury collisions: a geostatistical analysis. *Prevention Science*, 2 (2), 123-134.

Lee, J., Abdel-Aty, M., Choi, K., & Huang, H. (2015). Multi-level hot zone identificationfor pedestrian safety. *Accident Analysis & Prevention*, 76, 64–73. doi: 10.1016/j.aap.2015.01.006

Lee, S. Y., Carle, S. F., & Fogg, G. E. (2007). Geologic heterogeneity and a comparison of two geostatistical models: Sequential Gaussian and transition probability-based geostatistical simulation. *Advances in water resources*, 30 (9), 1914-1932. doi: 10.1016/j.advwatres.2007.03.005

Lequeux, Q. (2016). *What about the seatbelt use? Results of the seatbelt behaviour measurement 2015*. Belgian Road Safety Institute – Knowledge Centre Road Safety. Brussels, Belgium. Retrieved from https://www.vias.be/publications

Li, J., & Heap, A. D. (2008). *A review of spatial interpolation methods for environmental scientists*. Geoscience Australia, Record 2008/23. Retrieved from https://d28rz98at9flks.cloudfront.net/68229/Rec2008_023.pdf

Li, Z., Wang, W., Liu, P., Bigham, J. M., & Ragland, D. R. (2013). Using Geographically Weighted Poisson Regression for county-level crash modeling in California. *Safety Science*, 58, 89-97. doi: 10.1016/j.ssci.2013.04.005

Lindner, A., & Pitombo, C. S. (2018). A conjoint approach of spatial statistics and a traditional method for travel mode choice issues. *Journal of Geovisualization and Spatial Analysis*, 2:1 doi: 10.1007/s41651-017-0008-0

Lindner, A., Pitombo, C. S., Rocha, S. S., & Quintanilha, J. A. (2016). Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. *Geospatial Information Science*, 19 (4), 245-254. doi: 10.1080/10095020.2016.1260811

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

Lotz, C., Brookhuis, K., Bauer, A., Wiethoff, M., Marchau, V. A. W. J, & De Waard, D. (2006). IN-SAFETY—Towards 'Forgiving road environments': implementation scenarios for road design measures and ITS solutions. In *TRA, Europe 2006 Conference*.

Majumdar, A., Noland, R.B., & Ochieng, W.Y. (2004). A spatial and temporal analysis of safety-belt usage and safety-belt laws. *Accident Analysis & Prevention,* 36 (4), 551-560. doi: 10.1016/S0001-4575(03)00061-7

Manepalli, U. R. R., & Bham, G. H. (2011). Crash prediction: evaluation of empirical bayes and kriging methods. In *3rd International Conference on Road Safety and Simulation*, Indianapolis, IN.

Martin, J. L. (2002). Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention*, 34 (5), 619-629. doi: 10.1016/S0001-4575(01)00061-6

Matheron, G. (1963). Principles of Geostatistics. *Economic Geology*, 58 (8), 1246–1266. doi: 10.2113/gsecongeo.58.8.1246

Matheron, G. (1971). *The theory of regionalized variables and its applications*. *Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau*. Paris: École Nationale Supérieure dês Mines.

Matkan, A. A., Mohaymany, A. S., Mirbagheri, B. & Shahri, M. (2011). Explorative spatial analysis of traffic accidents using GWPR model for urban safety planning. In *3rd International Conference on Road Safety and Simulation*, Indianapolis, IN.

Matsumoto, P. S. S., & Flores, E. F. (2013). Aplicações de Geoestatística na saúde: acidentes de trânsito em Presidente Prudente SP. In *Anais do III Simpósio de Geoestatística Aplicada em Ciências, Botucatu* (III Simpósio de Geoestatística Aplicada em Ciências Agrárias, Botucatu).

Mazzella, A., Piras, C., & Pinna, F. (2011). Use of kriging technique to study roundabout performance. *Transportation Research Record: Journal of the Transportation Research Board*, 2241. doi: 10.3141/2241-09, 2011

Mcmillan, G. P., Hanson, T. E. & Lapham, S. C. (2007). Geographic variability in alcohol-related crashes in response to legalized Sunday packaged alcohol sales in New Mexico. *Accident Analysis & Prevention*, 39 (2), 252-257. doi: 10.1016/j.aap2006.07.012

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

Miaou, S., Song, J. J., & Mallick, B. K. (2003). Roadway traffic crash mapping: a space–time modeling approach. *Journal of Transportation and Statistics*, 6 (1), 33–57.

Mitra, S., & Washington, S. (2012). On the significance of omitted variables in intersection crash modeling. *Accident Analysis & Prevention,* 49, 439-448. doi: 10.1016/j.aap.2012.03.014

MOBIEL VLAANDEREN. (2018). Mobiliteit en openbare werken. Retrieved from http://www.mobielvlaanderen.be

Molla, M. M., Stone, M. L, & Lee, E. (2014). *Geostatistical Approach to Detect Traffic Accident Hot Spots and Clusters in North Dakota*, DP-276. North Dakota State University, Fargo: Upper Great Plains Transportation Institute.

Movig, K. L. L., Mathijssen, M.P.M., Nagel, P. H. A., Van Egmond, T., de Gier, J. J., Leufkens, H. G. M., Egberts, A.C.G. (2004). Psychoactive substance use and the risk of motor vehicle accidents. *Accident Analysis & Prevention*, 36 (4), 631-636. doi: 10.1016/S0001-4575(03)00084-8

Nilsson, G. (2004). *Traffic safety dimensions and the power model to describe the effect of speed on safety*. (Doctoral thesis. Lund institute of technology. Retrieved from http://portal.research.lu.se/ws/files/4394446/1693353.pdf

Odgen, K. W. (1996). *Safer roads: a guide to road safety engineering*. Aldershot: Avebury Technical.

OECD. (2016). *Road Safety Annual Report*. OECD Publishing, Paris. Retrieved from http://dx.doi.org/10.1787/irtad-2016-en

Openshaw, S. & Taylor, P. J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In *Statistical Applications in the Spatial Sciences* (pp. 127-144). Pion, London: N. Wrigley.

Openshaw, S. & Taylor, P. J. (1981). The modifiable areal unit problem. In *Quantitative Geography: A British View*. (pp. 60-69). Edited by N. Wrigley and R. Bennett. Routledge and Kegan Paul, London.

Openshaw, S. (1984). *The modifiable areal unit problem*. Norwich, England: GeoBooks.

Openstreetmap. (2018). Retrieved from https://www.openstreetmap.org

Orton, T. G., Pringle, M. J., & Bishop, T. F. A. (2016). A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. *Geoderma,* 262, 174-186. doi: 10.1016/j.geoderma.2015.08.013

Pearce, J. L., Rathbun, S. L., Aguilar-Villalobos, M., & Naeher, L. P. (2009). Characterizing the spatiotemporal variability of PM2.5 in Cusco, Peru using kriging with external drift. *Atmospheric Environment,* 43(12), 2060–2069. doi: 10.1016/j.atmosenv.2008.10.060

Pei, X., Wong, S. C., & Sze, N. N. (2012). The roles of exposure and speed in road safety analysis. *Accident Analysis & Prevention*, 48, 464-471. doi: 10.1016/j.aap.2012.03.005

Petridou, E., & Moustaki, M. (2000). Human factors in the causation of road traffic crashes. *European Journal of Epidemiology*, 16 (9), 819-826. doi: 10.1023/A:1007649804201

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). doi: 10.1061/(ASCE)TE.1943-5436.0000680

Pirdavani, A., Brijs, T., Bellemans, T., & Wets, G. (2013b). Spatial analysis of fatal and injury crashes in Flanders, Belgium: application of geographically weighted regression technique. In *92th Annual Meeting of Transportation Research Board, Washington, DC.*

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., & Wets, G. (2013a). Evaluating the road safety effects of a fuel cost increase measure by means of zonal crash prediction modeling. *Accident Analysis & Prevention*, 50, 186-195. doi: 10.1016/j.aap.2012.04.008

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., Wets, G. (2012). Application of different exposure measures in development of planning-level zonal crash prediction models. *Transportation Research Record: Journal of the Transportation Research Board,* 2280, 145 - 153. doi: 10.3141/2280-16

Pirdavani, A., Daniels, S., van Vlierden, K., Brijs, K., & Kochan, B. (2017). Socioeconomic and sociodemographic inequalities and their association with road traffic injuries. *Journal of Transport & Health*, 4, 152-161. doi: 10.1016/j.jth.2016.12.001

Pitombo, C. S., Salgueiro, A. R.., Costa, A. S. G., & Isler, C. A. (2015). A two-step method for mode choice estimation with socioeconomic and spatial information. *Spatial Statistics*, 11, 45-64. doi: 10.1016/j.spasta.2014.12.002

Pulugurtha, S. S., Duddu, V. R., & Kotagiri, Y. (2013). Traffic analysis zone level crash estimation models based on land use characteristics. *Accident Analysis & Prevention,* 50, 678-687. doi: 10.1016/j.aap.2012.06.016.

Qin, X., Ivan, J. N. & Ravishanker, N. (2004). Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis & Prevention,* 36 (2), 183-191. doi: 10.1016/S0001-4575(02)00148-3

Qin, X., Ivan, J. N., & Ravishanker, N. (2006). Bayesian estimation of hourly exposure functions by crash type and time of day. *Accident Analysis & Prevention*, 38 (6), 1071 – 1080. doi: 10.1016/j.aap.2006.04.012

Quddus, M. A. (2008). Modeling area-wide count outcomes with spatial autocorrelation and heterogeneity: an analysis of london crash data. *Accident Analysis & Prevention,* 40 (4), 1486-1497. doi: 10.1016/j.aap.2008.03.009

Redelmeier, D. A., & Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. *New England Journal Of Medicine*, 336 (7), 453-458. doi: 10.1056/NEJM199702133360701

Rengarasu, T. M., Hagiwara, T., & Hirasawa, M. (2007). Effects of road geometry and season on head-on and single-vehicle collisions on rural two lane roads in hokkaido, Japan. *Journal of the Eastern Asia Society for Transportation Studies,* 7, 2860-2872.

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

Richter, M., Pape, H. C., Otte, D., & Krettek, C. (2005). Improvements in passive car safety led to decreased injury severity – a comparison between the 1970s and 1990s. *International Journal of Care of the Injured,* 36 (4), 484-488. doi: 10.1016/j.injury.2004.10.001

Robertson, L. S. (1996). Reducing death on the road: the effects of minimum safety standards, publicized crash tests, seat belts, and alcohol. *American Journal of Public Health,* 86 (1), 31- doi: 34. 10.2105/AJPH.86.1.31.

Saha, D., Alluri, P., & Gan, A. (2015). Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. *Accident Analysis & Prevention*, 79, 133–144. doi: 10.1016/j.aap.2015.03.011

Shankar, V., Mannering, F. & Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention,* 27 (3), 371-389. doi: 10.1016/0001-4575(94)00078-Z

Shariat-Mohaymany, A., Shahri, M., Mirbagheri, B., & Matkan, A. A. (2015). Exploringspatial non-stationarity and varying relationships between crash data andrelated factors using Geographically Weighted Poisson Regression. *Transportation in GIS,* 19(2), 321–337.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Tamayo-Mas, E., Mustapha, H., & Dimitrakopoulos, R. (2016). Testing geological heterogeneity representations for enhanced oil recovery techniques. *Journal of Petroleum Science and Engineering*, 146, 222-240. doi: 10.1016/j.petrol.2016.04.027

Taylor, M. C., Lynam, D. A., Baruya, A., (2000). *The effects of drivers' speed on the frequency of road accidents*. Transport Research Laboratory Report 421.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–40. doi: 10.2307/143141

Tukey, J.W. (1977). *Exploratory Data Analysis*. Massachusetts: Addison Wesley.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., & Barbone, F. (2002). Risk factors for fatal road traffic accidents in Udine, Italy. *Accident Analysis & Prevention*, 34 (1), 71-84. doi: 10.1016/S0001-4575(00)00104-4

VLAANDEREN. (2018). Flanders State of Art. Retrieved from https://www.vlaanderen.be/en

Wackernagel, H. (2003). *Multivariate Geostatistics*. 3rd ed. Berlin, Heidelberg: Springer-Verlag.

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. doi: 10.1016/j.aap.2009.04.002

Washington, S., Karlaftis, M., & Mannering, F. (2010). S*tatistical and Econometric Methods for Transportation Data Analysis*. 2nd ed. Boca Raton: Chapman & Hall.

Webster R., & Oliver M. A. (2007). *Geostatistics for environmental scientists*. 2nd ed. Chichester, UK: John Wiley & Sons.

Wheeler, D., & Calder, C. A (2007). An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems,* 9 (2), 145-166. Doi: 10.1007/s10109-006-0040-y

WHO. (2013). WHO|Global status report on road safety. Retrieved September, 2014, from http://www.who.int/violence_injury_prevention/road_safety_status/2013/en

WHO. (2015). WHO|Global Status Report on Road Safety. Retrieved September, 2017, from http://www.who.int/violence_injury_prevention/road_safety_status/2015/en

Wiethoff, M., Brookhuis, K., de Waard, D., Marchau, V., Walta, L., Wenzel, G., ... Macharis, C. (2012). A methodology for improving road safety by novel infrastructural and invehicle technology combinations. *European Transport Research Review,* 4 (2), 67-77. doi:10.1007/s12544-011-0065-2

Williamson, M., Jalayer, M., Zhou, H., & Pour-Rouholamin, M. (2015). A sensitivity analysis of crash modification factors of access management techniques in Highway Safety Manual. *Access Management Theories and Practices*, 76. doi: 10.1061/9780784413869.008

Xu, C., Liu, P., Wang, W., & Li, Z. (2012). Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention,* 47, 162 – 171. doi: 10.1016/j.aap.2012.01.020

Xu, P., Huang, H., Dong, N., & Wong, S. C. (2017). Revisiting crash spatial heterogeneity: A bayesian spatially varying coefficients approach. *Accident Analysis & Prevention,* 98, 330 – 337. doi: 10.1016/j.aap.2016.10.015

Yamamoto, J. K., & Landim, P. M, B. (2013). *Geoestatística: Conceitos e Aplicações*. São Paulo: Oficina de textos.

Yau, K. K. W. (2004). Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. *Accident Analysis & Prevention,* 36 (3), 333-340. doi: 10.1016/S0001-4575(03)00012-5

Zhang, D. & Wang, X. (2013). *Traffic volume estimation using network interpolation techniques: an application on transit ridership in NYC Subway System*. New York: Final Report, 2013.

Zhou, M., & Sisiopiku, V. (1997). Relationship between volume-to-capacity ratios and accident rates. *Transportation research record: Journal of the Transportation Research Board*, 1581, 47-52. doi: 10.3141/1581-06

Zou, H., Yue, Y., Li, Q., & Yeh, Ago. (2012). An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network. *International Journal of Geographical Information Science*, 26(4), 667– 689. doi: 10.1080/13658816.2011.609488

APPENDIX A

Table A.1 - Detailed codification for pedestrian fatalities

Code	Description
V01	Pedestrian injured in collision with pedal cycle
V02	Pedestrian injured in collision with two-or-three wheeled motor vehicle
V03	Pedestrian injured in collision with car, pick-up truck or van
V04	Pedestrian injured in collision with heavy transport vehicle or bus
V05	Pedestrian injured in collision with railway train or railway vehicle
V06	Pedestrian injured in collision with non-motor vehicle
V09	Pedestrian injured in other and unspecified transport accidents

Table A.2 - Detailed codification for cyclist fatalities

Code	Description
V10	Pedal cyclist injured in collision with pedestrian or animal
V11	Pedal cyclist injured in collision with other pedal cycle
V12	Pedal cyclist injured in collision with two-or-three wheeled motor vehicle
V13	Pedal cyclist injured in collision with car, pick-up truck or van
V14	Pedal cyclist injured in collision with heavy transport vehicle or bus
V15	Pedal cyclist injured in collision with railway train or railway vehicle
V16	Pedal cyclist injured in collision with non-motor vehicle
V17	Pedal cyclist injured in collision with fixed or stationary object
V18	Pedal cyclist injured in non-collision transport accident
V19	Pedal cyclist injured in other and unspecified transport accidents

Table A.13- Detailed codification for motorcycle rider fatalities

Code	Description
V20	Motorcycle rider injured in collision with pedestrian or animal
V21	Motorcycle rider injured in collision with pedal cycle
V22	Motorcycle rider injured in collision with two-or-three wheeled motor vehicle
V23	Motorcycle rider injured in collision with car, pick-up truck or van
V24	Motorcycle rider injured in collision with heavy transport vehicle or bus
V25	Motorcycle rider injured in collision with railway train or railway vehicle
V26	Motorcycle rider injured in collision with non-motor vehicle
V27	Motorcycle rider injured in collision with fixed or stationary object
V28	Motorcycle rider injured in non-collision transport accident
V29	Motorcycle rider injured in other and unspecified transport accidents

Table A.14 - Detailed codification for occupant of three-wheeled motor vehicle fatalities

Code	Description
V30	Occupant of three-wheeled motor vehicle injured in collision with pedestrian or animal
V31	Occupant of three-wheeled motor vehicle injured in collision with pedal cycle
V32	Occupant of three-wheeled motor vehicle injured in collision with two-or-three wheeled motor vehicle
V33	Occupant of three-wheeled motor vehicle injured in collision with car, pick-up truck or van
V34	Occupant of three-wheeled motor vehicle injured in collision with heavy transport vehicle or bus
V35	Occupant of three-wheeled motor vehicle injured in collision with railway train or railway vehicle
V36	Occupant of three-wheeled motor vehicle injured in collision with non-motor vehicle
V37	Occupant of three-wheeled motor vehicle injured in collision with fixed or stationary object
V38	Occupant of three-wheeled motor vehicle injured in noncollision transport accident
V39	Occupant of three-wheeled motor vehicle injured in other and unspecified transport accidents

Table A.5 - Detailed codification for car occupant fatalities

Code	Description
V40	Car occupant injured in collision with pedestrian or animal
V41	Car occupant injured in collision with pedal cycle
V42	Car occupant injured in collision with two-or-three wheeled motor vehicle
V43	Car occupant injured in collision with car, pick-up truck or van
V44	Car occupant injured in collision with heavy transport vehicle or bus
V45	Car occupant injured in collision with railway train or railway vehicle
V46	Car occupant injured in collision with non-motor vehicle
V47	Car occupant injured in collision with fixed or stationary object
V48	Car occupant injured in non-collision transport accident
V49	Car occupant injured in other and unspecified transport accidents

Table A.6 - Detailed codification for occupant of pick-up truck or van fatalities

Code	Description
V50	Occupant of pick-up truck or van injured in collision with pedestrian or animal
V51	Occupant of pick-up truck or van injured in collision with pedal cycle
V52	Occupant of pick-up truck or van injured in collision with two-or-three wheeled motor vehicle
V53	Occupant of pick-up truck or van injured in collision with car, pick-up truck or van
V54	Occupant of pick-up truck or van injured in collision with heavy transport vehicle or bus
V55	Occupant of pick-up truck or van injured in collision with railway train or railway vehicle
V56	Occupant of pick-up truck or van injured in collision with non-motor vehicle
V57	Occupant of pick-up truck or van injured in collision with fixed or stationary object
V58	Occupant of pick-up truck or van injured in non-collision transport accident
V59	Occupant of pick-up truck or van injured in other and unspecified transport accidents

Table A.7 - Detailed codification for occupant of heavy transport vehicle fatalities

Code	Description
V60	Occupant of heavy transport vehicle injured in collision with pedestrian or animal
V61	Occupant of heavy transport vehicle injured in collision with pedal cycle
V62	Occupant of heavy transport vehicle injured in collision with two-or-three wheeled motor vehicle
V63	Occupant of heavy transport vehicle injured in collision with car, pick-up truck or van
V64	Occupant of heavy transport vehicle injured in collision with heavy transport vehicle or bus
V65	Occupant of heavy transport vehicle injured in collision with railway train or railway vehicle
V66	Occupant of heavy transport vehicle injured in collision with non-motor vehicle
V67	Occupant of heavy transport vehicle injured in collision with fixed or stationary object
V68	Occupant of heavy transport vehicle injured in non-collision transport accident
V69	Occupant of heavy transport vehicle injured in other and unspecified transport accidents

Table A.8 - Detailed codification for bus occupant fatalities

Code	Description
V70	Bus occupant injured in collision with pedestrian or animal
V71	Bus occupant injured in collision with pedal cycle
V72	Bus occupant injured in collision with two-or-three wheeled motor vehicle
V73	Bus occupant injured in collision with car, pick-up truck or van
V74	Bus occupant injured in collision with heavy transport vehicle or bus
V75	Bus occupant injured in collision with railway train or railway vehicle
V76	Bus occupant injured in collision with non-motor vehicle
V77	Bus occupant injured in collision with fixed or stationary object
V78	Bus occupant injured in non-collision transport accident
V79	Bus occupant injured in other and unspecified transport accidents

APPENDIX B

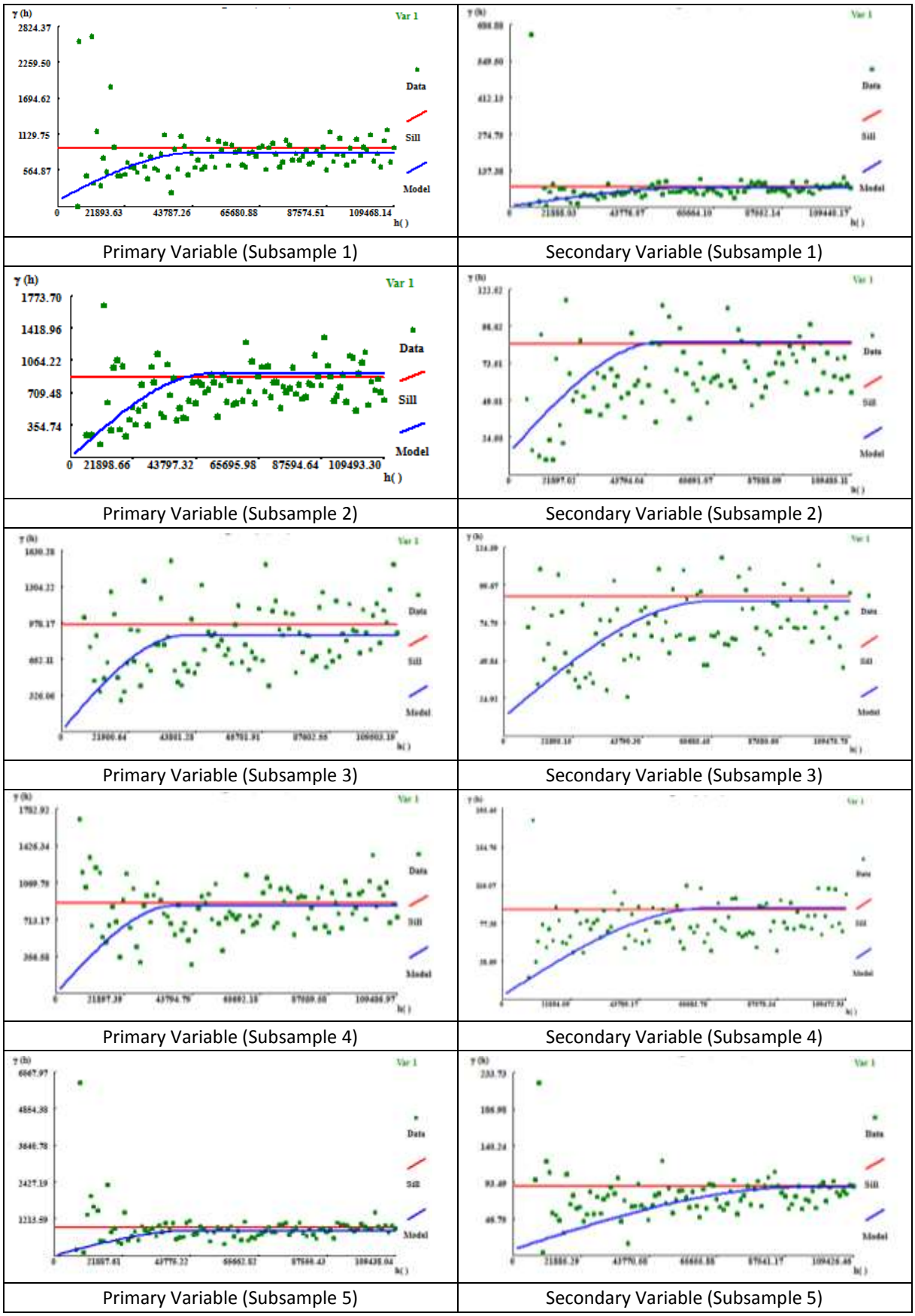
Table B.1 - Descriptive statistics of variables collected for São Paulo (including São Paulo)

	Variable	Average	Min	Max	SD[a]
Fatalities	**Active transport**	10.71	0	2179	88.656
	Motorized transport	14.95	0	1991	83.042
Network	**Link length for AT**	155.67	5.04	9523.95	425.35
	Link length for MT	168.35	5.29	10026.43	450.22
	Area	384.8	5.4	1977	319.99
Socioeconomic and demographic	**Population**	63972.4	805	11253503	454386.49
	Male population	31128.49	422	5328632	215407.9
	Female population	32843.92	383	5924871	238986.98
	Population density	302.13	3.73	12519.10	1198.31
	AAGR	1.03	-2.15	10.92	1.25
	Percentage male population	50.52	45.76	81.09	2.52
	Percentage female population	49.48	18.91	54.24	2.52
	Percentage proportion population	102.97	84.36	428.86	17.88
	Urban population	61372.48	627	11152344	45374.2
	Rural population	2599.92	0	101159	5302.31
	HDI	0.739	0.639	0.862	0.032
	GNP	22531.58	7131.54	287646.17	18420.10
	Employed people	20563.82	155	5098791	203422.19
	Occupied people	24054.99	211	5899412	235320.29
Vehicle fleet	Motorcycle	5973.61	24	797405	33069.35
	Microbus	138.98	0	31192	1252.88
	Car	20674.22	133	4617635	185230.73
	Truck	903.51	11	128606	5267.19
	Bus	196.71	3	39397	1580.75
	Total of vehicles	27887.03	220	5614235	225963.56
Fuel consumption	Gasoline (liters)	11484135.92	0	2276740223	91999934.82
	Diesel oil (liters)	17933402.11	0	1686036682	73439737.11
	Fuel oil (liters)	880143.13	0	44127640	3407289.73
	GLP (liters)	2854920.32	0	357590947	5199613.25
	Ethanol (liters)	12859838.16	0	2017823810	83035134.79

[a]SD: Standard Deviation

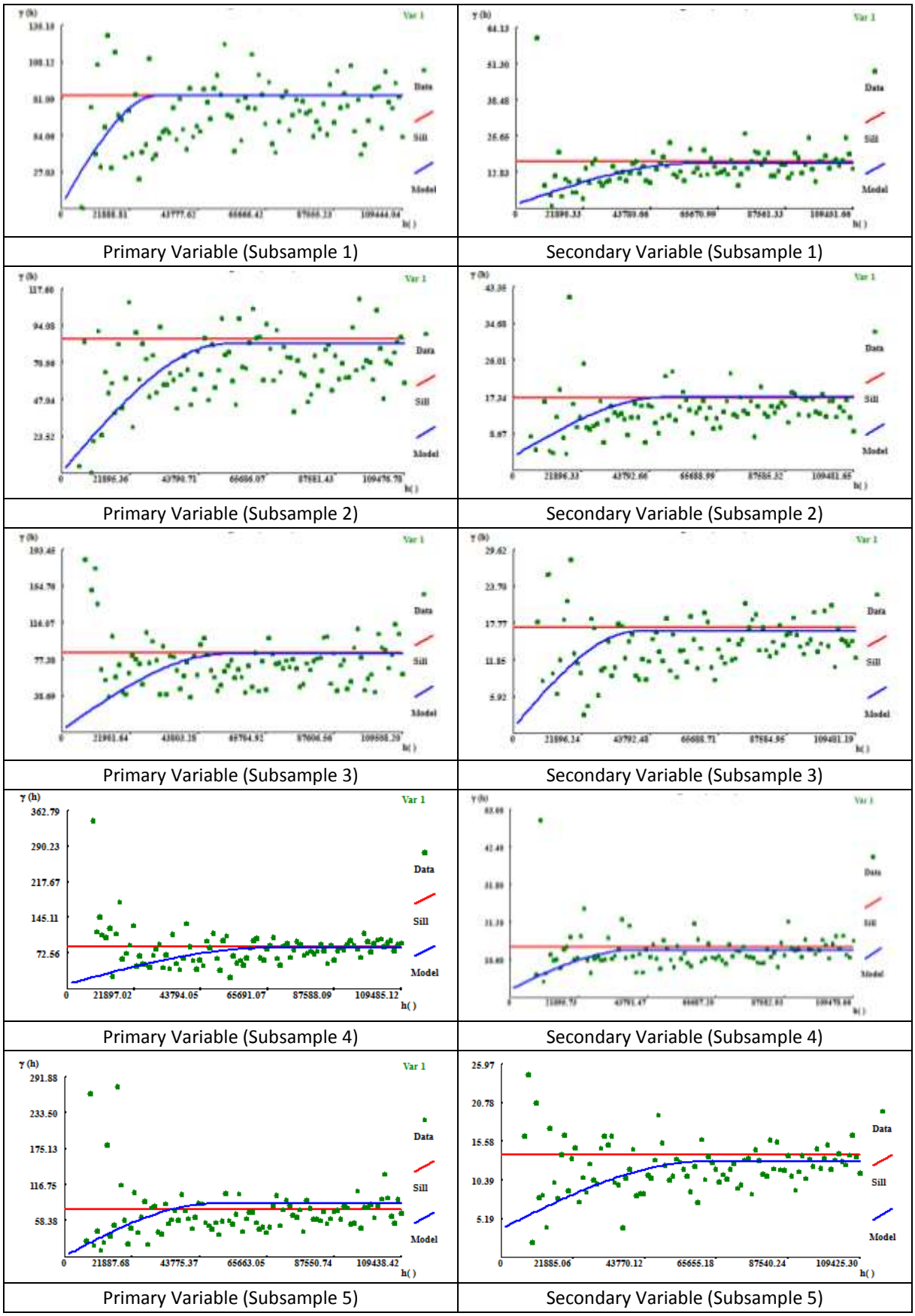
APPENDIX C

Figure C.1 - Theoretical semivariograms for motorized transport (model validation)



Primary Variable (Subsample 1)	Secondary Variable (Subsample 1)
Primary Variable (Subsample 2)	Secondary Variable (Subsample 2)
Primary Variable (Subsample 3)	Secondary Variable (Subsample 3)
Primary Variable (Subsample 4)	Secondary Variable (Subsample 4)
Primary Variable (Subsample 5)	Secondary Variable (Subsample 5)

Figure C.2 - Theoretical semivariograms for active transport (model validation)



Primary Variable (Subsample 1) | Secondary Variable (Subsample 1)

Primary Variable (Subsample 2) | Secondary Variable (Subsample 2)

Primary Variable (Subsample 3) | Secondary Variable (Subsample 3)

Primary Variable (Subsample 4) | Secondary Variable (Subsample 4)

Primary Variable (Subsample 5) | Secondary Variable (Subsample 5)