MODELLING CRASH FREQUENCY AND SEVERITY USING GPS TRAVEL DATA: EVALUATION OF SURROGATE SAFETY MEASURES AND DEVELOPMENT OF A NETWORK SCREENING MODEL

Joshua Stipancic

Department of Civil Engineering and Applied Mechanics
McGill University, Montreal

November 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Engineering.

To my girls, Amber and Adelaide.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Improving road safety requires accurate network screening methods to identify and prioritize sites to maximize effectiveness of implemented countermeasures. In the screening phase, hotspots are commonly identified using statistical models based on historical crash data. However, collision databases are subject to errors and omissions and crash-based methods are reactive. With the arrival of Global Positioning System (GPS) trajectory data, surrogate safety methods, proactive by nature, have gained popularity. Although GPS-enabled smartphones can collect reliable and spatio-temporally rich driving data from regular drivers using an inexpensive, simple, and user-friendly tool, few studies to date have analyzed large volumes of smartphone GPS data and considered surrogate-safety modelling techniques for network screening. The main objective of this thesis is to propose and validate a GPS-based network screening modeling framework dependent on surrogate safety measures (SSMs).

First, methods for collecting and processing GPS and associated data sources are presented. Data, collected in Quebec City and capturing 4000 drivers and 21,000 trips, was processed using map matching and speed filtering algorithms. Spatio-temporal congestion measures were proposed and extracted and techniques for visualizing congestion patterns at aggregate and disaggregate levels were explored. Results showed that each peak period has an onset period and dissipation period lasting one hour. Congestion in the evening is greater and more dispersed than in the morning. Congestion on motorways, arterials, and collectors is most variable during peak periods.

Second, various event-based and traffic flow SSMs are proposed and correlated with historical collision frequency and severity using Spearman's correlation coefficient and pairwise Kolmogorov-Smirnov tests, respectively. For example, hard braking (HBEs) and accelerating events (HAEs) were positively correlated with crash frequency, though correlations were much stronger at intersections than at links. Higher numbers of these vehicle manoeuvres were also related to increased collision severity. Considered traffic flow SSMs included congestion index (CI), average speed ($\bar{V}$), and coefficient of variation of speed (CVS). CI was positively correlated with crash frequency and showed a non-monotonous relationship with severity. $\bar{V}$ was negatively correlated with crash frequency and had no conclusive statistical relationship with crash severity. CVS was positively related to increased crash frequency and severity.

Third, a mixed-multivariate model was developed to predict crash frequency and severity incorporating GPS-derived SSMs as predictive variables. The outcome is estimated using two models; a crash frequency model using a Full Bayes approach and estimated using the Integrated Nested Laplace Approximation (INLA) approach and a crash severity model integrated through a fractional Multinomial Logit model. The results are combined to generate posterior expected crash frequency at each severity level and rank sites based on crash cost. Negative Binomial models outperformed alternative models based on a sample of the network, and including spatial effects showed improvement in model fit. This crash frequency model was shown to be accurate at the network scale, with the majority of proposed SSMs statistically significant at 95 % confidence. In the crash severity model, fewer variables were significant, yet the effect of all significant variables was consistent with previous results. Correlations between rankings predicted by the model and by the crash data were adequate for intersections (0.46) but were poorer for links (0.25). The inclusion of severity, which is an independent dimension of safety, is a substantial improvement over many existing studies, and the ability to prioritize sites based on GPS data and SSMs rather than historical crash data represents a substantial contribution to the field of road safety.

RÉSUMÉ

L'amélioration de la sécurité routière nécessite des méthodes de dépistage des réseaux routiers précises pour identifier et hiérarchiser les sites (segments et carrefours) afin de maximiser l'efficacité des mesures d'amélioration mises en œuvre. Dans la phase de dépistage, les points noirs sont généralement identifiés à l'aide de modèles statistiques basés sur des données historiques d'accidents. Cependant, les bases de données des accidents sont sujettes à des erreurs et des omissions, et les méthodes basées sur les accidents sont réactives. Avec l'arrivée de les données de trajectoire du Système de Positionnement Global (« Global Positioning System », GPS), les méthodes substituts de la sécurité, par nature proactives, ont gagné en popularité. Bien que les smartphones compatibles GPS puissent collecter des données sur les conducteurs fiables et riches d'un point de vue spatio-temporel avec cet outil peu coûteux, simple et convivial, peu d'études à ce jour ont analysé de grandes quantités de données GPS provenant de téléphones intelligents et considéré des modèles substituts de la sécurité pour le dépistage du réseau. L'objectif principal de cette thèse est de proposer et de valider un cadre de modélisation pour le dépistage du réseau basé sur les données GPS et les mesures substituts de la sécurité (« surrogate safety measures », SSM).

Premièrement, les méthodes de collecte et de traitement des données GPS et des autres de données associées sont présentées. Les données, recueillies à Québec et couvrant 4000 conducteurs et 21000 voyages, ont été traitées à l'aide d'algorithmes d'affectation au réseau et de filtrage des vitesses. Des mesures de congestion spatio-temporelle ont été proposées et extraites et des techniques de visualisation de la congestion à des niveaux agrégés et désagrégés ont été explorées. Les résultats ont montré que chaque période de pointe a une période d'apparition et une période de dissipation d'une heure. La congestion en soirée est plus importante et plus dispersée que le matin. La congestion sur les autoroutes, les artères et les collecteurs est plus variable pendant les périodes de pointe.

Deuxièmement, différentes SSMs basées sur les événements et des variables de trafic sont proposées et mises en corrélation avec la fréquence et la gravité des collisions en utilisant respectivement coefficient de corrélation de Spearman et le test de Kolmogorov-Smirnov pour deux échantillons. Par exemple, les événements de freinage et accélération brusques (respectivement « hard braking events », HBEs et « hard accelerating events », HAEs) sont corrélés positivement avec la fréquence des accidents, bien que les corrélations ne soient beaucoup

plus fortes aux carrefours que sur les segments. Un nombre plus élevé de véhicules est également associé à une gravité accrue des collisions. Les SSM des variables de trafic considérées comprennent l'indice de congestion (CI), la vitesse moyenne ($\bar{V}$) et le coefficient de variation de la vitesse (CVS). CI est corrélée positivement avec la fréquence des accidents et présente une relation non monotone avec la gravité. $\bar{V}$ a une corrélation négative avec la fréquence des accidents et n'a aucune relation statistique concluante avec la gravité des accidents. CVS est positivement lié à la fréquence et à la gravité accrue des accidents.

Troisièmement, un modèle multivarié mixte a été mis au point pour prédire la fréquence et la gravité des collisions en intégrant les SSM dérivées des données GPS comme variables prédictives. Deux modèles sont estimés, à savoir un modèle de la fréquence des collisions avec une approche complètement bayésienne estimé à l'aide de l'approche « Integrated Nested Laplace Approximation » (INLA) et un modèle de la gravité des collisions intégré en utilisant un modèle logit multinomial fractionnel. Les résultats sont combinés pour générer la fréquence postérieure attendue des collisions à chaque niveau de gravité et classer les sites en fonction du coût total des collisions. Les modèles binomiaux négatifs surpassent les modèles alternatifs sur un sous-ensemble du réseau et les effets spatiaux montrent une amélioration de l'ajustement du modèle. Ce modèle de la fréquence des collisions s'avèrent précis à l'échelle du réseau, la majorité des SSM proposées étant statistiquement significatifs à un niveau de confiance de 95 %. Dans le modèle de la gravité des collisions, moins de variables sont significatives, mais l'effet de toutes les variables significatives est cohérent avec les résultats antérieurs. Les corrélations entre les classements prédits par le modèle et les données de collision sont de bonne qualité pour les carrefours (0,46) mais sont plus faibles pour les segments (0,25). L'inclusion de la gravité, qui est une dimension indépendante de la sécurité, constitue une amélioration substantielle par rapport à des nombreuses études existantes et la capacité à hiérarchiser les sites en fonction des données GPS et des SSM plutôt que des données historiques d'accidents est une contribution importante au domaine de la sécurité routière.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my supervisors Luis Miranda-Moreno and Nicolas Saunier. Your guidance and input have not only improved this project but have made me a more curious student and a more competent researcher. Additionally, a special thanks to Aurelie Labbe who contributed substantially to the statistical analysis herein and proved that statisticians and engineers can work together peacefully.

In addition to my main collaborators, I must also acknowledge Spencer McNee for his assistance in data processing, the Ministère de Transport de Québec (MTQ) for providing the required crash data, and the City of Québec for collecting the GPS data using the application developed by BriskSynergies. A very special thanks to my funding sources at the Natural Sciences and Engineering Research Council (NSERC) and to Les Vadasz through the McGill Engineering Doctoral Award (MEDA).

Although there are too many to name individually, thanks to my friends and colleagues at McGill University who made coming to school every day for the past five years, if not always enjoyable, then at least bearable.

Thank you to my parents, Carol and Joe, who, from an early age, demonstrated the importance hard work and a good education, and to my sister, Kaila, for a healthy sibling rivalry that no doubt contributed to our success in our respective fields. Finally, thank you to my wife, Amber, for your love, faithfulness, and patience throughout this process.

PREFACE AND CONTRIBUTIONS OF AUTHORS

This thesis is largely the combination of five journal articles written in collaboration with three additional authors, including the two project supervisors. Dr. Luis Miranda-Moreno and Dr. Nicolas Saunier provided methodological guidance and editorial revision throughout all journal articles and the writing of this thesis. In addition, Dr. Aurelie Labbe provided assistance on three of the five journal articles listed below. Across all related projects, the author of this thesis is the sole student responsible for the preparation, analysis, and writing of this work.

Stipancic, J., L. Miranda-Moreno, A. Labbe, and N. Saunier. Measuring and Visualizing Space-Time Congestion Patterns in an Urban Road Network Using Large-Scale Smartphone-Collected GPS Data. *Transportation Letters*, published online, 2017.

Stipancic, J., L. Miranda-Moreno, and N. Saunier. Vehicle Manoeuvers as Surrogate Safety Measures: Extracting Data from the GPS-Enabled Smartphones of Regular Drivers. *Accident Analysis and Prevention*, no. 115, 2018, pp. 160-169.

Stipancic, J., L. Miranda-Moreno, and N. Saunier. Impact of Congestion and Traffic Flow on Crash Frequency and Severity: An Application of Smartphone-Collected GPS Travel Data. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2659, 2017, pp. 43-54.

Stipancic, J., L. Miranda-Moreno, N. Saunier, and A. Labbe. Surrogate Safety and Network Screening: Modelling Crash Frequency Using GPS Travel Data and Latent Gaussian Spatial Models. *Accident Analysis and Prevention*, no. 120, 2018, pp. 174-187.

Stipancic, J., L. Miranda-Moreno, N. Saunier, & A. Labbe. Network Screening for Large Urban Road Networks: Using GPS Data and Surrogate Measures to Model Crash Frequency and Severity. *Accident Analysis and Prevention*, in revision, 2018.

GLOSSARY OF TERMS

average speed the mean of all observed speeds on a single network link over a specified period of time

Bayesian techniques methods for crash modelling in which the relationships between the covariates (independent variables) and crash frequency and severity (dependent variables) are defined by probability distributions (Bayesian inference), and the probability for a hypothesis is updated as more information becomes available

behavioural techniques safety analyses which aim to identify individual driver behaviours for use as surrogate safety measures

coefficient of variation of speed the standard deviation of all observed speeds divided by the mean of all observed speeds for a single link over a specified period of time

congestion index a measure of congestion, quantified as the difference between actual and free flow speed divided by the free flow speed

crash assignment the process of locating crashes with respect to a reference coordinate system based on text-based fields provided in police reports, assigning historical crash data to the links and intersections in the road network to obtain crash counts at each link and intersection

crash-based methods techniques for safety analysis, either network screening or site diagnosis, which rely on historical crash data to develop quantitative measures of crash frequency and severity

crash frequency the number or rate of crashes occurring at a specific site within a specific period of time

crash severity the maximum severity of an injury sustained by a road user involved in a given crash, most frequently categorized as property-damage-only (no injury), minor injury, major injury, and fatal

crash modelling the process of developing statistical models to establish correlations between crash frequency/severity and covariates related to environment, driver, geometry, or surrogate safety measures and predict crash frequency and severity based on those covariates

crash precursors traffic parameters useful in identifying the potential for collisions, including variation in speed or density

decision parameter a quantitative measure which systematically combines the results from crash models in order to rank sites

evasive manoeuvre a subset of vehicle manoeuvres, actions taken by road users in an attempt to avoid collisions during near-crash events, including braking, accelerating, and steering

event-based techniques safety analyses that consider the occurrence of individual "near-crash" events, including traffic conflicts, interactions between road users, or vehicle manoeuvres, as surrogate measures of safety

floating car see *moving observer method*

free flow speed the travel speed of vehicles when unimpeded, or the average travel speed measured outside of the peak period

functional classification a categorization of roadway links based on their role within the larger road network, including freeways, arterials and collectors (further divided into primary, secondary, and tertiary) and local or residential streets

hard accelerating event a measured acceleration rate which exceeds a specific acceleration threshold

hard braking event a braking rate between consecutive GPS observations which exceeds a specific braking threshold

high-risk sites or hotspots, blackspots, and hazardous road locations, locations within the network with a potential for crash reduction

instrumented vehicle or probe vehicle, a vehicle with internal instrumentation capable of continuous spatio-temporal tracking, including GPS, accelerometer, or gyroscope, acting as moving sensors and providing traffic monitoring

integrated nested Laplace approximation an approach for Bayesian inference which uses a combination of Laplace approximations and numerical integration to estimate the posterior marginal of the latent field

jerk the rate of change of acceleration, the second derivative of speed

k-fold cross validation a method for model validation in which observations are randomly divided into k-folds before the model is estimated k times, with a different fold set aside for validation in each fold

Kolmogorov-Smirnov test a statistical test used to test equality between two probability distributions, by comparing two cumulative distribution functions and returning the maximum difference between the distributions

latent Gaussian model a subclass of structured additive models, in which the response variable (in this case, the number of crashes) for each subject is assumed to follow a distribution from the exponential family

machine learning methods methods for crash modelling in which the relationships between the covariates (independent variables) and crash frequency and severity (dependent variables) are determined using models and algorithms which independently learn from the data in order to improve estimation or prediction

map matching techniques for taking positions reported in raw GPS travel data and explicitly linking them to the road network, thereby reducing positional variability present in the raw data

moving observer method a technique for collecting traffic data requiring human observers travelling within a test vehicle or vehicles (i.e. floating cars), with average traffic flow measures derived from the number of vehicles overtaken by and overtaking the test vehicle

naturalistic driving data spatio-temporally rich driving data collected unobtrusively from regular drivers in crashes, near crashes, and under normal conditions

network screening the low-cost examination of all sites in the road network (intersections and links) to identify a smaller subgroup, often called hotspots, blackspots, hazardous road locations, or high risk sites, for detailed investigation

OpenStreetMaps an open, freely editable, and collaboratively-developed map of the world

percent deviation the proportion of observations which two ranked lists do not share in common, the relative error between modelled and observed results

probe vehicle see *instrumented vehicle*

regression models methods for crash modelling in which the relationships between the covariates (independent variables) and crash frequency and severity (dependent variables) are defined by a single value (frequentist inference)

site diagnosis detailed investigation of identified hotspots for the purpose of identifying causes and/or risk factors contributing to crash occurrence and severity in order to design and implement effective countermeasures

Spearman's rank correlation coefficient the correlation between the ordered ranks of two variables, indicates how strongly the dependency between two variables is described by a monotonic function

speed filtering the application of an algorithm, such as the Savitzky-Golay filter, to improve the signal-to-noise ratio present in the reported speeds in the raw GPS travel data

speed uniformity a quantitative measure of the uniformity between observed speed measurements within a specific time or location, see *coefficient of variation of speed*

surrogate measures of safety or surrogate safety measures, non-crash measures that are physically (causally or logically) and predictably related to traffic crashes

TrackMatching a commercially available web-based map matching service

traffic conflicts interactions between road users occurring within such a space and time that a collision is likely unless one makes an evasive manoeuvre

traffic flow techniques safety analysis methods which consider macroscopic measures of traffic flow, including combinations of volume, speed, and density, as surrogate safety measures, assuming that certain traffic conditions indicate a potential for collisions to occur

vehicle manoeuvres actions taken by drivers in crashes, near crashes, and under normal conditions

vehicle reidentification the process of matching vehicles from one point on the roadway to the next based on a reproducible vehicle feature or signature

LIST OF SYMBOLS AND ABBREVIATIONS

AADT annual average daily traffic

AVI automatic vehicle identification

BNN Bayesian neural network

BPNN back propagated neural network

BYM Besag-York-Mollič model

CDF cumulative distribution function

CI congestion index

CORR correlation coefficient

CVS coefficient of variation of speed

DIC deviance information criteria

EB empirical Bayes

FB full Bayes

FFS free flow speed

FHWA Federal Highway Administration

FMNL fractional multinomial logit model

GLONASS global navigation satellite system

GNSS global navigation satellite system

GPS global positioning system

GIS geographic information system

HAE hard accelerating event

HBE hard braking event

INLA integrated nested Laplace approximation

K-S Kolmogorov-Smirnov test

LGM latent Gaussian model

MCMC Monte Carlo Markov Chain simulation

MNL multinomial logit model

MSE mean square error

MTMDET Ministère des Transports, de la Mobilité durable et de l'Électrification des transports

MTQ Ministry of Transportation Quebec

NB negative binomial model

OSM OpenStreetMaps

PET post-encroachment time

RFID radio-frequency identification

SSM surrogate safety measure

SVM support vector machine

TTC time-to-collision

TTI travel time index

$\bar{V}$ average speed

VRI vehicle reidentification

CHAPTER 1: INTRODUCTION

"There is nothing more difficult to take in hand, more perilous to conduct, or more uncertain in its success, than to take the lead in the introduction of a new order of things."

Niccolo Machiavelli, Diplomat

1.1 Introduction

Despite the recent decline in collision rates demonstrated for Canada in Figure 1-1, road safety remains a concern due to the high economic and social costs of crashes (*1*). In 2016 alone, 1898 Canadians were killed and 160,000 were injured as a result of road traffic crashes (*2*). The process of managing and improving road safety begins with network screening. In general, screening is "the low-cost examination of all entities of a population" to identify a smaller subgroup for detailed investigation (diagnosis) (*3*). The goal of network screening is to identify sites where design or operation may "create an increased risk of unforeseeable accidents" (*4*), prioritizing them for remediation. These sites, often called hotspots, blackspots, hazardous road locations, or high-risk sites, have potential for crash reduction (*5*) and are subject to further investigation during the site diagnosis phase, where effective countermeasures are designed and implemented.

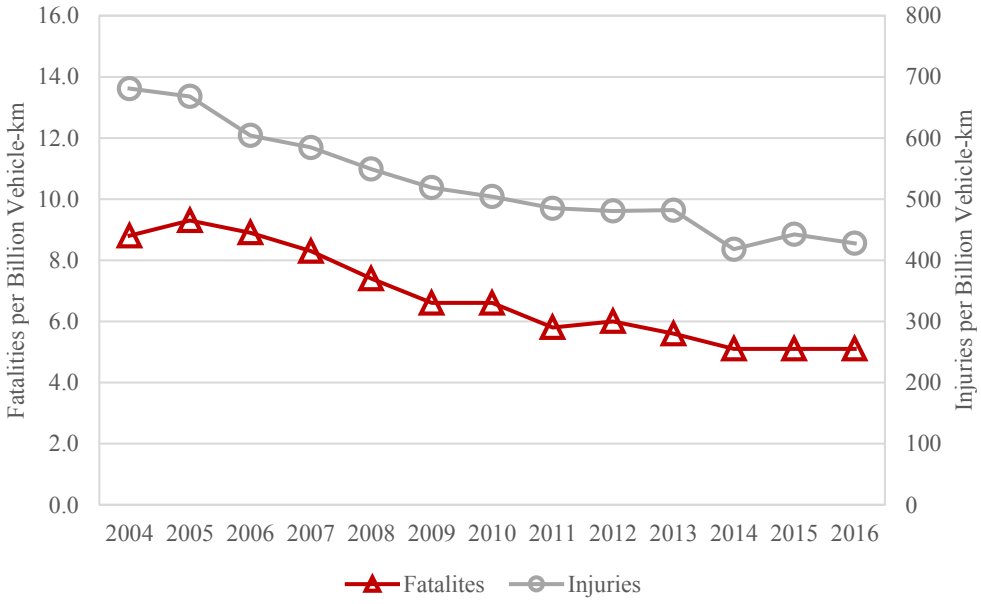


FIGURE 1-1 Canadian road traffic fatality and injury rates, 2004-2016 **(*2*)**

In both screening and diagnosis, safety must be objectively quantified, yet the selection of appropriate safety measures is often inhibited by subjective or qualitative perceptions of safety (*6*). Debate regarding preferred techniques for safety analysis remains. Traditionally, historical crash data is used to establish relationships between attributes of traffic, geometry, environment, and driver (*7*) and collision frequency or severity (*6*). Typical network screening methods use

regression to estimate the expected number of collisions based on crash, geometry, and traffic data. Though popular, crash-based methods are subject to errors and omissions in collision databases and are sensitive to crash underreporting (*8*). As traffic collisions are relatively rare events, long collection periods are required to accumulate the necessary volume of crash data for analysis (*9*). Critically, crash-based methods are reactive, requiring crashes to occur before hazardous sites are identified and improvements are made (*4*). Additionally, crashes themselves are not perfect predictors of safety (crashes occurring today may not reveal where crashes will occur tomorrow). A proactive approach to road safety should reduce dependence on crash data. Surrogate measures of safety (or surrogate safety measures, SSMs) are non-crash measures that are physically and predictably related to traffic crashes (*10*) and have been studied since the 1960s as alternatives to crash-based techniques (*11*). SSMs allow practitioners and researches to undertake proactive safety analyses before crashes occur as they are, by definition, surrogate to crash-based measures of safety. Automated safety analysis using SSMs remains in its nascence and is one of the most promising fields within safety research (*10*). Work to date has largely focussed on traffic conflict techniques using video-based sensors and computer vision techniques, a method proven powerful for site-specific diagnosis. However, it is impossible to conduct network-level safety analyses with video sensors alone, as the required number of cameras (and, in fact, other road traffic sensors) would be impractical and data processing would be resource intensive.

An alternative to crash-based screening requires an alternative data source from which crash frequency and severity can be constantly and systematically estimated across the entire road network. Probe vehicles are perhaps the only data source which currently meets these criteria and is feasible for large scale implementation. Probe vehicles, or instrumented vehicles, act "as moving sensors, continuously feeding information about traffic conditions" (*12*) through instrumentation and tracking, allowing for precise traffic monitoring from vehicles operating within normal traffic (*13*). Among the most promising methods for instrumenting vehicles, smartphones with connection to a Global Navigation Satellite System (GNSS) can collect spatio-temporally rich naturalistic driving data from regular drivers in crashes, near crashes, and under normal conditions. The term GNSS is used to refer to several satellite navigation systems, predominantly the Global Positioning System (GPS) in the United States and Russia's Global Navigation Satellite System (GLONASS). Considering the North America setting of this research, the term GPS will be used throughout, although GPS could be replaced by any other fully-functioning GNSS system.

Smartphones are an inexpensive, simple, and user-friendly data collection method that eliminate the need for external sensors. Through extracted SSMs, smartphones would allow for proactive safety analyses to be carried out continuously, rather than periodically. Furthermore, the entire process of collecting data, processing data to extract SSMs, and predicting crash frequency and severity could be implemented in a framework which automates the safety audit process.

Though instrumented vehicles could support the development of screening methods based on SSMs, there remains a lack of agreement on the type of data to collect and how to collect it, which SSMs are the best indicators of crash frequency and severity, and which techniques are most appropriate for modelling crashes based on observed SSMs. Additional shortcomings are apparent in the existing literature. Few studies have collected GPS travel data across a large urban road network using smartphones. Much of the work concerning SSMs has focussed on manual or video-based conflict analysis for site diagnosis, with little consideration for the network-wide collection of SSMs. Work on SSMs has rarely compared proposed SSMs to large amounts of historical crash data, instead comparing results to self-reported safety data or to near misses. Additionally, SSMs have rarely been validated with respect to both crash frequency and severity. With regards to crash modelling, very few studies have incorporated SSMs into statistical models of crash frequency, and none have attempted to add SSMs to models of crash severity. Despite the widespread use of Bayesian methods for screening, work to date has largely ignored the Integrated Nested Laplace Approximation (INLA) approach for applications in road safety.

1.2 Purpose and Objectives

The purpose of this thesis is to propose and validate a GPS-based network screening method dependent on SSMs. The methodology includes the development of models predicting crash frequency and severity from SSMs extracted from smartphone GPS data of regular drivers. This purpose is accomplished by addressing the following three specific objectives.

Objective 1: Propose a methodology for data processing and integration. This research aims to introduce an approach for processing large amounts of GPS travel data to generate and visualize network performance measures. This methodology involves processing raw GPS trips including the steps of map matching and speed filtering to eliminate or reduce signal noise. Techniques for processing associated data sources, including map data and crash databases, for use in GPS data

analysis are also presented. The extraction of network performance measures from GPS data is demonstrated through a visualization of network-wide congestion patterns.

Objective 2: Propose and validate SSMs in an urban road network. The second objective of this thesis is to propose several SSMs that can be extracted from GPS travel data and to validate their relationship with observed crash data across an entire network. The proposed SSMs are evaluated based on a GPS database and historical crash data. As part of this objective, several SSMs are extracted from GPS data and compared to multiple years of crash data at both the link and intersection levels at the network scale. Importantly, SSMs are evaluated with respect to both crash frequency and severity, a level of analysis absent from many of the existing surrogate safety studies. Lastly, the potential strengths and weaknesses of various GPS-based SSMs for measuring crash frequency or severity are discussed.

Objective 3: Model crash frequency and severity using a mixed-multivariate model for GPS-based network screening. This work aims at developing statistical models capable of modelling and predicting the frequency and severity of road traffic crashes. Recognizing that all existing models are based on actual crash data, the first step under this objective is to develop a statistical model of crash frequency and severity using only GPS-based SSMs, along with minimal geometric or network data, as predictive variables. In this thesis, crashes are modelled in two steps. First, a Bayesian model of crash frequency is estimated using a state-of-the-art technique for Bayesian inference (INLA) to yield crash counts at the link and intersection levels for an entire urban road network. As in the existing literature, the importance of incorporating spatial correlations in the model formulation is demonstrated. Second, a discrete choice model is used to estimate the proportion of crashes of several injury severity levels before sites are ranked and the model is validated.

1.3 General Literature Review

Though each successive chapter contains a detailed literature review pertaining to the specific subject matter within the chapter, a general literature review is provided to summarize the main topics within this thesis. Namely, these topics are instrumented vehicles and GPS data, the

development and implementation of SSMs, and techniques for crash modelling. The shortcomings present in the existing literature are also summarized.

1.3.1 Collecting Probe Vehicle Data

Collecting data using normal vehicles operating within traffic is not new. The moving observer method, or test vehicle technique, used since the 1920s requires an observer travelling in a test vehicle to record travel times between checkpoints along a route or corridor (*14*). Human observation has been replaced by the development of electronic sensors and devices, thereby improving objectivity and decreasing labour requirements. As the popularity of smartphones (which themselves contain many of the same sensors used in instrumented vehicles)has increased, so to the availability of data which can be gathered from regular drivers and the sample size which can be studied (*15*). The following sections highlight the progress from moving observers to instrumented vehicles and smartphone data collection.

Moving Observer Method: Traditionally, test vehicles and moving observers were a popular method for estimating traffic volumes, speeds, and travel times (*14*). Techniques for estimating travel time depend on the number of vehicles utilized. One approach uses the average travel time from a relatively large number of floating cars operating within the same time and space. Volume, density and mean speed are derived from the number of vehicles overtaking and overtaken by the test vehicle (*16*). Though popular, the labour requirement associated with test vehicles is high (*17*), requiring at least two individuals (one driver and one passenger) to be present (*14*). As such, approaches have been developed to use relatively fewer vehicles with statistical adjustments to extrapolate floating car travel time to mean travel time (*18*).

Instrumented Vehicles: Vehicles capable of continuous spatio-temporal tracking have become popular for measuring traffic conditions (*19*) and represent substantial improvement over test vehicles. GPS tracking provides latitude and longitude on a second-by-second basis and enables measurement of traffic parameters including travel time, speed, and delay (*14*). Initial work by D'Este, Zito, and Taylor (*20*) found that GPS was "a relatively cheap, efficient and effective means" of collecting travel data. Li and McDonald (*13*) proposed an approach using the driving pattern of a single vehicle to estimate the difference between the vehicle and average traffic

conditions. In addition to GPS positions, additional data may be collected with other sensors. The I-880 field experiment utilized four instrumented vehicles which, in addition to GPS, were instrumented with a compass, rate-of-turn indicator, and speedometer (*21*). To ensure accurate travel data is collected across a road network, a relatively high proportion of probe vehicles must be achieved (*22*). However, the need to physically instrument and operate individual probe vehicles constrains practical sample sizes.

Several methods have been proposed to increase probe sample sizes. First, the use of taxis or other fleet vehicles has been widely explored. Most large taxi fleets in the developed world are outfitted with GPS and other sensors to report "their positions, status, and trip records continuously to a central server" (*23*). Balan, Khoa, and Jiang (*23*) describe a real-time trip information system developed for a fleet of 15,000 GPS-equipped taxis in Singapore. Similarly, Moreira-Matias et al. (*24*) developed a routing algorithm for a fleet of 441 taxis in Portugal using GPS location data from the taxi fleet as an input. Despite the relatively small fleet studied in the paper, taxis made up approximately 4 % of all vehicles outside of the peak period, demonstrating the high penetration rates possible with taxi fleet data (*24*). Besides fleet vehicles, kinematic vehicle data, including position, speed, and acceleration are becoming increasingly available as automakers integrate data collection and communications technologies into smart and connected vehicles (*25*). Kluger et al. (*25*) demonstrated how SSMs could be extracted from kinematic vehicle data and supplementary information from vehicle-mounted cameras and radars. The authors successfully identified 78 % of all near-crash events based on vehicle acceleration data alone (*25*).

Smartphones and GPS: Though taxis and other instrumented vehicles present interesting opportunities for instrumentation, both are inherently biased towards specific segments of the population. GPS, accelerometer, gyroscopic, or other data collected from the smartphones of regular drivers enables the use of many probe vehicles without the high labor costs associated with traditional floating cars and may better represent average traffic conditions. While smartphone users are not truly representative, as not all drivers own smartphones or use applications that share location data, collecting data from the smartphones of regular drivers represents the least biased method currently feasible for collecting large volumes of GPS data. Data collection using smartphones is relatively inexpensive, user-friendly, and takes advantage of technology that is already widespread in the driving population. Additionally, a smartphone-based data collection

system “exploits the extensive coverage provided by the cellular network, the high accuracy in position and velocity measurements provided by GPS devices, and the existing infrastructure of the communication network” (*26*). The Mobile Century Field Experiment (*26*) used virtual trip lines to extract position and speed data from passing smartphones. The study showed that a 2-3 % penetration rate was adequate for accurately measuring average travel speed along a given link (*26*). Dunlop et al. (*15*) used smartphone GPS and accelerometer data to extract sudden braking events and identify dangerous road sections. The authors demonstrated that locations with sudden decelerations were correlated with locations that had experienced recent crashes. (*15*). In addition to safety outcomes, smartphones enable detailed transportation planning analysis, as trip distribution, route choice, trip purpose, and even travel mode can be identified without active user input (*27*).

1.3.2 Surrogate Measures of Safety

Considering that SSMs should be physically and predictably related to crashes, surrogate safety indicators must demonstrate both validity and reliability. According to Laureshyn (*28*), validity is “the property of an indicator to describe the quality that it is intended to represent”. In other words, an SSM should measure what it is supposed to measure (safety) and not some other phenomenon (*11*). Various methods have been proposed to separate events of interest from normal driving behaviour (*29*). Reliability is “the property of an indicator to be measured with the same accuracy and objectivity” independent of spatio-temporal or environmental conditions (*28*). Manual collection of SSMs, in particular using one of the traditional traffic conflict techniques, was criticized due to potential subjectivity of human observers. Although proven reliable (*28*), automating conflict analysis has further increased objectivity and ease of analysis (*30*). Methods for surrogate safety can be categorized as either event-based techniques, traffic flow techniques, or behavioural techniques, as summarized in the following paragraphs.

Event-Based Techniques: Analyses that rely on the occurrence of individual “near-crash” events are considered event-based techniques. Individual events can be identified as either traffic conflicts, interactions between road users, or vehicle manoeuvres. Severe conflicts are events that are sufficiently close to real crashes. Although initially studied using human observation, more recently the process has been automated using video data and computer vision techniques (*31*, *32*),

and conflicts have been measured using objective indicators like time-to-collision (TTC) or post-encroachment time (PET) (*10*). Though video-based sensors provide high temporal resolution (*4*) and rich positional data (*33*), the analysis of video data has spatial limitations (*4*, *34*) and potential biases in speed measurement (*35*), creating a need to implement event-based techniques using other data sources.

Rather than requiring video, other methods identify near-crash events using evasive manoeuvres captured by other sensors, also called "avoidance activities" (*4*), made by drivers in an attempt to avoid a collision. Vehicle manoeuvres of steering, braking, or accelerating have been considered and extracted from probe vehicles using various sensors (*36*). Fazeen et al. (*37*) used smartphone accelerometer data to classify 'safe' accelerations and decelerations from 'unsafe' ones, though no evidence was provided demonstrating 'unsafe' behaviour led to increased risk. Jun, Ogle, and Guensler (*38*) analyzed the relationship between spatio-temporal driving behavior activity and likelihood of crash involvement, finding that drivers involved in crashes tended to travel longer distances at higher speeds and "engaged in hard deceleration events" more frequently. Agerholm and Larhmann (*4*) stated that deceleration is the most intuitive evasive action to consider, noting that "braking was the evasive action […] in 88 % of the accidents in built-up areas" (*4*). Jerk, the rate of change of acceleration, was correlated with accident occurrence across both drivers and sites (*4*). Using GPS, accelerometer, radar, and self-reported collision data, Bagdadi (*29*) proposed a jerk-based surrogate measure that correctly identified 86 % of near misses. To date, most probe vehicle studies have relied on small sample sizes.

Traffic Flow Techniques: Rather than consider individual events, macroscopic measures of traffic flow, including combinations of volume, speed, and density, may be used as SSMs (*39*). Traffic flow techniques assume that certain traffic conditions indicate a potential for collisions to occur (*9*, *40*). Traffic conditions are typically measured using roadside point sensors such as inductive loops or radar (*40*, *41*, *42*). Oh et al. (*40*) considered the average and variance of flow, occupancy, and speed as potential indicators of traffic state, correlating the standard deviation of speed with disruptive traffic conditions and a higher likelihood of collisions (*40*). Lee et al. (*42*) used the term "crash precursors" for traffic parameters useful in identifying the potential for collisions, showing that variation in speed and density were significantly correlated with crash frequency (*42*). Abdel-Aty and Pande (*7*) classified traffic conditions as either leading to or not leading to a crash, finding

that crash prone conditions could be characterized by speed variation in congested traffic (increased speed variation increases the likelihood of a crash). Golob et al. (*41*) concluded that "the key elements … affecting safety are not only mean volume and speed, but also variations in volume and speed" (*41*). Though successful on freeways, traffic flow variables are not sufficient to predict accidents on urban streets. Furthermore, it is impractical and costly to implement roadside sensors across an entire urban network (*26*), and traffic flow measures have yet to be proven as reliable SSMs in urban networks with at-grade intersections.

Behavioural Techniques: Driver behaviour is the most prevalent cause of road crashes worldwide (*43*), and behavioural techniques aim to identify individual driver behaviours and quantify their relationship with crash likelihood (*36*). Yielding behaviour has been used frequently to evaluate the safety of pedestrians at crosswalk locations. For example, Schroeder and Rouphail (*44*) found that driver yielding behaviour was influenced by both the crosswalk treatment and the behaviour of the crossing pedestrian. Similarly, Turner et al. (*45*) found that yielding behaviour was highly depending on crossing type, but that behaviour was also influenced by site specific characteristics such as speed limit or number of lanes. In terms of driving infractions, Parker et al. (*46*) showed that driving violations, driver errors, and lapses in judgement were related to higher occurrences of certain crash types. Ulleberg and Rundmo (*47*) demonstrated the relationships between certain personality traits, including aggression and anxiety, and risky driving behaviours. Socio-demographic factors have also been linked to driving behaviour and road safety. Young drivers tend to have more driving infractions and exhibit risky driving behaviour, leading to a higher risk of traffic crashes (*48*), while working professionals are more prone to drowsy driving and are more likely to use their phone while driving (*43*).

1.3.3 Methods for Modelling Crash Frequency and Severity

Methods for crash modelling in the existing literature are varied. Regression models have been the most popular, while Empirical and Full Bayesian methods have seen increasing use. Other methods have employed machine learning approaches. Though the specific approaches vary, crash models which consider only crash frequency are the most common (*49*). Lord and Mannering (*50*) provide a detailed summary on methods for crash frequency modelling. Yet, as with Vision Zero, some have suggested that reducing injuries and fatalities, rather than overall crashes may be a more

effective means of improving safety (*51*), and frequency and severity should both be considered. Severity has been incorporated into crash risk models through various methods (*3*). The main techniques for modelling crash frequency and severity are summarized below.

Statistical Regression Models: The main approach for crash modelling has been to use statistical count models, primarily Poisson regression (*52*), estimated from environment, roadway, and traffic variables. Traditional Poisson and negative binomial (NB) or Poisson-Gamma models (used to account for overdispersion) have been the most common (*53*). Newer models, including zero-inflated models, have been used to address the overabundance of sites with zero observations (*52*). Lord, Washington, and Ivan (*53*) provide a comparison of these three primary model types, noting that although zero-inflated models may provide a good model fit, their assumptions regarding the underlying crash process may be problematic. In addition to these basic models, other models have been developed to include random effects, multivariate outcomes, or hierarchical structures (*50*).

Early regression models of injury severity include binary choice models, where the outcome is selected from two alternatives (non-injury and injury crashes). Severity models have evolved to include multiple discrete outcomes which incorporate several injury severity categories, whether unordered (multinomial and nested logit) or ordered (ordered probit and logit) (*52*). Ye and Lord (*54*) compared the most common severity models in the multinomial logit, ordered probit and mixed logit. The ordered probit showed the best goodness-of-fit, and the mixed logit was superior to the multinomial logit. Yasmin and Eluru (*55*) examined differences between several ordered and unordered frameworks, showing the most flexible of each (the mixed generalized ordered logit and the mixed multinomial logit) performed similarly in predicting severity.

Bayesian Techniques: In contrast to regression techniques, in which coefficients take fixed values, Bayesian techniques assume that the coefficients are defined by a probability distribution (*56*). In Empirical Bayes (EB) models, the probability distribution is determined, in part, by using observed historical crash data (*57*). In Full Bayes (FB) techniques, the posterior distributions are typically determined by assuming a prior distribution and iteratively computing and updating the posterior marginal using a Monte Carlo Markov Chain (MCMC) simulation. However, this process is time and resource intensive requiring tens of thousands of iterations and often hours or days to complete the estimation procedure (*58*). Studies comparing EB and FB approaches (*59*, *60*) have shown their

superiority to basic regression models. Miaou and Lord (*61*) compared EB and FB techniques, noting that EB estimates deviated significantly from the FB estimates for some data sets. FB techniques have been extended by incorporating temporal and spatial correlations (*57*). Quddus (*62*) found that a Bayesian spatial model was most appropriate for modelling crash frequency.

The simplest method for incorporating crash severity into a Bayesian framework is to estimate a two-step model. Wang, Quddus, and Ison (*63*) developed a mixed multivariate model using an FB spatial model to estimate crash counts, and an unordered nominal response model to determine the proportion by severity type. In contrast, Miaou and Song (*64*) utilized a spatial multivariate model to simultaneously estimate the number of collisions at several injury severity levels, showing that consideration of spatial effects "significantly improved the overall goodness-of-fit". Aguero-Valerde and Jovanis (*5*) argued that severity levels should not be considered independently, and that doing so "may distort estimates of variance components and regression coefficients… and result in reduced efficiency and possibly biased parameter estimates" (*5*). The authors showed that including correlation between severity levels in the model specification increased goodness-of-fit. Similarly, Park and Lord (*65*) used a multivariate Bayesian model to demonstrate correlation across severity levels and generate more precise estimates compared to univariate models.

Machine Learning Methods: In addition to more traditional crash models, machine learning techniques have also been explored. Xie, Lord, and Zhang (*66*) showed that both a back-propagated neural network (BPNN) and Bayesian neural network (BNN) had a better goodness-of-fit and prediction capabilities than a traditional NB model. Li et al. (*67*) assessed the predicting power of a Support Vector Machine (SVM), which was more accurate than an NB model. BNNs have also been used to classify crashes according to injury severity, such as in de Ona, Mujalli, Calvo (*68*), where risk factors were also studied. Chang and Wang (*51*) used a classification and regression tree, which successfully predicted non-injury and injury collisions (85-95% accuracy) but failed to predict fatal collisions. Chong, Abraham, and Paprzycki (*69*) compared several machine learning approaches including BNN, decision trees, SVM, and a hybrid decision tree/neural network approach for modelling crash severity. The authors found that their hybrid approach outperformed other methods in both training and testing (prediction). Figure 1-2 summarizes the literature review.

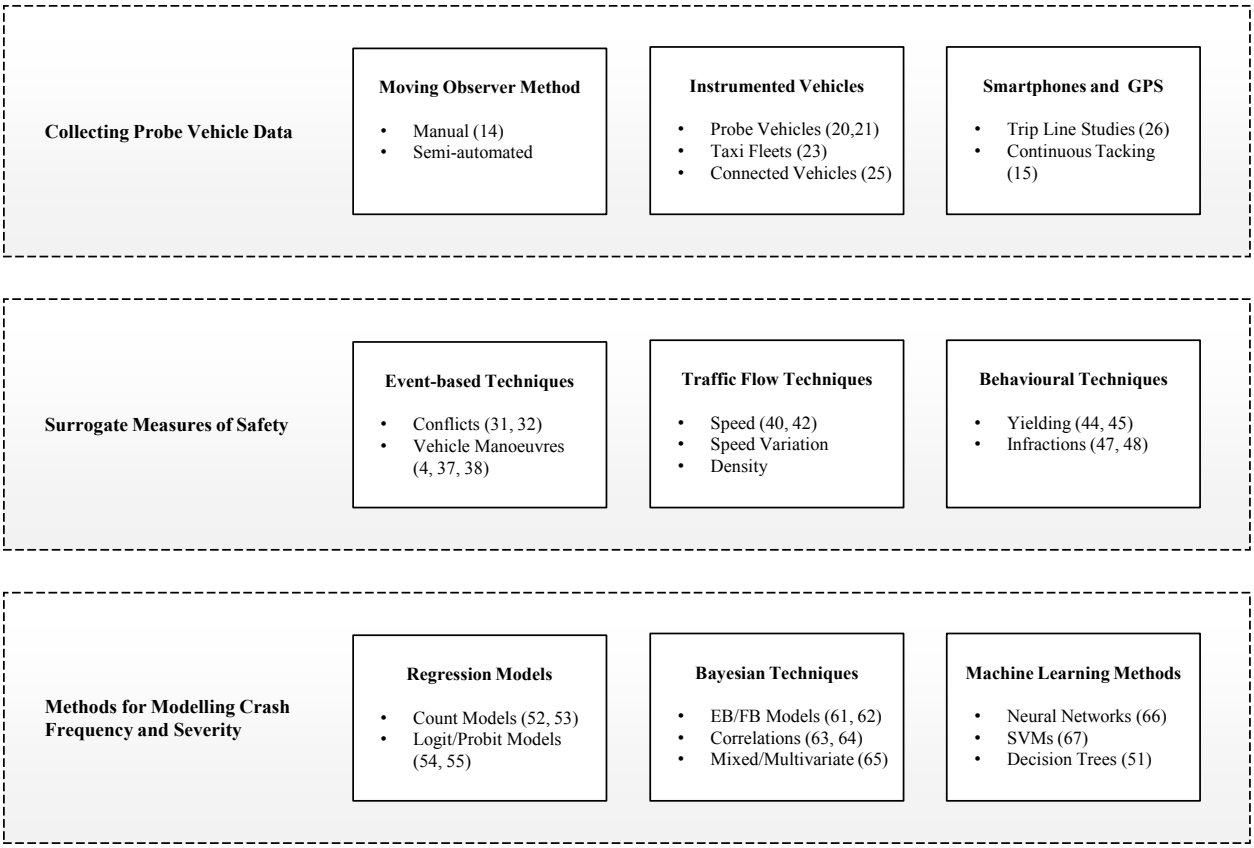


FIGURE 1-2 Summary of methods for crash modelling

1.3.4 Shortcomings

Considering the existing literature in the areas of instrumented vehicles, surrogate safety analysis, and crash risk modelling, several shortcomings become obvious. Many of these shortcomings fall at the intersection of two or three of these topics, and therefore are well suited for a comprehensive work such as this thesis. The gaps in the literature, which this thesis aims to address, are summarized below.

1) GPS and the network scale: Despite the emergence and promise of vehicles instrumented with GPS, few studies have explored the potential of collecting data across an entire urban road network. Dedicated probe vehicles are incapable of such a widespread data collection campaign, and data from vehicle fleets is inherently biased, making smartphone data collection the best option for collecting GPS travel data at the network scale. Despite the proliferation of GPS-enabled smartphones, few studies have attempted to collect network-wide data using the smartphone GPS of regular drivers alone.

2) Smartphones and SSMs: Although many methods for extracting and analyzing SSMs have been proposed, most of the work has focussed on video-based conflict analysis for site diagnosis. Though some studies have extracted event-based SSMs from instrumented vehicles, few have extracted such measures from smartphone GPS travel data. Furthermore, no studies have attempted to use instrumented vehicles (whether instrumented conventionally or using smartphones) to extract traffic flow SSMs.

3) SSMs and historical crash data: Few surrogate safety studies have compared the proposed SSMs of hard braking, congestion, speed, and speed variation to large amounts of historical crash data. Instead, results are often compared to self-reported safety data or to near misses. By definition, SSMs must be physically and predictably related to crashes. Therefore, more effort is needed to compare any proposed SSM with a reasonable amount of historical crash data to demonstrate that such a relationship exists.

4) SSMs as predictors of crash severity: Collision frequency and severity are independent dimensions of road safety. Although both frequency and severity have been studied extensively using crash-based techniques, few surrogate safety studies have explicitly considered injury severity. Though correlation between SSMs and crashes has been considered, measures may also have statistically significant relationships with injury severity. In fact, some measures may be better predictors of crash severity than of crash frequency. Therefore, the relationships SSMs share with both crash frequency and severity must be established and validated.

5) Crash modelling with SSMs: The existing literature is replete with approaches to modelling crash risk. Although more complex models continue to improve estimates of road traffic crashes, all existing models for predicting crashes are based on crashes. The best method for reducing dependence on crash data in network screening is to develop models capable of predicting crash counts and injury levels using SSMs as predictive variables. To date, few studies have incorporated SSMs into statistical models of crash frequency, and none have attempted to add SSMs to models of crash severity.

6) Advances in Bayesian inference: When it comes to estimating or predicting road traffic crashes, the most accurate and well-accepted approach is Bayesian modelling. MCMC approaches have been widely used in the safety literature to estimate these types of models. However, this approach is computationally expensive and therefore time consuming. This study takes advantage of recent advances in Bayesian inference, namely INLA, as a state-of-the-art method to solve a complex problem in the field of road safety.

1.4 Contributions

This thesis contributes to the existing state-of-practice in several key areas described in the shortcomings of existing literature and outlined in the following chapters. Some of these contributions are visible in Figure 1-3.

Data Processing and Visualization: An automated methodology for processing large quantities of GPS data is developed and presented, by combining existing and newly-developed techniques. These methods allow for network-wide analysis of GPS data, which has been rare in existing work. This thesis contributes further by demonstrating the types of data visualization which are possible through the large-scale, or network-level application of smartphone GPS data.

Extracting and Validating SSMs: Several SSMs are derived from GPS data, including traffic flow measures which have never been quantified using smartphone GPS data alone. One of the most substantial contributions of this research is the robust validation of these indicators. Not only are the proposed SSMs validated with respect to a large volume of historical crash data, not often seen in the existing literature, but techniques for validating SSMs with crash severity are also presented. The inclusion of the independent safety dimension of crash severity in the validation of SSMs is a significant improvement over existing surrogate safety studies.

Modelling Crash Frequency and Severity: An early application of Bayesian inference using the INLA technique for crash modelling represents a chief contribution to the field of road safety. Estimation using INLA is a substantial improvement in terms of estimation time compared to traditional MCMC approaches. Additionally, the crash model, which predominantly relies on measures derived from GPS data as the covariates, markedly reduces the necessity for crash data

in the network screening process and facilitates a major shift from reactive crash-based models to proactive models based on SSMs.

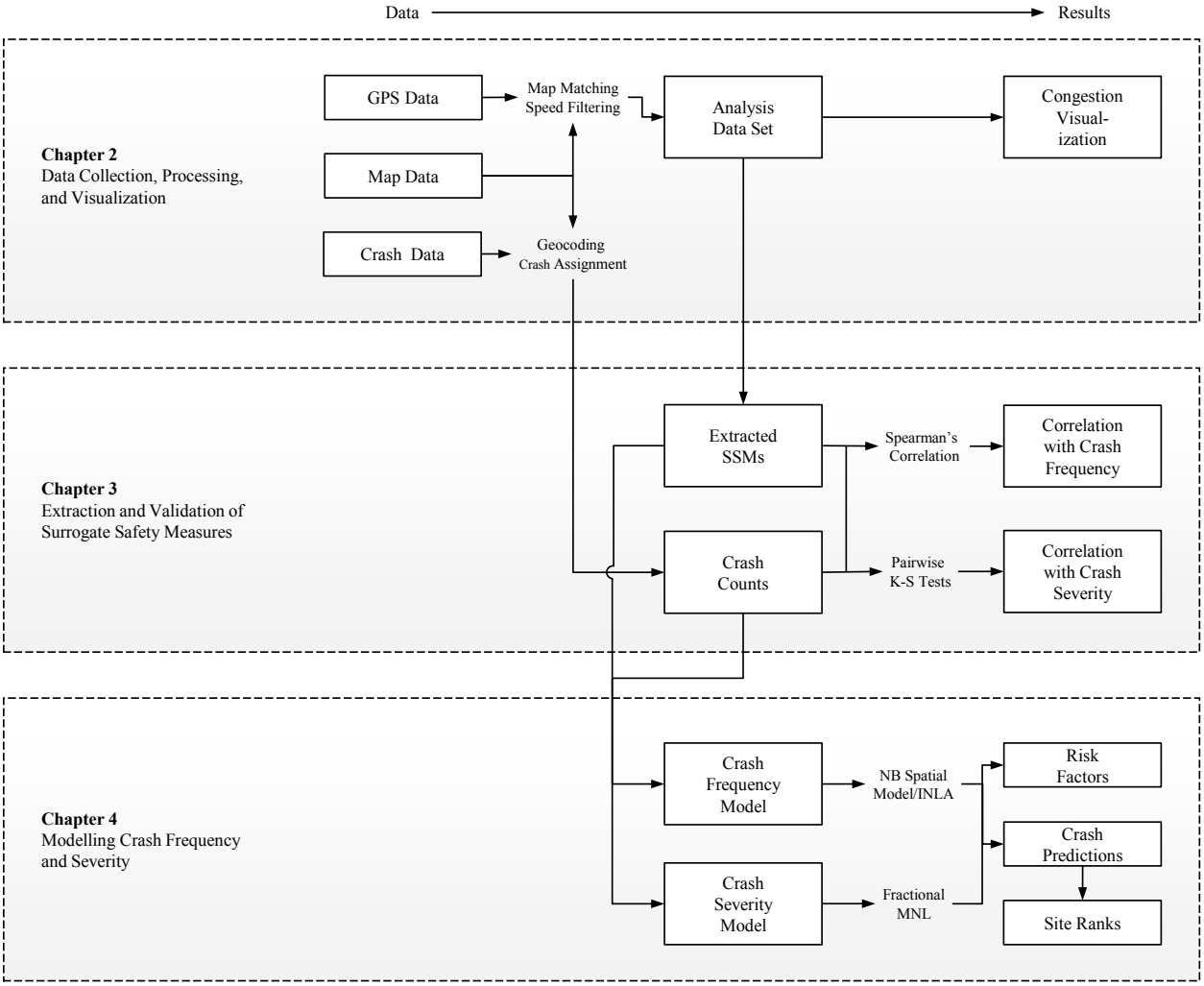


FIGURE 1-3 Thesis methodology and workflow

1.5 Organization

This thesis is organized in five chapters, including the introduction. Chapters 2, 3, and 4 are based on journal articles either published in or submitted to peer-reviewed journals. The general methodology and workflow for this thesis is outlined in Figure 1-3. Chapter 2 presents the methods for collecting and processing GPS trip data to yield meaningful traffic information, with a specific focus on measuring and visualizing spatio-temporal patterns of congestion. Congestion is a dynamic phenomenon with elements of space and time, and it is therefore a promising application

of probe vehicles instrumented with smartphones. This chapter introduces the data set, collected in Quebec City in the spring of 2014 using the Mon Trajet application (*70*). The data is processed and the congestion index is computed and visualized across time and space.

Chapter 3 focusses on the extraction and validation of SSMs. Recognizing that braking is one of the most common evasive manoeuvres in urban areas, event-based measures of HBEs and HAEs were extracted along with selected traffic flow measures of congestion, average speed, and speed variation. All extracted SSMs were compared to a large volume of historical crash data provided by the Ministry of Transportation Quebec (or MTQ, now the Ministère des Transports, de la Mobilité durable et de l'Électrification des transports, or MTMDET) from the years 2000 to 2010. Relationships with crash frequency were determined using Spearman's correlation coefficient, and relationships with crash severity were evaluated using pairwise K-S tests.

Chapter 4 presents the development of the crash model. Extracted SSMs are combined with observed trip counts and roadway functional classification and length as predictive variables. Several latent Gaussian models are presented and estimated using the INLA approach for a subset of links and intersections in the road network. These models are improved by incorporating spatial correlations before the model is scaled to the entire road network, and predicting power is validated. In Chapter 5, conclusions are summarized and the main contributions of this research are outlined. Several recommendations for future work are discussed.

CHAPTER 2: DATA COLLECTION, PROCESSING, AND VISUALIZATION

"In God we trust; all others bring data."

W. Edwards Deming, Engineer

2.1 Introduction

For transportation engineers and planners, accurate traffic data facilitates the development of quantitative performance measures necessary for operating existing networks and planning future facilities (*20*). For drivers, road network performance influences travel choices (*71*). The need for 'very accurate road traffic information' (*12*) has led to the development of new traffic sensors, including radar, magnetic, and video-based devices to supplement or replace traditional inductive loop detectors. In contrast to these fixed sensors, probe vehicles act 'as moving sensors' (*12*), and are instrumented using various techniques and sensors to provide precise traffic data for vehicles operating within normal traffic (*13*). Among the existing sensor types, GPS has been proven reliable in several applications (*38*). Whereas traditional floating car studies were limited in driver sample size and spatio-temporal coverage (*71*), GPS-enabled smartphones have the potential to increase the number of drivers sampled, increase temporal coverage to several weeks or months, and increase the spatial coverage to include the entire road network. Smartphones eliminate the need for dedicated devices by taking advantage of widespread technology and exploiting existing communication infrastructure (*26*).

GPS-equipped probe vehicles allow for the precise measurement of performance measures throughout the road network. Link or route travel time is easily calculated from GPS trip data and is beneficial for road users and practitioners alike (*13*). Although it is a key parameter defining traffic state (*72*) and is easily understood (*71*), travel time may not be the biggest factor influencing travel decisions. For example, road users are willing to accept longer travel times if they can be assured that they will usually arrive on time (*73*). Traffic congestion occurs on a roadway 'when demand … exceeds its ability to supply an acceptable level of service' (*20*) and greatly affects travel time reliability. Traffic congestion is worsening in many urban areas (*74*), where traditional expectations of peak period congestion are being replaced by congestion lasting throughout the day. Efforts to understand and reduce the 'extent, duration, and intensity' of congestion (*75*) should be a high priority (*74*). Furthermore, congestion 'is a dynamic phenomenon with elements of both space and time' (*20*) and requires temporal and spatial data coverage, making it a promising application of smartphone-equipped probe vehicles, one of the only methods currently capable of providing such data.

Although GPS probe vehicles have been successfully used in freeways (*19*), their application to urban environments requires additional consideration. Tall buildings can completely

block GPS signals or create spurious signals (*20*). Sufficient network coverage is practically limited by driver sample size, regardless of the collection method. Additionally, methods for automating data handling and analysis are required due to large data volumes obtained from smartphones. Other challenges include selecting appropriate performance measures which can be reliably extracted from GPS trip data. Despite technological advancement and increasing congestion levels, smartphone-based systems for measuring and monitoring traffic congestion are still rare in North American cities (*17*). Additionally, "literature in network level dynamics and congestion propagation is limited especially in large urban networks" (*76*). The purpose of this chapter, previously published in one journal article (*77*), is to present the main sources of data used in this thesis and to demonstrate a practical application of smartphone GPS data by measuring and visualizing the magnitude and variability of congestion across an urban road network. The three primary objectives of this chapter are to:

(1) Present and process network-wide travel data from smartphone GPS;
(2) Quantify congestion at the network scale using the Congestion Index (CI), and;
(3) Visualize changes in CI across time and space at aggregate and disaggregate scales.

2.2 Literature Review

Existing methods for estimating traffic conditions depend predominantly on the data source. Using fixed point sensors, the naïve method uses spot speeds to estimate traffic conditions (*78*). The naïve method introduces a systematic bias (*79*) by equating detector data "averaged over a fixed time period at a single point in space" to traffic conditions "averaged over a fixed distance and a variable amount of time" (*80*). In trajectory methods, trajectories of simulated vehicles are constructed based on traffic data observed by several consecutive fixed sensors (*81*). van Lint and van der Zijpp (*81*) improved traditional methods by assuming linear speed variation (rather than piecewise-constant variation), which more accurately represents flow. Coifman (*80*) constructed virtual trajectories based on several loop detectors to estimate travel time in a freeway environment, demonstrating that estimated travel times were within 10 % of actual travel times on average. Liu and Ma (*71*) fused loop data with signal phase information in urban corridors to estimate travel times generally within 5 % of ground truth. As with the naïve method, trajectory methods are limited because data is collected at discrete locations (*80*).

Vehicle reidentification (VRI) is "the process of matching vehicles from one point on the roadway … to the next" (*82*) based on a reproducible feature or vehicle signature (*83*). When a vehicle is identified at two locations within the network, the travel time between those locations can be estimated. VRI may be used interchangeably with automated vehicle identification (AVI), although AVI typically specifically refers to identifying vehicles using radio-frequency identification (RFID) transponders and receivers, or other unique identifiers. Vehicle signatures may be captured using license plate recognition (*84*) or media access control addresses captured from Bluetooth and Wifi devices within passing vehicles (*85*). Non-unique attributes of vehicle length (*86*) and magnetic signature (*87*) have also been used to define vehicle signature. Coifman and Cassidy (*86*) reidentified 20 % of vehicles based on length, and Sun et al. (*82*) used inductive loops and feature-based colour extracted from video stills to achieve an approximately 90 % match rate. Kwong et al. (*87*) used permanent wireless magnetic sensors installed across several intersections along an urban corridor. The authors estimate a successful matching rate of 65-75 % (*87*). One limitation of VRI is that the accuracy depends on the distance between sensors. As distance between sensors increases, so do the unknowns of the vehicle's path, decreasing the likelihood that the sensors measure travel time along the path of interest.

The moving observer method has traditionally been popular for measuring traffic conditions, though estimation techniques depend on the number of test vehicles utilized. Approaches using single vehicles must use statistical adjustments to extrapolate conditions experienced by the test vehicle to mean conditions (*18*). Conversely, if many test vehicles are available, traffic conditions can be estimated as the average from a relatively large number of floating cars operating within the same time and space. As the labour cost associated with floating cars is high (*17*), this technique is often impractical. Probe vehicles represent substantial improvement over methods using fixed sensors or floating cars. Li and McDonald (*13*) used GPS data and fuzzy logic to categorize the driving pattern of a test vehicle (fast, medium, or slow) to estimate the difference between the vehicle and average traffic conditions. The use of GPS is "a relatively cheap, efficient and effective means" of collecting traffic data (*20*), and smartphone GPS data collection enables the use of a large number of probe vehicles without the high labor costs associated with traditional floating cars. Additionally, data from regular drivers may better represent typical traffic conditions (*88*).

Though imperfect due to sample bias, vehicles instrumented with smartphones provide unparalleled ability to monitor the road network and estimate performance especially compared to alternative methods. Yet, appropriate performance measures for estimating traffic state using GPS data collection have not been determined. Measures of congestion are typically based on travel time or speed (*89*), such as the difference between actual and expected travel time (*90*, *80*) or the difference between actual and off-peak travel time, as in the travel time index (TTI) (*91*). These techniques can be used to derive historical trends and separate recurring and non-recurring congestion (*80*). Skabardonis, Varaiya, and Petty (*92*) utilized a delay-based approach for congestion measurement. With GPS probe vehicles, congestion measures based on instantaneous speed are preferred, as travel time estimation may be influenced by errors in reported positions and on assumptions of the start and end of a trip. In Washington State, mean speeds below 75 % of free flow speed define the onset of congestion (*91*). In Quebec, a threshold of 60 % is used (*93*). The Congestion Index (CI) was proposed by Dias et al. (*94*) as the difference between actual and free flow speed as a proportion of free flow speed. Although CI and other indices are limited to calculations on a particular link (route), they "can be used for an urban area wide application" (*95*).

Considering the recent advent of GPS data in transportation research, several shortcomings remain in this literature. Few studies have considered the rich source of data available from GPS-enabled smartphones. Congestion studies using probe vehicles have primarily focused on freeway corridors without consideration for estimating travel time or congestion at the network level.

2.3 Methodology

2.3.1 GPS Data Collection

For each trip i, logged into a smartphone application, GPS travel data is returned as a series of observations, O_{ij}, such as

$$trip_i = \begin{Bmatrix} O_{i0} \\ O_{i1} \\ \vdots \\ O_{ij} \\ \vdots \\ O_{in_i} \end{Bmatrix} = \begin{Bmatrix} i,\ c_{i0},\ t_{i0},\ x_{i0},\ y_{i0},\ z_{i0},\ v_{i0} \\ i,\ c_{i1},\ t_{i1},\ x_{i1},\ y_{i1},\ z_{i1},\ v_{i1} \\ \vdots \\ i,\ c_{ij},\ t_{ij},\ x_{ij},\ y_{ij},\ z_{ij},\ v_{ij} \\ \vdots \\ i,\ c_{in_i},\ t_{in_i},\ x_{in_i},\ y_{in_i},\ z_{in_i},\ v_{in_i} \end{Bmatrix}$$

where i is a unique trip identifier, O_{ij} is the j^{th} observation in trip i with $j = 0, \ldots, n_i$, c_{ij} is a

unique coordinate identifier, t_{ij} is the datetime, x_{ij}, y_{ij}, and z_{ij} are the latitude, longitude, and altitude, and v_{ij} is the speed. From each trip, several key pieces of trip information include the origin (x_{i0}, y_{i0}) and destination (x_{in_i}, y_{in_i}) and start (t_{i0}) and end times (t_{in_i}). Total travel time can also be computed $(t_{in_i} - t_{i0})$. The time between consecutive observations is typically between 1 and 2 seconds. Depending on the application, socio-demographic information may also be available. Once a trip has been collected and reported by the user, initial pre-processing of the data using methods like Kalman filtering (*96*) to reduce variability are typical. The data is then stored in a database from which observations are exported for analysis.

2.3.2 Data Processing

This thesis utilizes three main sources of data which must be processed prior to analysis. First, GPS travel data is processed using map-matching and speed filtering algorithms. Second, the road network data must be correctly redefined. Third, crashes must be assigned to the road network. These processes are described in the sections below.

Map Matching: Raw GPS traces contain positional variability even in cases where the data is pre-processed. Position is provided only in terms of latitude and longitude, and data is not linked spatially to the road network. If the goal is to determine traffic attributes at the link level, then it is necessary to explicitly match each trip to the travelled network links. Besides, map matching can eliminate positional noise by explicitly linking GPS observations to the road network. For this reason, map matching is preferred to other filtering methods that only smooth the data in terms of longitude and latitude. TrackMatching is a commercially available, cloud-based map-matching software service (*97*) that matches GPS trip data to the OpenStreetMap (OSM) road network (*98*). Before GPS data is sent to TrackMatching, the data is split into individual trips and formatted according to the software input requirements, including only the coordinate ID, timestamp, latitude, and longitude for each observation. The software requires no additional parameters to be set or input by the user. The software returns a new latitude and longitude, x'_{ij} and y'_{ij}, which correspond to a specific OSM link ID, l_{ij}, as shown below.

$$\{c_{ij},\ t_{ij},\ x_{ij},\ y_{ij}\} \rightarrow \text{TrackMatching} \rightarrow \{c_{ij},\ t_{ij},\ x'_{ij},\ y'_{ij},\ l_{ij},\ s_{ij},\ d_{ij}\}$$

x'_{ij} and y'_{ij} are chosen based on the Euclidean distance from the raw GPS points to the nearest link and on network topology (*99*). Track Matching also returns the source node s_{ij} and destination node d_{ij}, which can be used to identify direction of travel along the link. The algorithm generates a set of candidate paths and assigns the trip to the most probable path from origin to destination. After map-matching is completed, each observation corresponds to an exact location on the road network, and the series of matched links can be used to define the route from origin to destination. Based on observed trip sample, the algorithm accurately matched observations to the network.

Speed Filtering and Differentiation: A second filter is required to eliminate noise in the GPS speeds and derive accelerations. Although several filters have been proposed and tested, both Zaki, Sayed, and Shaaban (*100*) and Bagdadi and Varhelyi (*101*) found that the Savitzky-Golay filter was adequate for this application, noting that it is suitable for time series with fixed and uniform intervals and with limited discontinuities in the data (*100*). This digital filter is "a weighted moving average–based filter, with weighting described as a polynomial model of arbitrary degree" (*100*). In the Savitzky-Golay, both the degree of the fitted polynomial and the window size (the number of points to which the polynomial is fitted) can be varied to adjust the amount of filtering. There is always a compromise between maintaining the signal and eliminating the noise (*101*). Previous studies have suggested that a polynomial of degree two is adequate, although less guidance exists for selecting the window length. Windows of 3, 5, and 7 points were tested in Chapter 3.

An additional benefit of the Savitzky-Golay filter is the ability to filter not only the data, but also its derivatives. The acceleration rate can be determined for every observation by estimating the derivative of the filtered data. The results from the map matching and speed filtering algorithms are combined with the original data to yield the analysis data set. This contains the refined latitudes, longitudes, speed measurements (v'_{ij}), and acceleration rates (a_{ij}), and also ties each observation to a specific OSM link (l_{ij}) in the road network as shown below.

$$trip_i = \left\{ \begin{array}{c} i,\ c_{i0},\ t_{i0},\ x'_{i1},\ y'_{i0},\ z_{i0}, v'_{i0}, a_{i0}, l_{i0},\ s_{i0},\ d_{i0} \\ i,\ c_{i1},\ t_{i1},\ x'_{i1},\ y'_{i1},\ z_{i1}, v'_{i1}, a_{i1}, l_{i1},\ s_{i1},\ d_{i1} \\ \vdots \\ i,\ c_{ij},\ t_{ij},\ x'_{ij},\ y'_{ij},\ z_{ij}, v'_{ij}, a_{ij}, l_{ij},\ s_{ij},\ d_{ij} \\ \vdots \\ i,\ c_{in_i},\ t_{in_i},\ x'_{in_i},\ y'_{in_i},\ z_{in_i}, v'_{in_i}, a_{in_i}, l_{in_i},\ s_{in_i},\ d_{in_i} \end{array} \right\}$$

The process of collecting and filtering smartphone GPS data is illustrated in Figure 2-1.

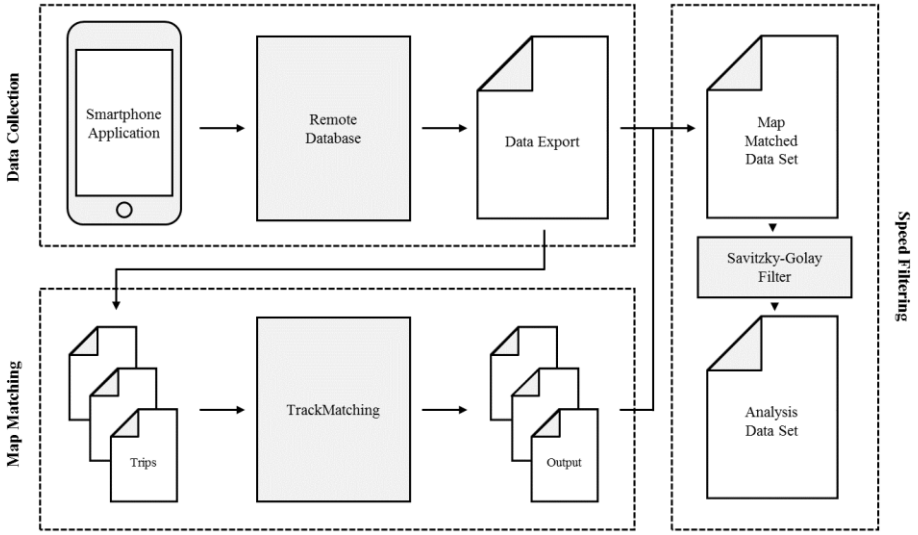


FIGURE 2-1 Collection and filtering of smartphone GPS data

Network Definition: The map-matching procedure links each observation to the OSM road network. Most of the network consists of five distinct functional classes: freeway, primary, secondary, tertiary, and residential (where primary, secondary, and tertiary are arterials and collectors classified by importance to the road network, with primary being most important). Ideally, these links would never be divided by an intersection (*75*). The OSM road network is generated non-systematically by users, and OSM links do not always meet this definition. It is desired to modify the network such that each link is properly defined between adjacent intersections. This modification requires several steps, which can be completed in any GIS software environment. The process, demonstrated in Figure 2-2, is as follows:

(1) Identify all nodes that represent an intersection in the road network. Nodes that define network topology, such as curves, are ignored.
(2) Split the road network at the identified nodes. Links connecting more than two intersections are broken into several smaller links. Properly defined links are unchanged.
(3) Rename each new link according to its original ID and the nodes on either end. Step 2 leaves several links with the same ID. To maintain a unique identifier for each link, the adjacent nodes are used to provide a unique ID.

(4) Remap the GPS observations to the new network. Travelled links in the GPS trip data are renamed using the same scheme as the map data, by concatenating the link ID, source node, and destination node into a unique identifier.

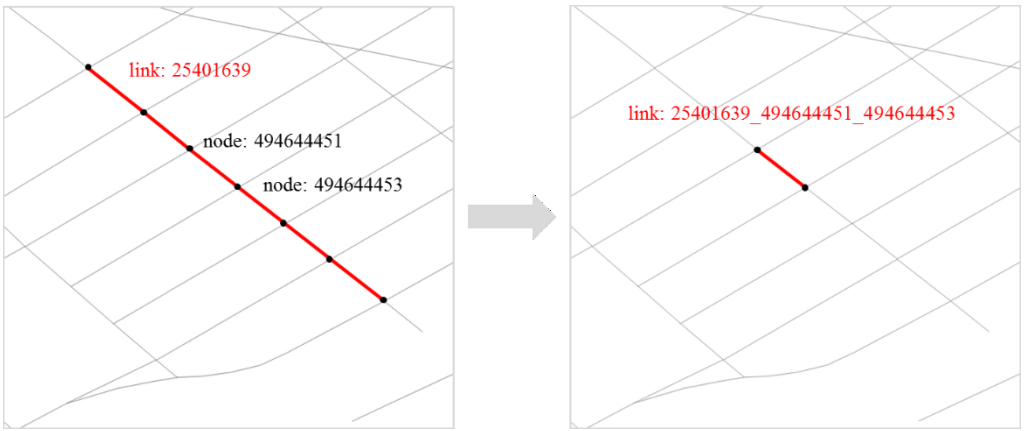


FIGURE 2-2 Redefinition of OSM links

Crash Assignment: The final step in data processing is assigning crash data to the links and intersections in the newly redefined road network to obtain crash counts at each link and intersection to model calibration and validation. Some collision reports may contain a latitude and longitude defining the location of the crash. However, for cases where coordinates are not provided, are provided inconsistently, or are inaccurate, a geocoding procedure must be used to determine the coordinates from text-based fields, including address or intersection. This project made use of a geocoding procedure developed previously by Burns et al. (*102*). Even with a latitude and longitude for each collision, crashes do not fall directly onto the network. In Chapter 3, buffers are used to assign crashes to individual links and intersections.

However, using buffers allows for the over-counting of crashes, which artificially creates spatial autocorrelations in crash models. One option to solve this problem is to assign crashes to the nearest link or intersection in the network. However, this method "increases the proportion of mislocated crashes, which can have serious implications in the modelling results. Instead, Chapter 4 uses 50 m buffers around links and intersections but does not allow link and intersection buffers to overlap. Although this method still allows for some overcounting due to adjacent links with overlapping buffers (less than 20 % at the link-level in the considered dataset), it improves the likelihood that crashes are assigned to the sites where they occurred. Though both methods

have benefits and limitations, preliminary testing, provided in Appendix C, showed that either method provided substantially similar results.

2.3.3 Computing and Visualizing Congestion

Computing the Congestion Index: The analysis data set can be used to compute various performance measures of interest to practitioners and/or the public, providing either a disaggregate view of link performance, or an aggregate view of network performance. As has been discussed, congestion is one critical measure of network performance that requires additional study. Aftabuzzaman (*95*) suggested that congestion measures meet several criteria including clarity, simplicity, comparability, and continuity. As link travel time is dependent on position, and because the precise latitude and longitude are untrustworthy (and are in fact removed as part of the map matching procedure), a congestion measure based on speed, rather than travel time, is preferred, even if those speeds are originally derived from the GPS positions. Dias et al. (*94*) proposed CI as one speed-based congestion measure, calculated as

$$\begin{aligned} CI &= \frac{\text{free flow speed} - \text{actual speed}}{\text{free flow speed}} && \text{if free flow speed} > \text{actual speed} \\ &= 0 && \text{otherwise} \end{aligned} \tag{2-1}$$

This formulation yields CI values ranging from 0 (speed equal to the free flow speed) and 1 (speed is zero) and meets several of the suggested criteria. The first necessary step is calculating the free flow speed on each link L. Free flow speed has been defined in numerous ways, though as congestion is generally constrained to the AM and PM peak periods, the speeds observed outside of these times can be used to estimate free flow speed. For this project, the morning peak period was defined as 6:00 to 10:00 AM, and the evening peak from 3:00 to 7:00 PM. The off-peak period T_{off} includes all other times. Free flow speed on a given link L is calculated as the average of all observed speeds on L during T_{off}, or

$$FFS_L = \frac{\sum_{i,j} v'_{ij}}{N} \quad \text{for all } O_{ij} \text{ where } l_{ij} = L \text{ and } t_{ij} \in T_{off} \tag{2-2}$$

where v'_{ij} is the speed for all observations where $l_{ij} = L$, and N is the number of those observations. Next, the congestion index for every observation can be computed according to

$$CI_{ij} = \frac{FFS_L - v'_{ij}}{FFS_L} \quad \text{if } FFS_L > v'_{ij}$$
$$= 0 \quad \text{otherwise} \tag{2-3}$$

where CI_{ij} is the congestion index for observation O_{ij}, FFS_L is the free flow speed on link l_{ij}, and v'_{ij} is the observed speed. As congestion levels vary across both distance and time, it is not only necessary to calculate CI at the link level, but also to calculate CI at different time intervals. The peak periods were divided into 60-minute time periods (one per hour) resulting in a total of eight periods. Therefore, the congestion index for link L during period T is calculated as:

$$CI_{LT} = \frac{\sum_{i,j} CI_{ij}}{N} \quad \text{for all } O_{ij} \text{ where } l_{ij} = L \text{ and } t_{ij} \in T \tag{2-4}$$

where CI_{ij} is the congestion index for all observations where $l_{ij} = L$, and N is the number of those observations on link L during T. To minimize noise, filters were added by setting a minimum acceptable number of trips and observations for CI calculation. For a valid CI_{LT}, L must contain at least 2 trips during T, each of which must have at least 2 observations on link L.

Visualizing Congestion: After the data is processed and CI is calculated, congestion can be visualized throughout the network, as has been discussed by several authors (*103*). Although CI can be computed for any single hour on any given day, a single instant in time does not provide general insight which would be beneficial to transportation professionals or to the driving public. Congestion levels vary significantly throughout the day (due to variation in demand) and vary significantly between days (due to variation in demand and non-recurrent phenomena including construction or crashes) and quantifying and/or visualizing this variation using only GPS travel data would be a significant contribution to existing research. Firstly, congestion can be visualized using a disaggregate approach, where each link is considered individually. First, CI was calculated for each hour of the peak periods, by pooling together all the weekday travel data in order to demonstrate hourly congestion variation on a typical weekday. Maps were generated by coloring

each link according to the congestion level observed during each hour. For this purpose, CI was divided into three categories; high congestion, CI of 0.30 or greater (consistent with Washington State (0.25) and Quebec (0.40) guidelines); moderate congestion, CI of 0.15 to 0.30, and low congestion, CI of 0.00 to 0.15. A detailed understanding of congestion should include both the average level of congestion and some indication of daily variability of congestion. To capture daily variation, CI was calculated for every hour of each weekday independently. The number of hours each link experienced CI above 0.30 were summed, out of a possible 120 hours (8 peak hours per day over 15 weekdays). Maps were generated to show not only which links were most congested, but also to show which links were congested most consistently over the 15 days of study.

Although one strength of this type of analysis is that each link can be viewed independently, it may be difficult or impossible to make meaningful conclusions about the behavior of the network in general. To facilitate observation of network-wide trends, it may be appropriate to aggregate the data in some way. Congestion does not occur all at once. Instead, it gradually builds and then subsides throughout the peak periods. Similarly, congestion does not occur across all network links simultaneously. This study aggregated in two different ways; first by distance to the city center, and second by roadway functional classification. Aggregation by distance has the potential to show how congestion propagates with respect to the downtown core. To accomplish this, the city center was defined as the location of Quebec City Hall. Bins of 200 m distances were defined with respect to City Hall. The average CI was then computed for each distance bin for each hour of the day, by pooling together all the weekday travel data. Next, links were categorized according to their OSM functional classification. The analysis was completed separately considering five distinct functional road classes: freeway, primary, secondary, tertiary, and residential. Using the pooled weekday data, the proportion of links experiencing high, moderate, and low congestion levels were computed and plotted for each hour for each of the five functional classes.

2.4 Results

2.4.1 Data Description

GPS travel data was collected in Quebec City, Canada using the Mon Trajet application (*70*) developed by Brisk Synergies (*104*). Screenshots from the application are shown in Figure 2-3. The application, which was available for Apple and Android devices, was installed voluntarily by drivers and allowed them to anonymously log trips into the application, which are automatically

uploaded and stored in a cloud-based platform. In total, approximately 5000 driver participants have logged nearly 50,000 trips using the application. The data used in this study is a sample of publicly available data, containing over 4000 drivers and 21,939 individual trips during the period between April 28 and May 18, 2014. Over the 21 days sampled, 19.7 million individual data points were logged. 11 years of crash data were obtained from the MTQ for the period between 2000 and 2010. In total, 14,278 minor injury, major injury, and fatal collisions involving at least one vehicle were identified during this period. Geometric map data was obtained in spring 2015 from OSM to ensure consistency with the map matching results.

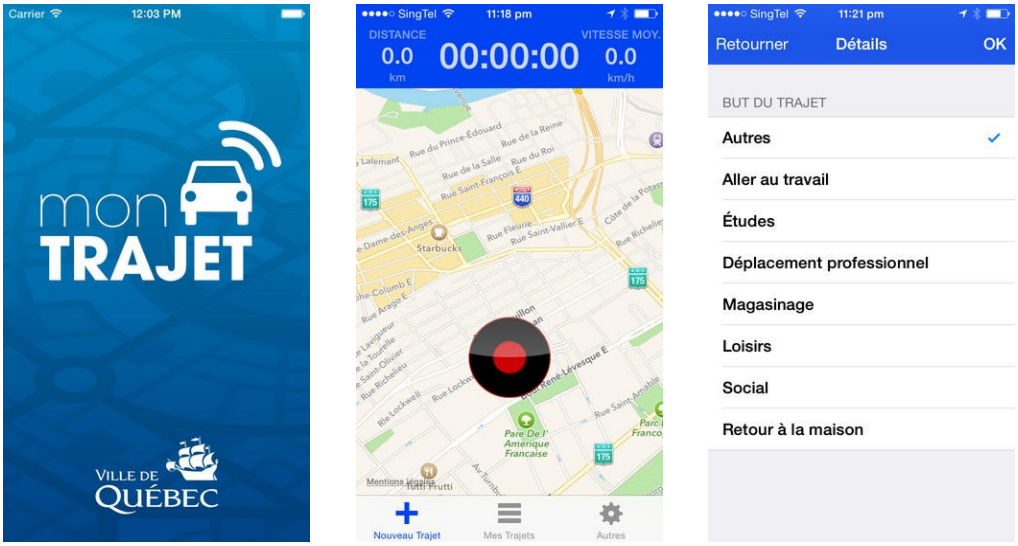


FIGURE 2-3 Smartphone application interfaces

2.4.2 Disaggregate Visualization

Figure 2-4 and Figure 2-5 present the level of congestion experienced throughout the road network for each of the eight AM and PM peak period hours. As CI for these maps is based on pooled weekday data, the figures demonstrate the level of congestion expected on a typical weekday. In this pooled data set containing fifteen weekdays of data, each peak hour contains CI measurements for between 6600 and 12,000 links. Considering the Quebec City network contains over 50,000 total links, this represents between 13 % and 24 % of the total road network. Although the collection campaign in Quebec City was large (the sample for this study contained nearly 22,000 trips), many network links contain no observations. Still, the most critical links, including major freeways, arterials, and collectors, are well populated, and adequately demonstrate congestion

trends. In fact, those links that are missing data tend to be low volume residential streets were the impact of congestion is low. As with the total number of trips in the population, the total number of GPS trips logged varies with time, first growing to a maximum and then subsiding throughout the peak periods. This phenomenon is clearly observed in Figure 2-4 and Figure 2-5. For example, the period beginning at 6:00 PM has far fewer observations than the period beginning at 4:00 PM.

This result largely corroborates intuition. At the beginning of the AM peak period, very few links are highly or moderately congestion (most have CI < 0.15, so are operating at or near free flow speed). By 7:00 AM, the freeways and major arterials become highly congested (CI exceeds 0.30). A similar pattern is observed at 8:00 AM. However, by 9:00 AM, congestion has largely subsided on the freeways and arterials, while some streets in the downtown core remain congested. A mirror image of this pattern is observed in the PM peak period, with congestion beginning to form at 3:00 PM, highly congested conditions at 4:00 PM and 5:00 PM, and dissipation at 6:00 PM. As a result, each 4-hour peak period can be viewed as having an onset period (lasting approximately one hour), the peak itself (lasting approximately two hours), and a dissipation period (lasting approximately one hour).

Rather than pooling all weekday data, every peak hour can be considered independently. Each individual weekday yields at least one CI measurement for between 2000 and 4000 links (about 6 % of the network on an average day). Each individual peak hour contains between 250 and 1750 links with a CI measurement (between 0.5 % and 3.5 % of the total road network). Figure 2-6 shows the total number of hours spent in the highly congested state for all links with at least one peak hour observation. Most links experiencing a highly congested state do so for about 30 hours over the 15 study days (about 2 hours per day). However, some links spend up to 61 hours in the highly congested state. Highly directional links, such as motorways and arterials which carry commuters towards the city center, experience high flow for one peak period (either AM or PM), relating to a possible 60 hours out of the total 120 peak period hours. A link with nearly 60 hours of high congestion could be considered chronically congested (consistently congested during either the AM or PM peak). The most consistently congested locations are Autoroute Felix-Leclerc, running east/west north of the city center, the interchange connecting Autoroute Felix-Leclerc with Autoroute Henry-IV, and several arterial links near Laval University west of the city center.

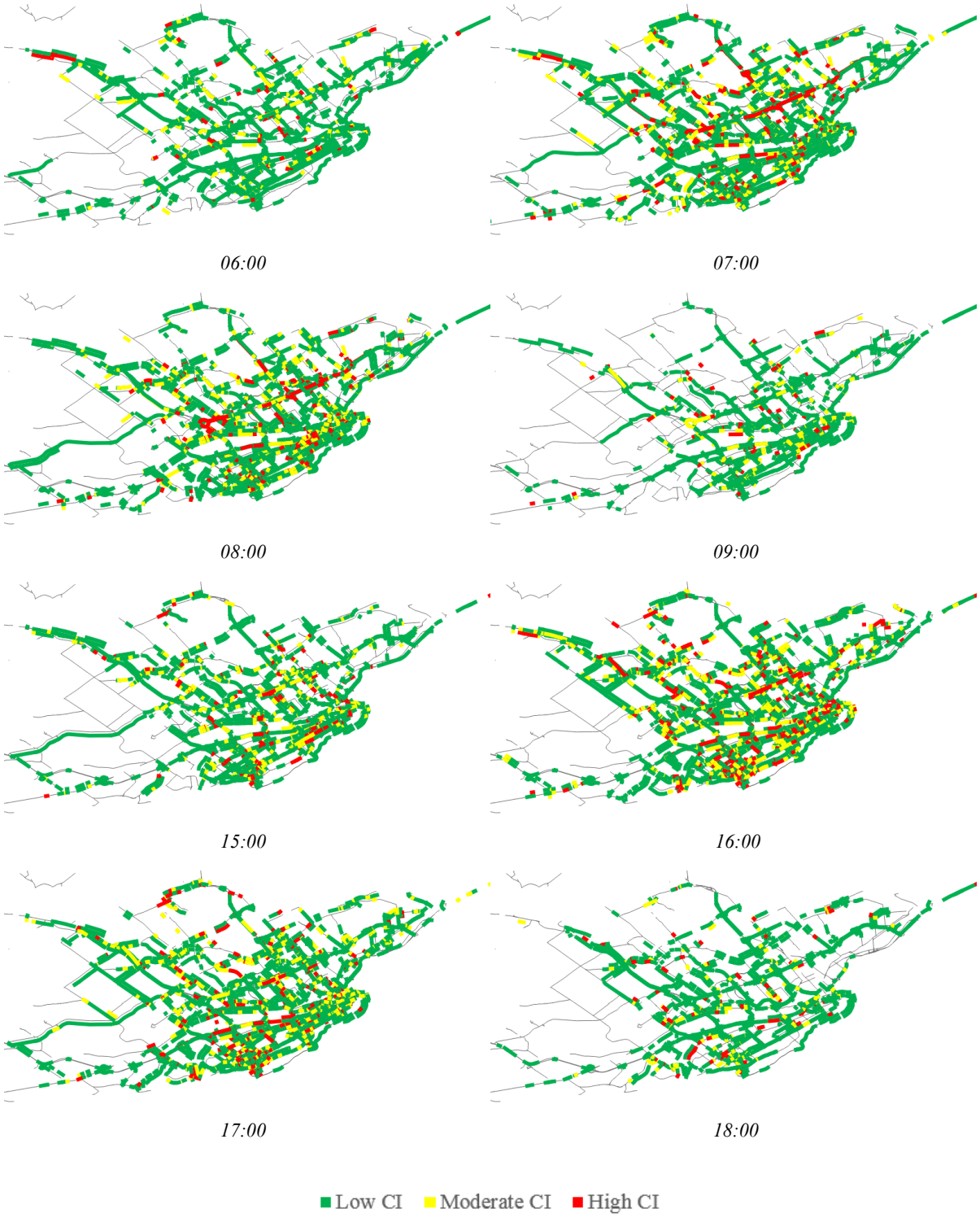


FIGURE 2-4 High, moderate, and low hourly CI levels for the network during peak periods

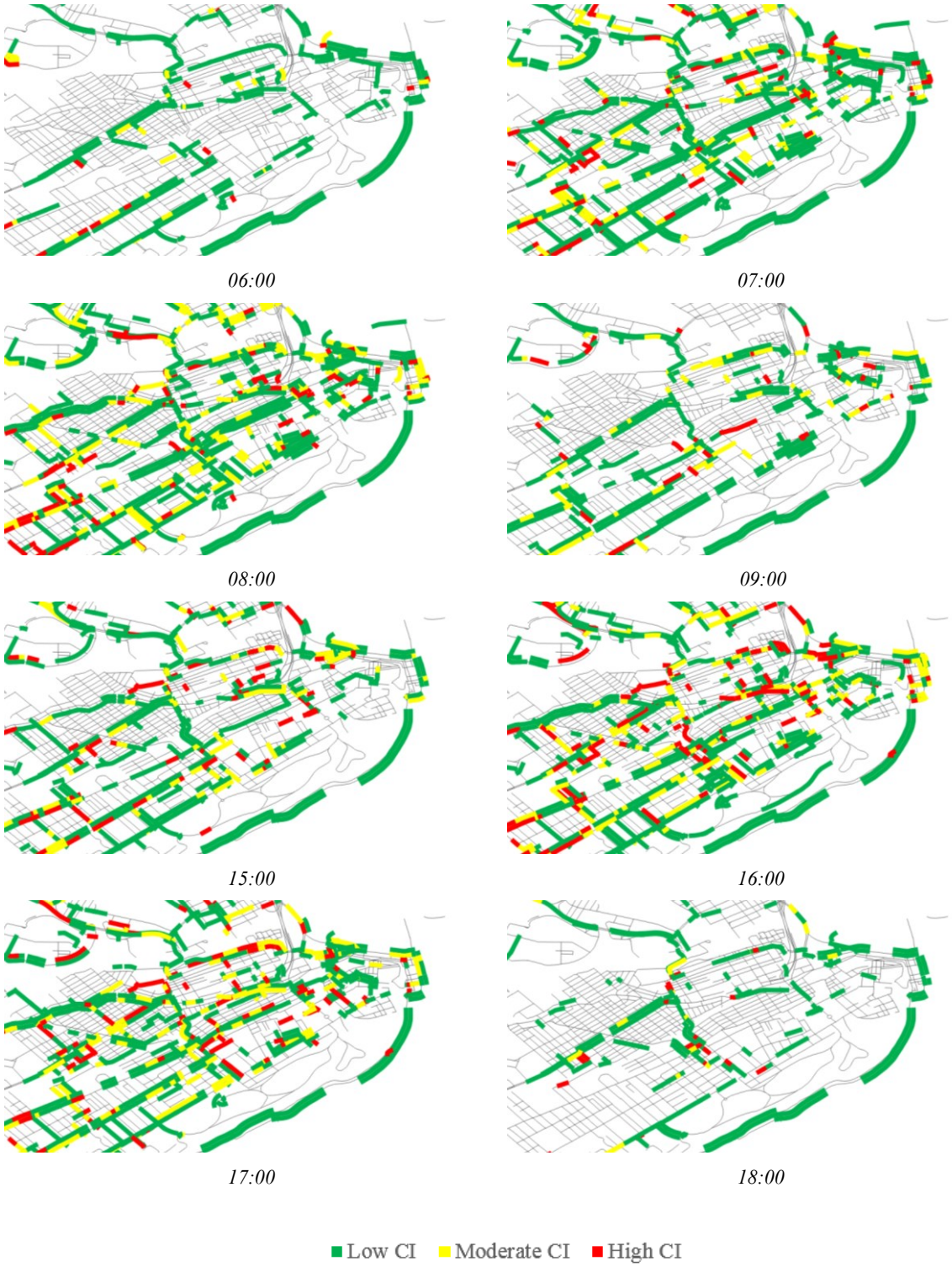


FIGURE 2-5 High, moderate, and low CI levels for downtown during peak periods

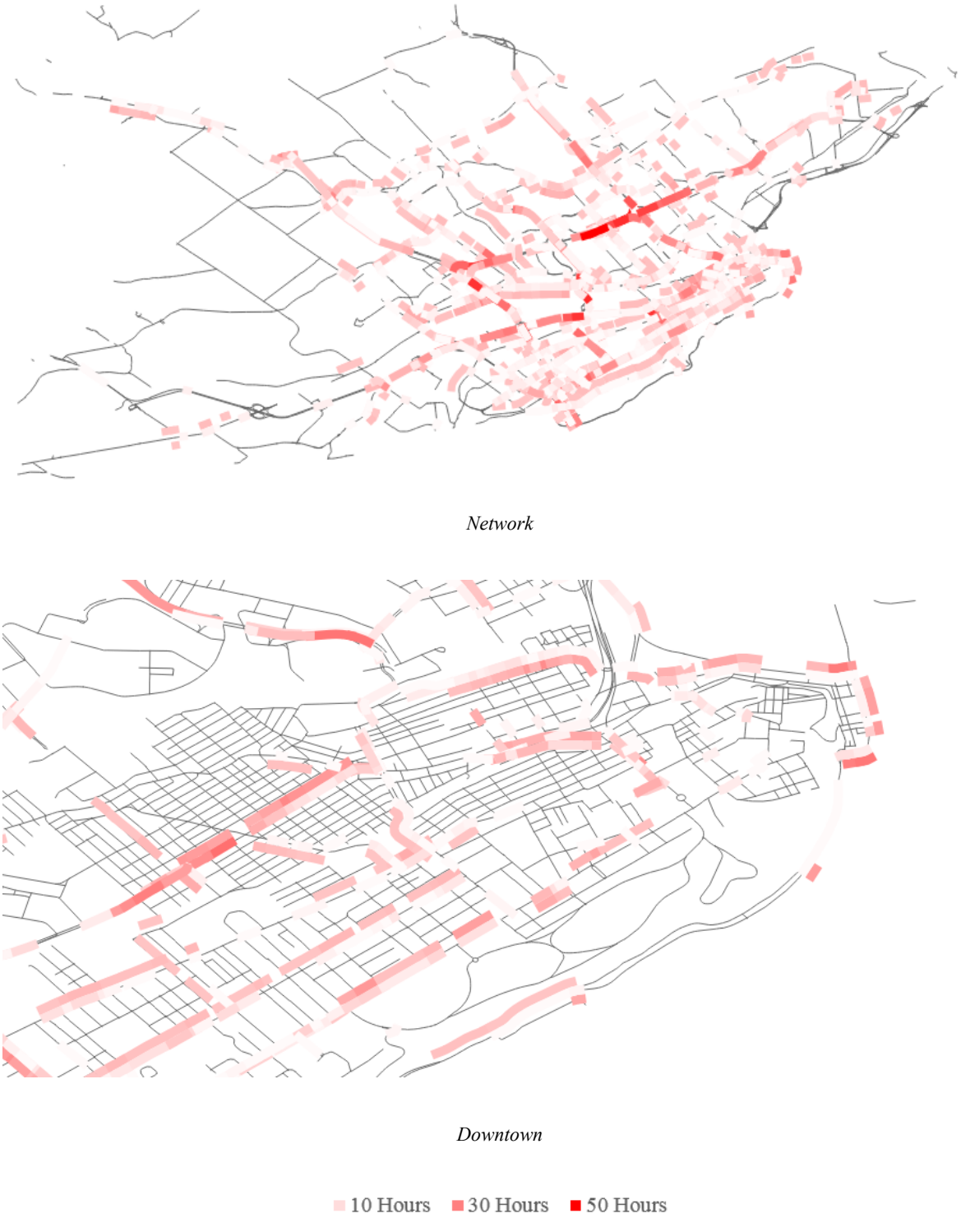


FIGURE 2-6 Total number of peak period hours exceeding CI levels of 0.3 over three weeks

2.4.3 Aggregate Visualization

To observe general trends related to the formation and propagation of congestion within Quebec City, links were aggregated according to distance to the city center and by functional classification. Figure 2-7 shows the average CI for each peak hour (calculated from the pooled weekday data) based on distance from the city center. Some results from the disaggregate analysis become further pronounced when aggregating by distance. Firstly, the onset and dissipation period in both the AM and PM, with relatively lower levels of congestion, are clearly seen (congestion is generally lower in the first and last hour of each peak period). Second, congestion appears to move towards the city center in the morning, and away from it in the evening. Third, congestion levels in the PM peak period tend be higher (particularly for links within 5 km of the city centre) and higher congestion levels are present further from the city center when compared to the AM peak.

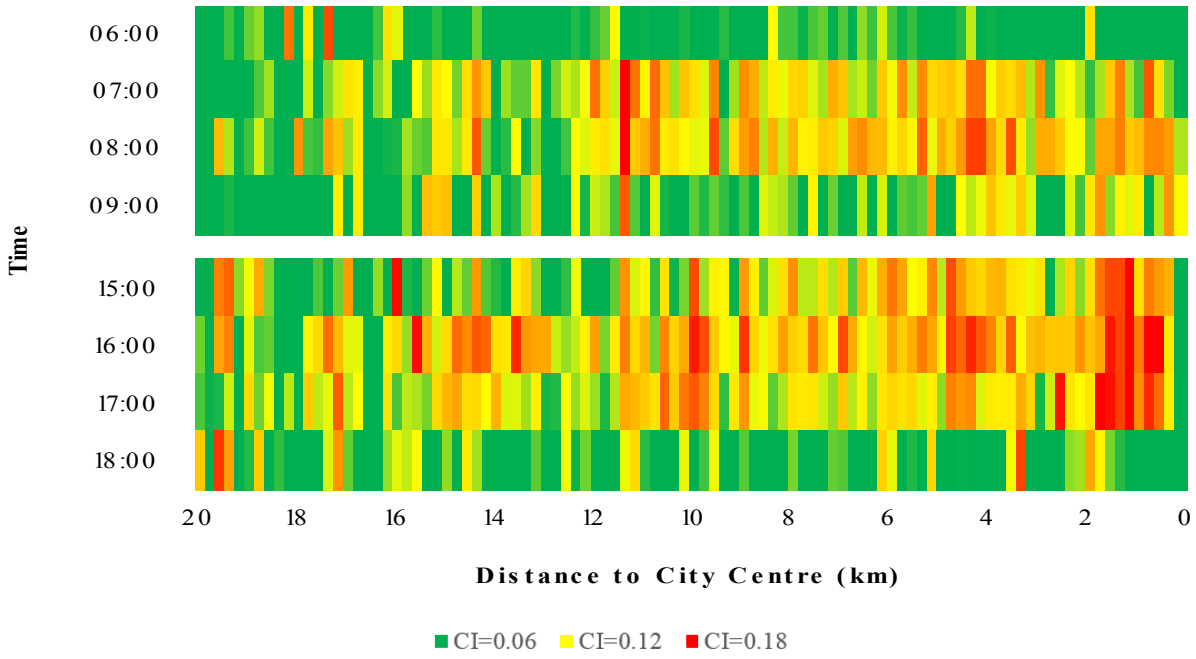


FIGURE 2-7 Average CI levels over peak periods with respect to distance from city center

When aggregating links by functional classification, the relative impact of different facility types on the magnitude of and variation of congestion can be detected. Figure 2-8 shows the proportion of links in each of the low, moderate, and high congestion states for five roadway functional classes. Motorways had the most severe and most pronounced variation in congestion. During the most congested hours, nearly 20 % of all motorway links were in the highly congested

state, and an additional 10 % to 15% were in the moderately congested state. The proportion of links with low congestion ranged between 66 % and 95 %. Primary, secondary, and tertiary links all share a similar pattern. Although variation in congestion levels was still observed, the variation on these arterials and collectors was less pronounced than for motorways. While the proportion of links in the highly congested state is more stable across the peak hour (between 2 % and 14 %), a greater proportion of links were at moderate congestion levels. In contrast, congestion of residential links was much more stable throughout the peak periods, with approximately the same proportion of links at all three congestion levels throughout. Several key results from the above analysis are summarized as follows:

(1) Each peak period can be viewed as having an onset period and dissipation period lasting one hour each. Between these periods, congestion levels are relatively stable.
(2) Congestion in the evening peak period is greater and higher congestion levels are present further from the city center than in the AM peak period.
(3) Chronically congested links include the major motorways and arterials which lead to the city center.
(4) Motorways, followed by major arterials and collectors, contribute most to peak period congestion. Residential links contribute little to peak period congestion levels.

2.5 Conclusions

The purpose of this chapter was to propose measures for estimating and visualizing congestion levels across time and space in an urban road network using data collected from the GPS-enabled smartphones of regular drivers. This chapter first presented the methodology for processing the GPS data and other associated data sources. Through map matching and network definition, observations are explicitly related to links in the road network. Speed filtering ensures that noise in the measured speed data is reduced. The measure and method for evaluating congestion proved to be easy to compute. Although the presented visualization is largely qualitative, the general trends observed are similar to the expected patterns of congestion at both the microscopic and macroscopic levels. Despite some limited spatio-temporal data coverage, enough data was available to calculate and visualize congestion for most major freeways, arterials, and collectors within Quebec City.

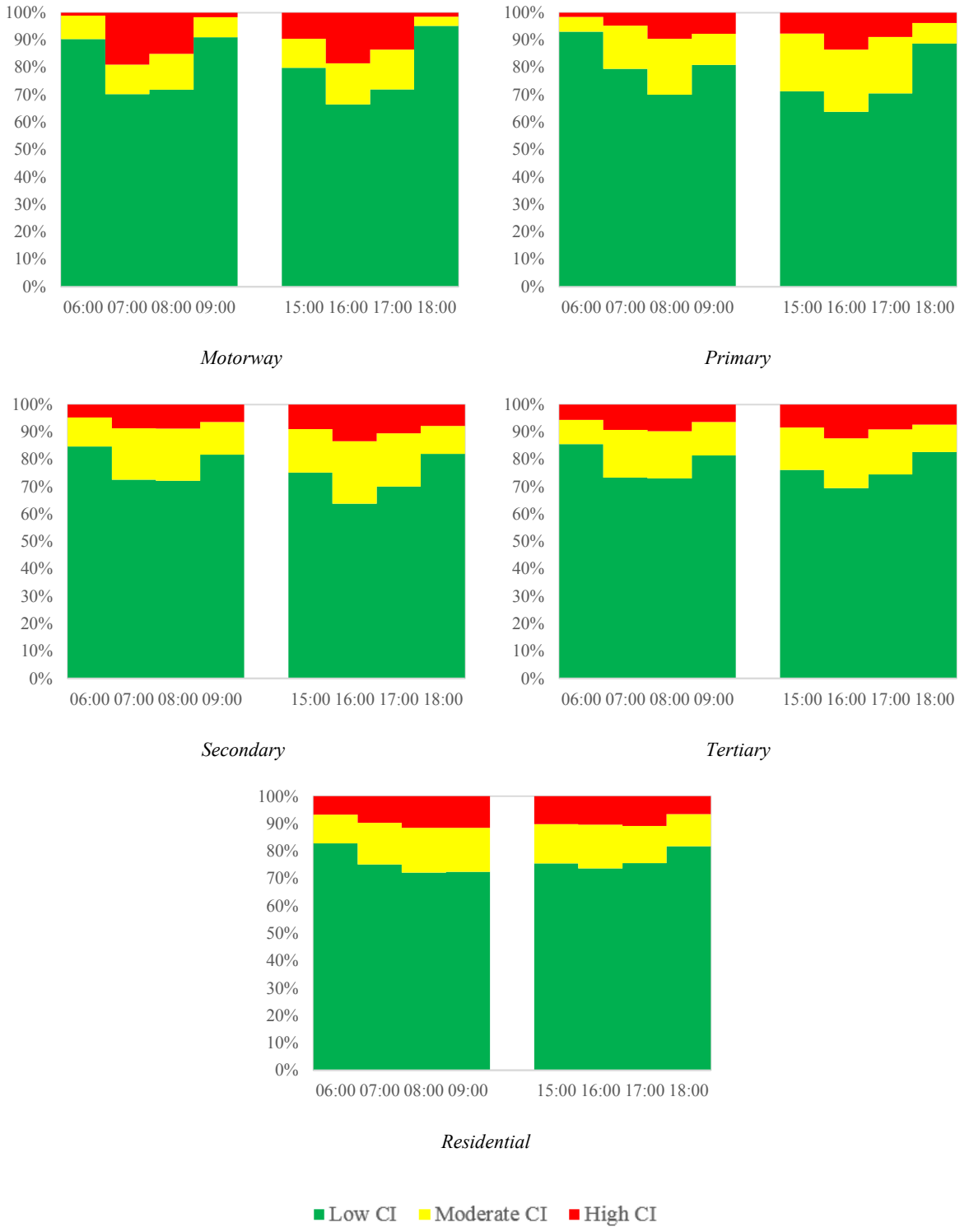


FIGURE 2-8 Proportions of links at high, moderate, and low CI levels segmented by functional classification

Several methods for visualizing congestion and its variation over time and space were explored. Disaggregate maps were generated, which ably demonstrated the rise and fall of congestion throughout the AM and PM peak periods on a typical weekday. Furthermore, several chronically congested links were identified by counting the hours each link spent in a highly congested state (CI > 0.30). While this type of work is beneficial for visualizing congestion and identifying sites for improvement, more aggregate analyses are required to observe general network trends. By aggregating links based on distance from the city center, the peak periods were clearly observed to have both an onset and dissipation period lasting approximately one hour each. Congestion was observed to move towards the city center in the morning, and away from it in the evening.

Perhaps the most surprising observation was that congestion is more severe and that higher congestion levels are present further from the city center in the PM peak. Finally, by aggregating links by functional classification, the relative influence of each facility type on peak period congestion trends was observed. Motorways had the most severe CI levels and the most variation within the peak periods, followed by collectors and arterials. Residential links had less variation in congestion levels across the peak periods. This bodes well for this type of analysis. As stated, GPS data is absent for a vast majority of residential links. However, based on the links for which data is available, the link with missing data are unlikely have a major impact on overall congestion trends, and so their absence from the data set is unlikely to skew the results.

CHAPTER 3: EXTRACTING AND VALIDATING SURROGATE SAFETY MEASURES

"Precaution is better than cure."

Edward Coke, Judge

3.1 Introduction

The safety of urban road networks is a serious concern that requires the continuous monitoring of crash frequency and severity. Considering limited budgets, the most dangerous sites in the network should be identified and prioritized for remediation to maximize the efficiency of countermeasures. An alternative screening method based on SSMs, which would represent an important response to the various issues associated with crash-based network screening, requires an alternative data source from which crash risk can be constantly and systematically estimated throughout the network. Instrumented vehicles are currently the only source capable of providing long periods of continuous data, albeit for a relatively small sample of road users (*4*). With the proliferation of GPS-enabled smartphones, which themselves contain many of the same sensors used for instrumenting vehicles, large volumes of reliable and spatio-temporally rich naturalistic driving data can now be collected passively from regular drivers (*38*) in crashes, near crashes, and under normal conditions (*29*, *105*). Though limited in terms of the studied population of drivers, the spatial coverage of GPS data creates the potential to reduce dependency on crash data in network screening (*34*) by supporting safety assessments based on SSMs rather than crash data.

Either event-based measures or traffic flow measures can be extracted from smartphone GPS data. Event-based measures, including traffic conflicts, interactions, or vehicle manoeuvres were first studied in the late 1960s based on human observation. Since then, video-based sensors and computer vision techniques have improved objectivity and increased the amount of data that can be processed. Though video-based sensors provide high temporal resolution (*4*) and rich positional data (*33*), the analysis of video data is potentially resource intensive and has spatial limitations (*34*, *4*), leading to a desire to implement event-based techniques using other data sources. Traffic flow techniques use measures of volume, mean speed, or density to estimate risk (*39*), typically requiring roadside point sensors such as loops or radar detectors (*40*, *41*, *42*). Though successful on freeways, it is impractical and costly to implement roadside sensors across an entire urban network (*26*). As traffic flow "is a dynamic phenomenon with elements of both space and time" (*20*), studies of flow are an ideal application for instrumented vehicles, and "there is increasing evidence that traffic conditions can be estimated accurately using only vehicular GPS data" (*106*). As mobility and safety are the two greatest priorities within any transportation system (*94*), simultaneous improvements to traffic flow and reduction in crashes are desired (*94*). However, newer theories contradict earlier beliefs that flow and crashes are positively linked (*94*).

Despite the advent of smartphone GPS data in the last years, few studies have thoroughly investigated the types of SSMs that can be extracted from such a data source. Accordingly, the purpose of this chapter, previously published in two studies (*107*, *108*) is to examine event-based measures (vehicle manoeuvres) and traffic flow measures (quantitative measures of congestion, speed, and speed uniformity) extracted from probe vehicle data collected by the GPS-enabled smartphones of regular drivers, and to correlate these measures with historic collision frequency and severity at the network scale. The specific objectives are to:

(1) Explore the potential surrogate measures available from smartphone data;
(2) Describe the procedure for extracting such measures, and;
(3) Investigate the relationship between SSMs and collision frequency and severity across different facility types.

3.2 Literature Review

Although probe vehicles have primarily been used in spatio-temporal applications such as traffic monitoring and origin-destination studies (*26*), they have been applied less frequently in studies of road safety. This underutilization can be partially attributed to the difficulties of collecting large volumes of data using dedicated GPS devices that are installed for a specific research purpose (*26*). Although speed is often regarded as an important surrogate measure, changes in speed (acceleration, the first derivative of velocity, or jerk, the second derivative) may be more important (*34*), and several studies have attempted to extract vehicle manoeuvres from probe vehicles as SSMs. Agerholm and Larhmann collected data from six drivers over a 3-month period using GPS devices and accelerometers. The authors stated that "braking was the evasive action […] in 88 % of the accidents in built-up areas" (*4*), making decelerations a logical indicator to extract. Jerk was found to be correlated with crash occurrence (*4*). Bagdadi (*29*) used GPS, accelerometer, and radar data from 109 participants and found that jerk could correctly identify self-reported near misses at an 86 % success rate (*29*). Jun, Ogle, and Guensler (*38*) analyzed spatio-temporal driving activity and crash involvement using dedicated GPS devices and self-reported safety data for 460 light-duty vehicles. The study found that drivers involved in crashes tended to travel longer distances and at higher speeds and "engaged in hard deceleration events" (greater than 2.7 m/s^2) more frequently (*38*). Though failing to show a causal link, the authors suggest that decelerations "may

be employed as roadway safety surrogate measures" (*38*). Fazeen et al. (*37*) used smartphone accelerometer data to classify 'safe' accelerations and decelerations from 'unsafe' ones (approximately 3 m/s^2 or greater), though failed to demonstrate whether 'unsafe' behaviour led to more collisions and used only a single smartphone. Ellison, Greaves, Johnson, and Trivedi (*109*) developed a system to distinguish non-aggressive and aggressive driving behaviour. Their system fused accelerometer, gyroscope, magnetometer, GPS, and video data from smartphones to monitor drivers. However, at the time of publication, the system had only been installed in three vehicles. Eren et al. (*110*) similarly studied maneuvering using the smartphone accelerometer and gyroscope data of 15 drivers. Guido et al. (*111*) attempted to evaluate time-to-collision (TTC) and deceleration rate as measured from smartphone GPS data as possible SSMs for rear-end collisions on a two-lane rural highway. The study used only three drivers and no attempt was made to correlate the results to actual collision risk.

The relationship between traffic flow and safety has largely been studied using roadside sensors. In terms of congestion, numerous studies have examined the relationship with road traffic safety using various methods and with various results. For example, Noland and Quddus (*112*) and Wang, Quddus, and Ison (*113*) found little effect on safety. Martin (*114*) found that collision frequency and severity were elevated during light traffic, while Zhou & Sisiopiku (*115*) found that the most collisions occurred at either low flow or high flow, though crash severity was greatest at low flow. Dias et al. (*94*) concluded that increasing congestion could also increase the probability of a collision. Wang et al. (*116*) found a negative (though statistically insignificant) relationship between congestion and collision severity, though CI values were consistently less than 0.5 (*94*). Other traffic flow measures have also been considered. For example, Oh et al. (*40*) hypothesized that the average and variance of flow, occupancy, and speed could be indicators of traffic state. The standard deviation of speed was correlated with disruptive traffic conditions and a higher likelihood of collisions (*40*). Lee et al. (*42*) used a log-linear statistical model to show that variation in speed and density were significantly correlated with crash frequency (*42*). Abdel-Aty and Pande (*7*) applied a Bayesian classifying approach to categorize traffic conditions as either leading to or not leading to a crash. Golob et al. (*41*) defined eight traffic flow regimes using speed and flow, concluding "the key elements … affecting safety are not only mean volume and speed, but also variations in volume and speed". Despite the limited literature on traffic flow SSMs derived from GPS data, if the volume of GPS data can be practically increased by leveraging the capabilities of

GPS enabled smartphones, traffic flow SSMs could be computed. Moreno and Garcia (*117*) present a methodology using GPS trackers in which speed profiles are used to calculate measures of speed uniformity and over speeding. Boonsiripant (*118*) utilized GPS data from the Commute Atlanta program to develop SSMs based on the variation in vehicle speed profiles.

Several shortcomings are apparent in the existing literature, which this chapter attempts to address. First, there has been no attempt to derive SSMs from smartphone-collected GPS data of regular drivers alone. Existing studies have used dedicated probe vehicles (resulting in sample sizes of 100 drivers or less) or dedicated GPS devices with supplemental accelerometer data. Studies using smartphones have used extremely few drivers, despite the potential for application to the population at large. Second, there has been no comprehensive comparison of GPS-based SSMs to large quantities of crash data at the network scale. Finally, there has been no application of GPS data across an entire urban network. Studies to date have focussed on specific corridors (namely freeways) despite that fact the GPS presents an opportunity for network-wide analysis.

3.3 Methodology

3.3.1 Extracting Surrogate Safety Measures

Hard Braking and Acceleration: Deceleration is perhaps the most common evasive manoeuvre in urban areas (*4*), and selecting hard braking events (HBEs) as a potential SSM is logical. Most studies focused on vehicle manoeuvres have used jerk, observed using accelerometers, as the surrogate indicator (*29*, *4*). Although the resolution of the GPS data is typically too coarse to calculate jerk, decelerations can also be used to detect unsafe behaviour. For example, Fazeen et al. (*37*) suggested using decelerations exceeding 3 m/s^2. Therefore, using a deceleration threshold may be sufficient to define HBEs. Some studies have suggested that hard acceleration events (HAEs) may also be good predictors of safety (*105*, *4*). Accelerations computed using the Savitsky-Golay filter are compared to a braking threshold, a_{min}, and acceleration threshold, a_{max}, to determine which observations correspond to HBEs or HAEs. The status of O_{ij} is determined using the following logic. For each series of consecutive negative (respectively, positive) accelerations, l_{dec} (l_{acc}), the minimum (resp. maximum) value is obtained. If this value is inferior (resp. superior) to the threshold, i.e. $\min(l_{dec}) < a_{min}$ (resp. $\max(l_{acc}) > a_{max}$), then consider that observation and only that observation (the minimum or maximum acceleration) an HBE (resp.

HAE). This ensures that a single deceleration spanning multiple observations is considered a single HBE, even if several observations exceed the threshold. This algorithm is described in Figure 3-1. Although the value of -3 m/s^2 was chosen for a_{min} as a starting point to develop GPS-based surrogate safety indicators (based on previous studies), thresholds of -2 m/s^2 and -4 m/s^2 were also tested. Similarly, values of 2 m/s^2, 3 m/s^2, and 4 m/s^2 were used for a_{max}.

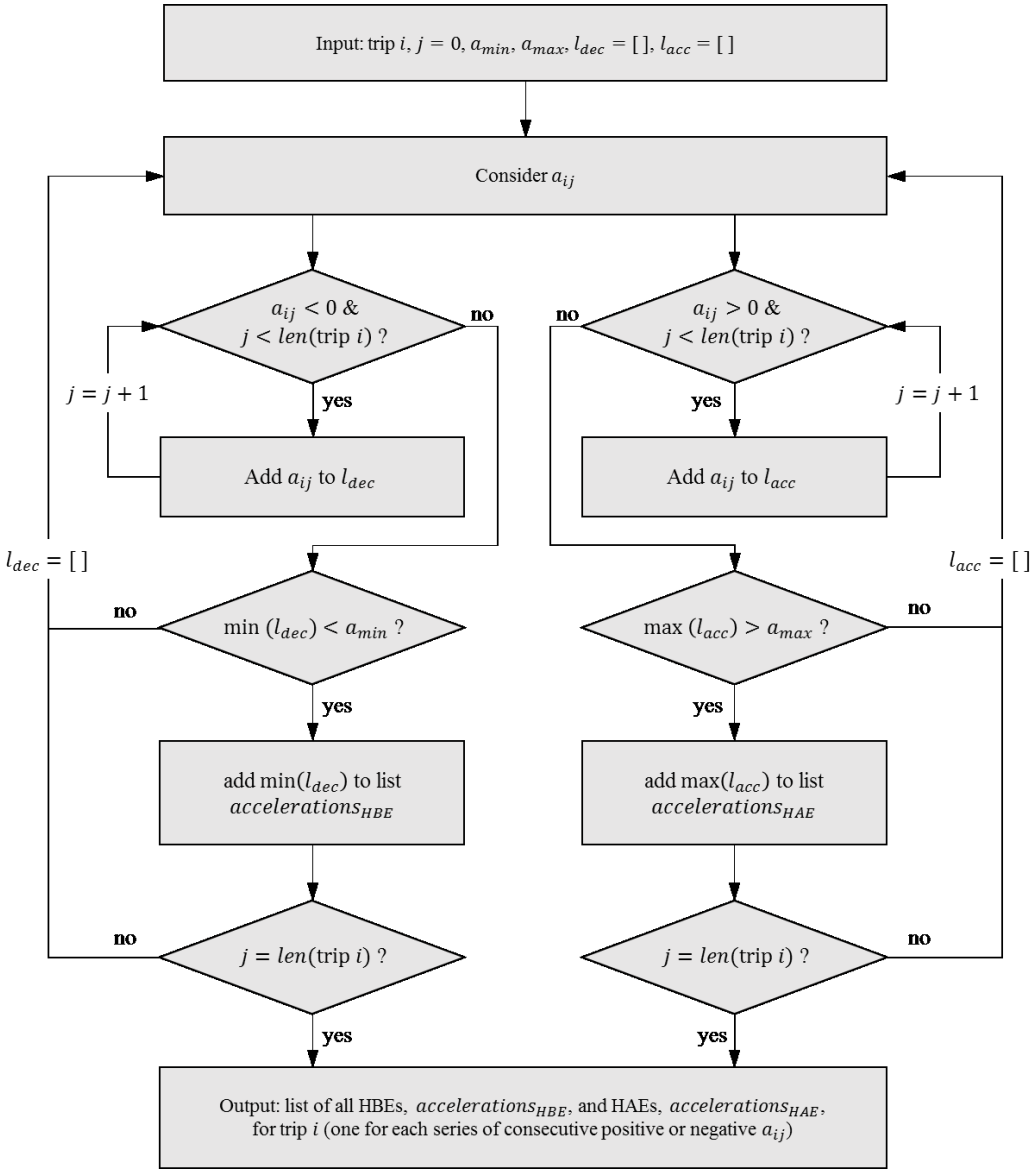


FIGURE 3-1 Algorithm for extracting vehicle manoeuvres from GPS trip data

Congestion Index: Like Boonsiripant (*118*), each proposed traffic flow SSM was calculated across several time periods throughout the day. In this chapter, the considered time periods are the AM peak (6:00 to 10:00 AM), PM peak (3:00 to 7:00 PM), the peak periods combined, and the off-peak trips. A single period was chosen for each SSM based on which time period was shown to be most strongly associated with collision frequency during analysis, as shown in Appendix A. CI is calculated as described in Chapter 2. Preliminary results indicated that CI during the PM peak had the strongest relationship with crash frequency and severity in this study.

Average Speed: Travel speed has long been believed to be an indicator of crash risk at the link level. Over speeding is regarded as a dangerous behaviour, and efforts to increase adherence to the speed limit have been implemented across the world. Locating facilities with high occurrence of over speeding would also be beneficial to law enforcement, who could increase the effectiveness of enforcement operations by targeting identified sites. Average speed is calculated for each link, by considering every observation falling on that link. The average speed is computed as

$$\bar{V}_L = \frac{\sum_{i,j} v'_{ij}}{N} \quad \text{for all } O_{ij} \text{ where } l_{ij} = L \tag{3-2}$$

where v'_{ij} is the speed for all observations where $l_{ij} = L$, and N is the number of those observations . The off-peak period was chosen for calculation to avoid collinearity with CI. In this case, $\bar{V}_L$ is exactly equal to FFS_L defined earlier. In practice, average speed can be computed for other time periods, so it is not strictly necessary that these measures be equal.

Speed Uniformity: Although the magnitude of speed is widely believed to contribute to crash occurrence, much of the existing literature supports that variation in speed may be a better predictor of risk. Although several measures of speed uniformity have been suggested, including standard deviation, coefficient of variation, and acceleration noise, Ko et al. (*119*) demonstrated that these measures are highly correlated. Although initial tests were done on several possible SSMs, shown in Appendix A, the coefficient of variation of speed (CVS) was determined to have the strongest relationship with safety. CVS is defined as the standard deviation of speed over the mean speed. For each link, CVS is computed as

$$CVS_L = \frac{\sigma(v'_{ij})}{\bar{V}_L} \quad \text{for all } O_{ij} \text{ where } l_{ij} = L \tag{3-3}$$

where $\sigma(v'_{ij})$ is the standard deviation speeds for all observations where $l_{ij} = L$ during the considered time period. CVS during the off-peak period was found to be most strongly related to crash frequency and severity for this study.

3.3.2 Validation of Measures

SSMs must be predictably related to crashes (*10*), and any proposed measure must demonstrate correlation with actual safety (collision frequency and/or severity). Although the measures above have been proposed as SSMs, the main objective of this chapter is to demonstrate the statistical relationship between the SSMs and collision frequency and severity by comparing the proposed indicators with historical collision data at the link and intersection levels. It is necessary to consider both frequency and severity, as the relationships of these independent dimensions of safety with SSMs are likely complex. Boonsipirant (*118*) suggested that separating facilities according to functional classification was necessary if measures were to adequately predict collision occurrence. Therefore, the analysis was carried out separately considering five distinct functional classes: freeway, primary, secondary, tertiary, and residential (where primary, secondary, and tertiary are arterials and collectors classified by importance to the road network, with primary being most important). These classes represented nearly all the travelled links in the GPS data.

Collision Frequency: Spearman's Rank Correlation Coefficient, or Spearman's rho, indicates how strongly the dependency between two variables is described by a monotonic function and is used by the FHWA for the evaluation of SSMs (*120*). Locations with the most collisions should also have the most vehicle manoeuvres if most collisions involve road users' attempt at evasive actions. A rho of 1.0 indicates perfect positive correlation, 0.0 indicates no correlation, and -1.0 indicates perfect negative correlation. Spearman's rho (ρ) between the collision frequency and the SSMs at the site level is calculated using

$$\rho = 1 - \frac{6\sum(w_L - v_L)^2}{M(M^2 - 1)} \tag{3-4}$$

where w_L is the rank of site L based on collision frequency, v_L is the rank of site L based on the SSMs, and M is the total number of sites. Ranks based on the SSMs were easy to determine because the GPS was previously map matched. To create ranks based on collision data, w_L, the collisions within either a 50 m or 100 m buffer surrounding each link were counted. Similarly, buffers of 100 m or 200 m were used for intersections. Ranks based on vehicles manoeuvres used the same buffer sizes. Although traffic flow SSMs are inherently geared towards analysis at the link level, an attempt was made to quantify correlation at the intersection level. The SSM at each intersection level was determined by averaging the value of the SSM on any intersecting links with data.

Collision Severity: A Kolmogorov-Smirnov test (K-S test) can be used to test equality between two probability distributions. The K-S test is preferred over other statistical techniques because it is nonparametric, requiring no assumption to be made about the probability distributions. The two-sample K-S test is used to compare the empirical cumulative distribution functions, and return the K-S statistic, D, which represents the maximum difference between the two cumulative distribution functions, computed as

$$\begin{aligned} D = max_{1<i<K}\big(E_1(i) - E_2(i)\big) \quad & \text{if } max_{1<i<K}\big(E_1(i) - E_2(i)\big) \\ > min_{1<i<K}|E_1(i) - E_2(i)| & \\ = min_{1<i<K}\big(E_1(i) - E_2(i)\big) \quad & \text{otherwise} \end{aligned} \tag{3-5}$$

where E_1 and E_2 are the empirical cumulative distribution functions of the two samples, and K is the maximum value of observations for which the empirical CDFs are defined. In order to evaluate if the extracted SSMs are statistically linked to collision severity, links and intersections were divided into three groups each; 1) links or intersections with at least one fatal collision; 2) links or intersections with at least one major injury collision, but no fatal collisions, and; 3) links or intersections with only minor injury collisions (the rest). For each function classification, a series of pairwise K-S tests were then performed between the distributions of the site SSMs for each pair of the above severity groups, to determine if any statistical differences exist at different levels of collision severity, to demonstrate the relationship between SSMs and collision severity.

3.4 Results for Event-Based Measures

3.4.1 Extracting Surrogate Safety Measures

Table 3-1 provides the number of vehicle manoeuvres (HBEs and HAEs) identified for each combination of acceleration threshold and filter window size. Both parameters were observed to greatly influence the number of identified vehicle manoeuvres. In the least restrictive case, accelerations and decelerations make up approximately 1 % of the total number of observations. In the most restrictive case, fewer than 500 of each type of vehicle manoeuvre were identified (0.025 % of the total data set). In general, hard braking appears to more common than hard accelerating, which supports previous evidence that braking is the most common evasive manoeuvre in built up areas (*4*). Table 3-1 also shows the number of links and intersections with at least one vehicle manoeuvre for each combination of parameters. Obviously, the number of events or sites decreases with higher thresholds and larger filter window size, as expected. The results of these tables are further illustrated in Figure 3-2. The number of HBEs identified at each intersection is shown in Figure 3-2a, while the data for links is provided in Figure 3-2b. Although many residential links are missing data (because too few trips were made there, if any), the coverage is sufficient to include most of the main highways, arterials, and collectors, as well as downtown Quebec City.

3.4.2 Collision Frequency

The correlation strength with historical collision frequency was calculated using Spearman's rho, for the five functional classifications, three filter window sizes, three acceleration thresholds, and two buffer sizes at both the link and intersection level. This yielded 360 unique test cases, which are summarized in Table 3-2 and in Table 3-3, where the strongest correlation for each functional class is bolded and red. Considering results at the intersection level, correlations between 0.53 and 0.64 were achieved across all functional classes. The performance was best for motorways and residential streets ($\rho > 0.60$). In all cases, the 200 m buffer provided stronger correlations compared to the 100 m buffer (though choosing 200 may be excessive in urban settings). Setting the acceleration threshold to ± 2 m/s^2 also provided the best results in all but one case (HBEs on motorways). Choosing a window size of 3 for the Savitzky-Golay filter also produced the strongest correlations. Although a window size of 3 results in unsmoothed speeds (a polynomial of degree 2 can be fit exactly to three points, therefore the measured speeds are preserved), the acceleration

rate is automatically calculated based on the derivative of the fitted polynomial. This result, that unsmoothed speeds produce the best results, could be due to initial preprocessing of the data. A window size of 5 also produced strong correlations (ρ was reduced by only 0.05 on average) and has the additional benefit of greatly reducing the number of vehicle manoeuvres to analyze (by nearly 50 % at ±2 m/s^2). Although accelerations are less common, the correlation strength is slightly stronger than for decelerations. The overall positive results indicate that intersections that experience a greater number of vehicle manoeuvres also experience a greater number of crashes.

TABLE 3-1 Number of vehicle manoeuvres and number of facilities with vehicle manoeuvres

Number of HBEs and HAEs

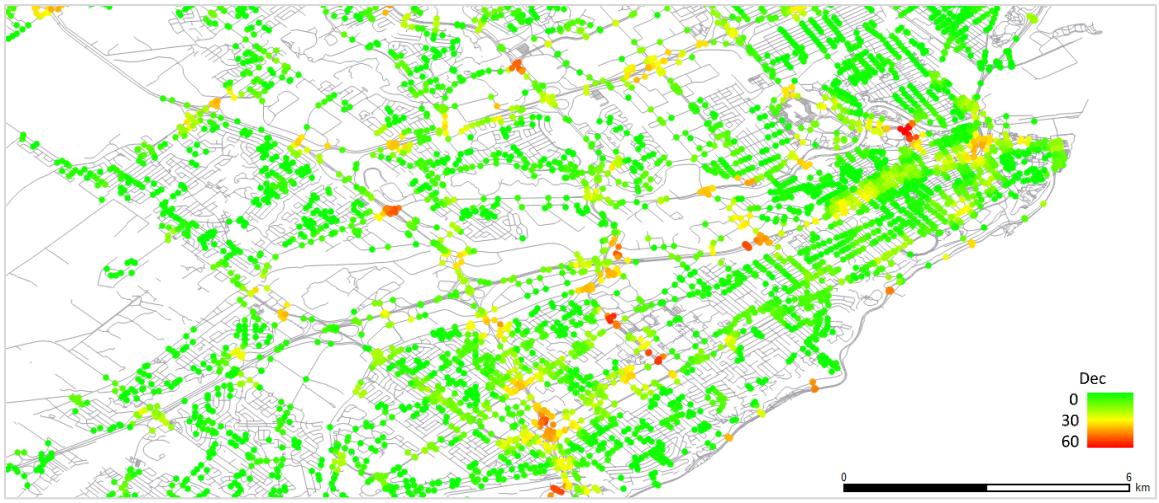
Number of HBEs				Number of HAEs			
	Deceleration Rate (m/s^2)				Acceleration Rate (m/s^2)		
Window	**-2.0**	**-3.0**	**-4.0**	**Window**	**2.0**	**3.0**	**4.0**
3	206788	66032	16725	3	143632	35077	8014
5	115260	16549	2040	5	74753	7725	1257
7	53567	3748	460	7	31565	2047	432

Number of intersections with at least one vehicle manoeuvre (based on 100 m buffer)

	Deceleration Rate (m/s^2)				Acceleration Rate (m/s^2)		
Window	**-2.0**	**-3.0**	**-4.0**	**Window**	**2.0**	**3.0**	**4.0**
3	9210	7717	5530	3	8733	6555	4030
5	8216	4710	1784	5	7476	3667	1336
7	6323	2227	589	7	5758	1810	567

Number of links with at least one vehicle manoeuvre

	Deceleration Rate (m/s^2)				Acceleration Rate (m/s^2)		
Window	**-2.0**	**-3.0**	**-4.0**	**Window**	**2.0**	**3.0**	**4.0**
3	7993	5255	2708	3	7263	4070	1866
5	5868	2192	552	5	5156	1674	465
7	3578	733	151	7	3295	681	187



(a)

(b)

FIGURE 3-2 Maps of the number of HBEs (with the threshold of -2.0 m/s^2, window size of 5 for intersections (a) and HBEs per meter for links (b)

TABLE 3-2 Number of Facilities for Determining Spearman's Rho

	Number of Facilities	
Classification	**Links**	**Intersections**
Motorway	1381	827
Primary	664	364
Secondary	1695	1100
Tertiary	1832	1090
Residential	15014	7340

TABLE 3-3 Spearman's Rho for Vehicle Manoeuvres at the Link and Intersection Levels (by window length, acceleration rate, and buffer size)

Decelerations at the Intersection Level

Classification	Window	100 m -2 m/s²	100 m -3 m/s²	100 m -4 m/s²	200 m -2 m/s²	200 m -3 m/s²	200 m -4 m/s²
Motorway	3	0.372	0.375	0.355	0.576	**0.603**	0.578
	5	0.397	0.319	0.171	0.597	0.555	0.385
	7	0.368	0.287	0.094	0.581	0.414	0.125
Primary	3	0.495	0.471	0.450	**0.540**	0.501	0.462
	5	0.441	0.400	0.250	0.462	0.386	0.252
	7	0.357	0.218	0.030	0.334	0.194	-0.093
Secondary	3	0.429	0.419	0.389	**0.532**	0.512	0.462
	5	0.407	0.324	0.196	0.495	0.378	0.189
	7	0.353	0.195	0.067	0.396	0.186	0.040
Tertiary	3	0.378	0.352	0.311	**0.573**	0.536	0.506
	5	0.344	0.283	0.115	0.522	0.410	0.248
	7	0.312	0.133	0.115	0.430	0.253	0.095
Residential	3	0.412	0.394	0.345	**0.615**	0.598	0.553
	5	0.397	0.319	0.171	0.597	0.555	0.385
	7	0.357	0.174	0.105	0.541	0.298	0.132

Accelerations at the Intersection Level

Classification	Window	100 m 2 m/s²	100 m 3 m/s²	100 m 4 m/s²	200 m 2 m/s²	200 m 3 m/s²	200 m 4 m/s²
Motorway	3	0.387	0.386	0.320	0.618	0.622	0.556
	5	0.393	0.320	0.170	**0.641**	0.576	0.341
	7	0.382	0.229	0.034	0.617	0.449	0.167
Primary	3	0.501	0.482	0.477	**0.554**	0.536	0.536
	5	0.469	0.430	0.306	0.498	0.487	0.425
	7	0.435	0.362	0.204	0.436	0.397	0.200
Secondary	3	0.443	0.409	0.387	**0.536**	0.497	0.445
	5	0.425	0.383	0.314	0.507	0.403	0.276
	7	0.390	0.298	0.230	0.438	0.309	0.194
Tertiary	3	0.404	0.377	0.338	**0.584**	0.543	0.513
	5	0.381	0.339	0.218	0.548	0.481	0.301
	7	0.351	0.257	0.090	0.483	0.360	0.224
Residential	3	0.475	0.440	0.352	**0.625**	0.594	0.506
	5	0.409	0.306	0.156	0.609	0.487	0.283
	7	0.362	0.170	0.062	0.550	0.306	0.167

Decelerations at the Link Level

Classification	Window	50 m -2 m/s²	50 m -3 m/s²	50 m -4 m/s²	100 m -2 m/s²	100 m -3 m/s²	100 m -4 m/s²
Motorway	3	0.046	0.064	0.067	**0.118**	0.117	0.110
	5	0.062	0.063	0.047	0.112	0.100	0.076
	7	0.057	0.069	0.019	0.106	0.092	0.058
Primary	3	0.245	0.245	0.227	**0.260**	0.254	0.218
	5	0.249	0.209	0.090	0.256	0.198	0.062
	7	0.219	0.112	-0.076	0.216	0.074	-0.087
Secondary	3	**0.261**	0.259	0.231	0.254	0.239	0.201
	5	0.252	0.216	0.076	0.240	0.186	0.039
	7	0.230	0.114	0.029	0.216	0.075	0.022
Tertiary	3	**0.213**	0.214	0.196	0.186	0.172	0.150
	5	0.200	0.186	0.096	0.164	0.142	0.071
	7	0.192	0.117	0.073	0.149	0.093	0.050
Residential	3	**0.270**	0.235	0.167	0.225	0.185	0.118
	5	0.239	0.144	0.055	0.191	0.100	0.042
	7	0.198	0.065	0.034	0.140	0.046	0.025

Accelerations at the Link Level

Classification	Window	50 m 2 m/s²	50 m 3 m/s²	50 m 4 m/s²	100 m 2 m/s²	100 m 3 m/s²	100 m 4 m/s²
Motorway	3	0.078	0.084	0.081	0.145	0.136	0.133
	5	0.100	0.067	0.049	**0.155**	0.116	0.087
	7	0.092	0.077	0.004	0.140	0.126	0.064
Primary	3	0.294	0.274	0.236	0.283	0.254	0.199
	5	**0.297**	0.223	0.094	0.272	0.193	0.086
	7	0.267	0.129	-0.016	0.230	0.110	0.002
Secondary	3	**0.333**	0.308	0.241	0.297	0.262	0.208
	5	0.320	0.233	0.148	0.269	0.191	0.123
	7	0.289	0.165	0.069	0.230	0.142	0.053
Tertiary	3	**0.244**	0.239	0.198	0.193	0.192	0.158
	5	0.233	0.182	0.124	0.175	0.149	0.097
	7	0.207	0.140	0.050	0.156	0.110	0.042
Residential	3	**0.256**	0.198	0.145	0.214	0.158	0.112
	5	0.225	0.118	0.052	0.184	0.088	0.039
	7	0.175	0.065	0.037	0.133	0.049	0.022

Considering the link-level results, the first observation is the correlation strength is significantly weaker when compared to intersections ($0.12 \leq \rho \leq 0.33$). Unlike at the intersection level, HBEs and HAEs had the poorest correlation along motorways ($\rho < 0.16$). Again, thresholds of ± 2 m/s^2 and a window size of 3 resulted in the strongest correlations in all but three cases. The 50 m buffer was observed to yield the best results. The overall positive results indicate that links that experience a greater number of vehicle manoeuvres also experience a greater number of crashes, and supports the results found in much of the existing literature (*101*, *38*, *4*). Also, the relatively high correlation strengths indicate that, if vehicle manoeuvres can be identified, they have the potential to be used as SSMs.

3.4.3 Collision Severity

The results of the collision severity testing are summarized in Figure 3-3. The full results of all K-S tests are provided in Table 3-4. Aggregate plots, which contain data for all intersections regardless of classification, are shown first, followed by an example of a single functional classification. These plots are intended as typical examples of the different classes. As the results for links and intersections are substantially similar, Figure 3-3 contains only results for intersections. A complete series of all plots is included in Appendix B. For decelerations at all intersections (shown in Figure 3-3a), it is observed that the distribution of the number of HBEs per site for intersections with minor injury only collisions is shifted to lower values compared to the distribution for intersections with major injury crashes. Furthermore, the distribution of major injury collisions is shifted to lower values compared to intersections with at least one fatal collision. The null hypothesis that the distribution of sites with minor and major injuries (and similarly, major injury and fatal crashes) are similar is rejected by K-S test at 90 % confidence. The same result is observed when considering only secondary class intersections, as illustrated in Figure 3-3b. This pattern, that an increase in the number of HBEs relates to an increase in crash severity, was confirmed to be statistically significant at the 90 % confidence level in 3 out of 10 test cases (5 functional classes at both the link and intersection level) and non-significant in an additional 3. In the remaining 4 test cases, the relative position of CDFs was inconsistent with this result. Similarly, for HAEs, the distribution of the number of HAEs per site for intersections with minor injuries was shifted to lower values than major injuries, which was again shifted to lower values compared to intersections with fatalities. This result was confirmed by K-S test for all

intersections (Figure 3-3c) and secondary intersections (Figure 3-3d). Again, this pattern of increasing severity with increasing number of HAEs was found to be statistically significant at the 90 % confidence level in 3 out of 10 test cases, and non-significant in an additional 3 cases. As before, the remaining 4 cases, the pattern was inconsistent.

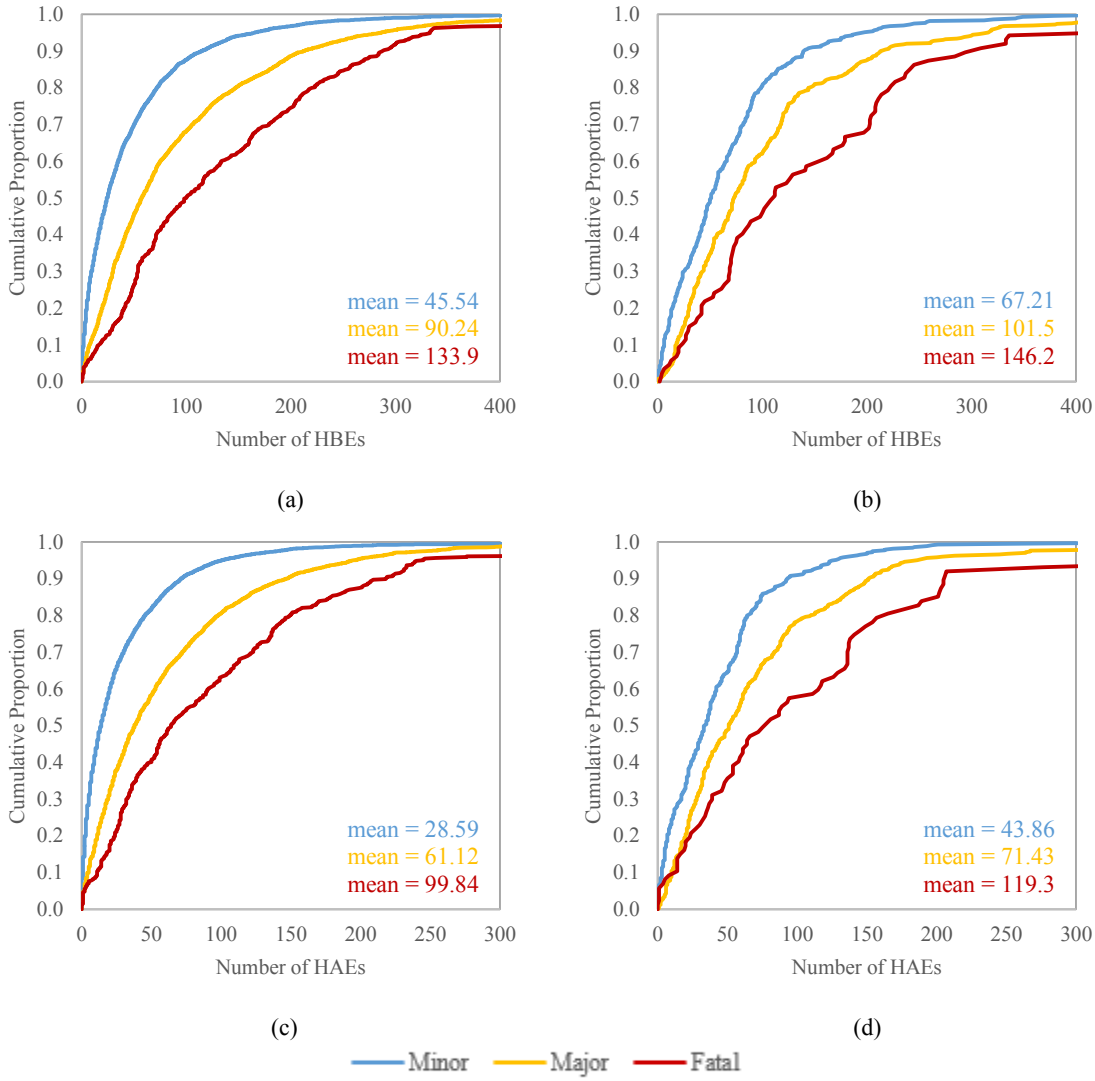


FIGURE 3-3 Cumulative distributions for decelerations, all intersections (a) and secondaries (b), and accelerations, all intersections (c) and secondaries (d) with 200 m buffers

TABLE 3-4 Results of the Pairwise K-S Tests for Crash Severity

INTERSECTIONS		All		Motorway		Primary		Secondary		Tertiary		Residential	
		D	P-value	D	P-value	D	P-value	D	P-value	D	P-value	D	P-value
HBEs	Mi/Ma	**0.251**	**0.000**	**0.305**	**0.000**	**0.408**	**0.000**	**0.184**	**0.000**	**0.181**	**0.000**	**0.271**	**0.000**
	Ma/F	**0.195**	**0.000**	-0.157	0.106	-0.146	0.225	**0.222**	**0.001**	**0.189**	**0.025**	**0.234**	**0.000**
	Mi/F	**0.432**	**0.000**	**0.400**	**0.000**	**0.478**	**0.000**	**0.352**	**0.000**	**0.355**	**0.000**	**0.485**	**0.000**
HAEs	Mi/Ma	**0.268**	**0.000**	**0.417**	**0.000**	**0.430**	**0.000**	**0.195**	**0.000**	**0.221**	**0.000**	**0.287**	**0.000**
	Ma/F	**0.184**	**0.000**	-0.234	0.007	-0.123	0.344	**0.221**	**0.001**	**0.216**	**0.008**	**0.227**	**0.000**
	Mi/F	**0.425**	**0.000**	**0.268**	**0.000**	**0.480**	**0.000**	**0.364**	**0.000**	**0.400**	**0.000**	**0.478**	**0.000**

LINKS		All		Motorway		Primary		Secondary		Tertiary		Residential	
		D	P-value	D	P-value	D	P-value	D	P-value	D	P-value	D	P-value
HBEs	Mi/Ma	**0.087**	**0.000**	0.084	0.225	0.066	0.378	**0.069**	**0.051**	0.062	0.137	**0.071**	**0.000**
	Ma/F	0.047	0.204	-0.958	0.462	0.069	0.690	-0.066	0.581	0.107	0.338	0.100	0.133
	Mi/F	**0.121**	**0.000**	0.053	0.735	0.124	0.303	0.089	0.348	0.149	0.100	**0.164**	**0.003**
HAEs	Mi/Ma	**0.117**	**0.000**	0.084	0.225	**0.102**	**0.098**	**0.119**	**0.000**	**0.109**	**0.002**	**0.084**	**0.000**
	Ma/F	**0.059**	**0.077**	-0.096	0.462	-0.117	0.346	0.077	0.473	0.123	0.239	0.083	0.252
	Mi/F	**0.174**	**0.000**	0.053	0.735	0.102	0.449	**0.168**	**0.024**	**0.202**	**0.014**	**0.159**	**0.004**

Notes: *Statistically significant values at 90 % confidence are in red. Crash severity levels have been abbreviated as Mi (minor), Ma (major), and F (fatal). A positive D statistic means the distribution of the first member of the pair is shifted to lower values, while a negative D statistic means the second of the pair is shifted to lower values.*

3.5 Results for Traffic Flow Measures

3.5.1 Extracting Surrogate Safety Measures

Based on the results for the event-based measures, the analysis of traffic flow measures set the parameters of the speed filter (window size of 5, degree of 2) and buffer size (50 m for links, 100 m for intersections) accordingly. At the link level, there was sufficient data to compute each of the three SSMs on 4912 links. Though the number of links was equal for all three measures, it is not required that a single link be represented in all three data sets. CI requires both peak and off-peak data to be available, but accepts 0 as a legitimate value, whereas the others require only off-peak observations but must be greater than 0. Figure 3-4 illustrates the network coverage for the three considered SSMs. Although many residential links are missing data (they experienced few, if any, trips), the coverage is sufficient to include most of the main highways, arterials, and collectors, as well as fair coverage in downtown Quebec City. When aggregating data to the intersection level, CI was evaluated at 4540 intersections, while for $\bar{V}$ and CVS, data was available at 4818 intersections. Linear correlations between the measures are provided in Table 3-5.

TABLE 3-5 Linear correlations between Traffic Flow SSMs at the Link and Intersection Level

	Link Level				**Intersection Level**		
	CI	**$\bar{V}$**	**CVS**		**CI**	**$\bar{V}$**	**CVS**
CI	1.00	0.28	-0.17	**CI**	1.00	0.27	-0.16
$\bar{V}$	-	1.00	-0.57	**$\bar{V}$**	-	1.00	-0.59
CVS	-	-	1.00	**CVS**	-	-	1.00

3.5.2 Collision Frequency

The average values and correlation strength with collision frequency for the three SSMs are provided in Table 3-6. CI was positively correlated with crash frequency at both the link and intersection levels for all functional classifications. This result (that increased congestion leads to more crashes) supports results found in several past studies (*94*, *41*, *89*). The indicator performed best on primary streets at the link level ($\rho = 0.21$) and performed better at the link level in general. Compared to the other SSMs, CI had the lowest correlation strength, and performed very poorly on motorways. In contrast, $\bar{V}$ showed the strongest correlation with crash frequency, although the direction of the correlation was consistently negative across functional classes.

(a)

(b)

(c)

FIGURE 3-4 Maps of congestion index (a), average speed (b), and coefficient of variation of speed (c)

TABLE 3-6 Averages and Spearman's rho For Traffic Flow SSMs at the Link and Intersection Level

Average Values for CI, $\bar{V}$ (km/h), and CVS

Link Level				Intersection Level			
Classification	**CI**	**$\bar{V}$**	**CVS**	**Classification**	**CI**	**$\bar{V}$**	**CVS**
Motorway	0.116	75.52	0.199	Motorway	0.118	75.39	0.195
Primary	0.104	45.74	0.326	Primary	0.103	43.50	0.338
Secondary	0.102	43.94	0.305	Secondary	0.102	42.97	0.308
Tertiary	0.087	38.84	0.314	Tertiary	0.086	38.42	0.320
Residential	0.065	30.74	0.347	Residential	0.083	34.14	0.334

Spearman's rho for CI, $\bar{V}$ (km/h), and CVS

Link Level				Intersection Level			
Classification	**CI**	**$\bar{V}$**	**CVS**	**Classification**	**CI**	**$\bar{V}$**	**CVS**
Motorway	0.05	-0.27	0.17	Motorway	0.02	-0.14	0.20
Primary	0.21	-0.35	0.16	Primary	0.18	-0.45	0.38
Secondary	0.11	-0.41	0.10	Secondary	0.11	-0.37	0.36
Tertiary	0.12	-0.22	0.16	Tertiary	0.15	-0.18	0.20
Residential	0.08	0.05	0.15	Residential	0.09	0.00	0.13

This result implies that, within a given roadway class, links and intersections with higher FFS have fewer crashes than those with lower speeds. The indicator performed well on primary, secondary, and tertiary facilities ($-0.50 \leq \rho \leq -0.30$) but performed poorest on residential streets ($\rho \leq -0.20$). Although the negative correlation opposes most existing literature (*121*), several things should be noted. Existing literature focusses largely on mean speed, while this work considers FFS. Besides functional class, there is no controlling for other factors which may themselves influence speed (geometry, distance from the city centre, volumes etc.). This study considers the network scale, while many previous studies considered individual corridors or links. The time scale of analysis may also be important. CVS was consistently positively correlated with crash frequency, and generally performed well on all facility types ($0.20 \leq \rho \leq 0.40$), except for residential facilities ($\rho < 0.20$). This result supports the specific findings by Lee et al. (*9*) that increased CVS is associated with more collisions, and the general findings by other authors that speed variation and crash frequency are positively linked (*41*, *7*, *117*). In general, correlation strengths were weak to moderate, and traffic flow measures perform poorer compared to the proposed event-based measures.

3.5.3 Collision Severity

The results of the collision severity testing are summarized in Figure 3-5. Aggregate results for all links are provided in the left column, and an example for one functional class is provided in the right. All plots are at the link level, as the results for intersections were substantially similar. The full results of all K-S tests are provided in Table 3-7. With regards to CI, the distributions for links (shown in Figure 3-5a) and intersections with minor injury and fatal collisions was shifted to significantly lower values than links with major injury collision (by K-S test at 90% confidence). This pattern was verified for several different functional classifications such as motorways, as illustrated in Figure 3-5b. In fact, this was the most common pattern for CI, being observed to be statistically significant (at 90 % confidence) in 4 out of 10 test cases (5 classes at both the link and intersection levels), and non-significant in 4 additional cases (see Table 3-7). This result suggests that congestion could be used to define two categories: high congestion, which have more major injury crashes, and low congestion which can have either minor injury or fatal crashes, as decided by other factors. Importantly, this result highlights the complex relationship between congestion and crash severity.

For $\bar{V}$, no statistically significant difference was found in the distributions for all links (Figure 3-5c) or all intersections. When dividing the data by functional classifications, no consistent patterns were observed. For example, on secondary links (shown in Figure 3-5d), the distribution of speed for links with major injuries was shifted to lower values compared to distributions for those with minor and fatal collisions (by K-S test at 90% confidence). Decreasing severity with increasing speed was observed to be significant in only two cases. FFS shares, at best, a weak relationship with crash severity at the network level in this data set. CVS had the strongest relationship with crash severity. The cumulative distributions of all links are provided in Figure 3-5e. For all links and intersections, the distribution of links with major injuries was shifted towards higher values compared to the distribution for links with major injuries (by K-S test at 90% confidence). Furthermore, the distribution for links and intersections with fatal crashes was shifted to higher values compared to both other distributions (at 90% confidence). This pattern was observed to be statistically significant in 5 out of 10 test cases, and non-significant in an additional 4 (see Table 3-7). One of the clearest examples of this pattern was for tertiary links, provided in Figure 3-5f. In general, higher values of CVS were related to increased crash severity for most functional classes.

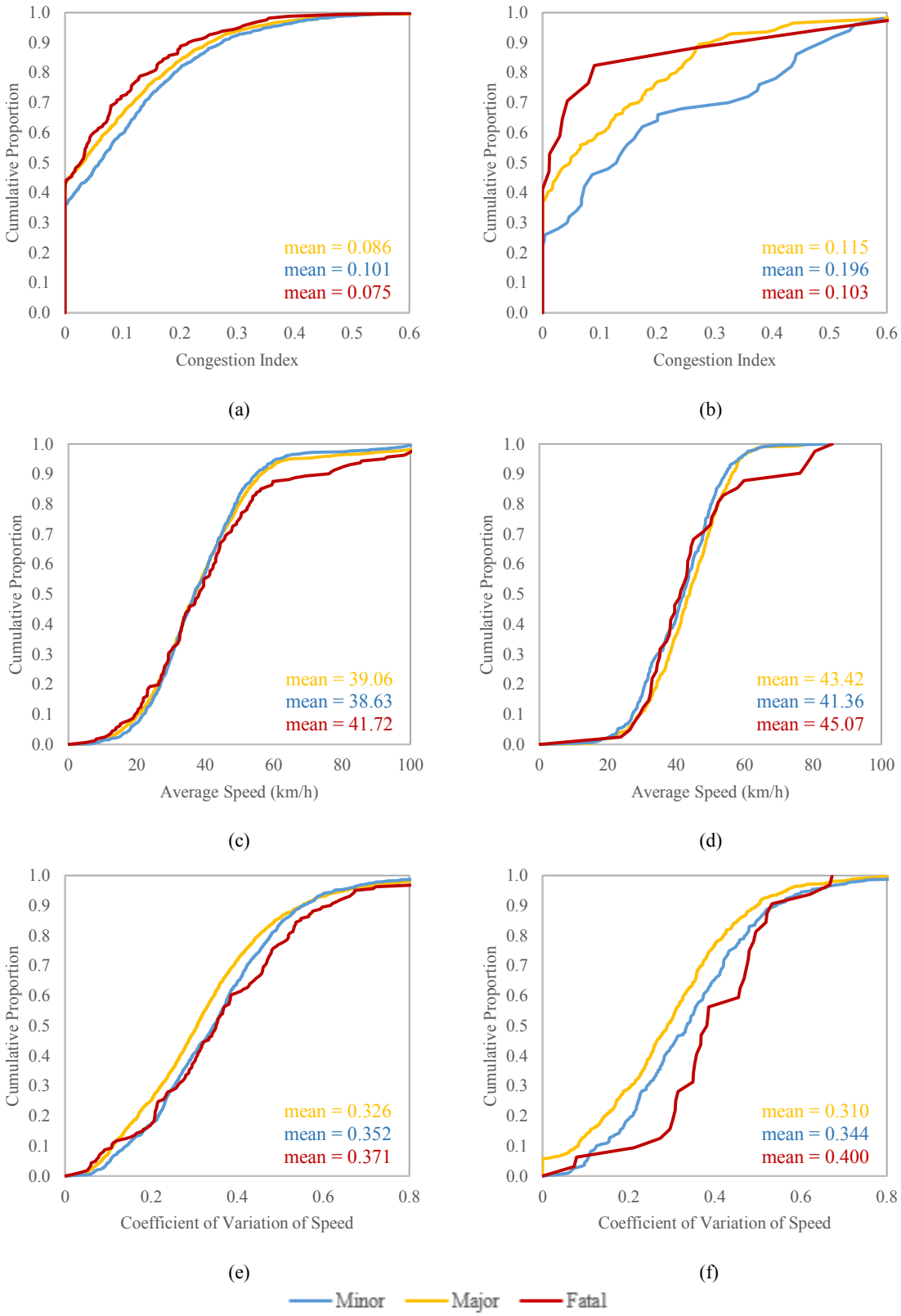


FIGURE 3-5 Cumulative distributions for CI, all links (a) and motorways (b), $\bar{V}$, all links (c) and secondaries (d), and CVS, all links (e) and tertiaries (f)

TABLE 3-7 Results of the Pairwise K-S Tests for Crash Severity

LINKS		All		Motorway		Primary		Secondary		Tertiary		Residential	
		D	P-value	D	P-value	D	P-value	D	P-value	D	P-value	D	P-value
CI	Mi/Ma	**0.086**	**0.000**	**0.244**	**0.010**	**0.183**	**0.023**	-0.060	0.289	0.054	0.413	0.049	0.417
	Ma/F	**-0.136**	**0.006**	**-0.406**	**0.015**	**-0.297**	**0.021**	-0.068	0.720	-0.108	0.521	**-0.231**	**0.040**
	Mi/F	-0.071	0.219	-0.247	0.152	-0.229	0.102	-0.101	0.464	-0.082	0.670	**-0.199**	**0.070**
V	Mi/Ma	-0.031	0.288	-0.167	0.116	**-0.250**	**0.001**	**-0.111**	**0.015**	**-0.127**	**0.007**	-0.078	0.109
	Ma/F	0.090	0.108	0.253	0.197	0.215	0.132	0.119	0.365	**-0.254**	**0.027**	0.124	0.393
	Mi/F	0.073	0.201	0.206	0.270	-0.125	0.507	-0.139	0.237	**-0.282**	**0.008**	-0.129	0.325
CVS	Mi/Ma	**0.107**	**0.000**	0.124	0.308	**0.215**	**0.006**	**0.175**	**0.000**	**0.105**	**0.034**	**0.137**	**0.001**
	Ma/F	**0.093**	**0.095**	-0.273	0.151	-0.147	0.388	-0.147	0.212	**0.295**	**0.008**	**0.238**	**0.032**
	Mi/F	**0.144**	**0.002**	-0.235	0.181	0.152	0.366	**0.181**	**0.087**	**0.349**	**0.001**	**0.237**	**0.023**

INTERSECTIONS		All		Motorway		Primary		Secondary		Tertiary		Residential	
		D	P-value	D	P-value	D	P-value	D	P-value	D	P-value	D	P-value
CI	Mi/Ma	**0.109**	**0.000**	**0.119**	**0.082**	**0.262**	**0.001**	**0.116**	**0.006**	**0.155**	**0.000**	**0.090**	**0.000**
	Ma/F	**-0.070**	**0.094**	-0.156	0.218	**-0.227**	**0.065**	**-0.139**	**0.096**	0.154	0.136	-0.056	0.460
	Mi/F	**0.092**	**0.013**	0.113	0.389	0.177	0.225	0.124	0.164	**0.274**	**0.001**	0.092	0.111
V	Mi/Ma	**-0.065**	**0.000**	-0.106	0.133	**-0.268**	**0.000**	**-0.143**	**0.000**	**-0.142**	**0.001**	**-0.094**	**0.000**
	Ma/F	-0.056	0.212	-0.114	0.436	-0.141	0.320	-0.110	0.227	**-0.213**	**0.022**	-0.093	0.108
	Mi/F	**-0.109**	**0.002**	**-0.194**	**0.058**	**-0.279**	**0.018**	**-0.217**	**0.004**	**-0.313**	**0.000**	**-0.150**	**0.002**
CVS	Mi/Ma	**0.102**	**0.000**	**0.193**	**0.001**	**0.228**	**0.003**	**0.126**	**0.002**	**0.092**	**0.050**	**0.091**	**0.000**
	Ma/F	**0.106**	**0.004**	0.115	0.427	-0.114	0.479	**0.155**	**0.054**	**0.208**	**0.026**	**0.186**	**0.000**
	Mi/F	**0.170**	**0.000**	**0.231**	**0.017**	**0.313**	**0.006**	**0.233**	**0.002**	**0.269**	**0.002**	**0.215**	**0.000**

Notes: *Statistically significant values at 90 % confidence are bolded. Crash severity levels have been abbreviated as Mi (minor), Ma (major), and F (fatal). A positive D statistic means the distribution of the first member of the pair is shifted to lower values, while a negative D statistic means the second of the pair is shifted to lower values.*

3.6 Conclusions

The purpose of this chapter was to examine potential SSMs extracted from the GPS-enabled smartphones of regular drivers, and to correlate these measures with historical collision frequency and severity at the network scale. The statistical relationship between SSMs and collision frequency was determined using Spearman's rank correlation coefficient. Relationships with crash severity were analyzed using pairwise K-S tests. Both HBEs and HAEs were shown to be positively correlated with frequency. Correlations between 0.53 and 0.64 were observed, dependent on the functional classification at the intersection level. Results at the link level were worse, with correlations between 0.12 and 0.33 depending on functional classification. These vehicle manoeuvres are more strongly correlated with collision frequency at or around intersections, making HBEs and HAEs more capable of identifying dangerous intersections than identifying dangerous links. Furthermore, an increase in either HBEs or HAEs was shown to be related to an increase in crash severity. In general, the distributions of the number of vehicle manoeuvres at links and intersections that have experienced at least one fatal crash were shown to be shifted to higher values compared to links or intersections with, at worst, major injury collisions, which were in turn shifted higher than facilities with only minor injury collisions (though this relationship was not statistically significant in some cases). As these results indicate, not only are vehicle manoeuvres related to a greater number of crashes, but more braking and accelerating may also be related to increased collision severity (at least for some functional classes).

Congestion was shown to be positively correlated with crash frequency. According to the data in this study, links and intersections having higher levels of evening congestion also tend to have a greater number of collisions. This result supports findings in existing literature. However, CI had the poorest correlation strength of the three considered SSMs. When considering collision severity, the relationship with CI was found to be non-monotonous. The distributions of links with, at worst, major injury crashes were found to be shifted towards higher values compared to links with minor or fatal collisions. This complex relationship can be potentially explained considering the relationship between flow, density, and speed. In general, fatal crashes are rare. Although scenarios with the lowest CI levels have the high speeds required to increase crash severity, the low vehicle densities make the occurrence of a fatal crash exceedingly rare. Therefore, injury crashes are much more likely. As density initially increases, the number of vehicular interactions increases (expected number of crashes) and speeds remain high (severity of crash). This increase

in density at the beginning of the fundamental diagram creates the environment necessary to produce fatal crashes, where high severity (due to speed) and high probability (due to density) coexist. This, along with other geometric factors could explain why uncongested facilities are split into two groups: minor injury only and fatal. As congestion continues to increase, the number of conflicts and crashes also increases, but speed decreases rapidly. Although the high number of interactions is associated with more crashes, speed reduction reduces the probability of fatal crashes.

Speed was found to be negatively correlated with crash frequency (in fact, the strongest correlations of all proposed traffic flow SSMs) and had no conclusive statistical relationship with crash frequency. There are two possible explanations for these results. First, the only factor controlled for in the analysis was functional class of the roadway. Additional features which themselves are related to speed, such as geometry and traffic volumes, were not considered. If the speed measure is correlated with another factor with a causal relationship to lower crash frequency, this would mask the true effect of speed on safety. A second possible explanation is the scale of the analysis. To date, most studies considering speed as an indicator of risk have done so using a single link or corridor. As has been shown in existing literature, it is clear that for a single link, increasing speed should increase crash severity. However, at the network scale, little if any work has been done, and it is possible that the relationship between speed and crash frequency and severity is different.

Speed uniformity was observed to be positively correlated with crash frequency, and statistically related to increased crash severity. According to the data utilized in this study, links and intersections with more speed variation experience, not only a greater number of crashes, but also more severe crashes. High CVS implies speed variation across both space (vehicles of different speeds interacting) and time (changing traffic conditions). This could mean that traffic flow is more complex, with more maneuvering, creating more opportunities for conflicts and crashes. High relative speed differences between conflicting vehicles could also lead to more major and fatal crashes, compared to facilities with less variation in speed. This result supports several past studies, which identified variation in speed as an important predictor of risk. In general, the strength of the correlations with respect to crash frequency is weak to moderate. Traffic flow SSMs may be stronger indicators of crash severity than crash frequency.

CHAPTER 4: MODELLING CRASH FREQUENCY AND SEVERITY

"Far better an approximate answer to the right question …
than an exact answer to the wrong question"
John Tukey, Statistician

4.1 Introduction

Most traditional screening techniques use statistical (regression) models, safety performance functions, and Bayesian statistics (*122*, *49*) to estimate the expected number of crashes at each location in the road network based on historical crash data and factors related to traffic, geometry, and environment (*51*). Through these techniques, the risk factors contributing to crash occurrence can also be uncovered (*51*) and various risk measures can be derived, including the posterior probability of excess and posterior of ranks among others (*123*, *124*). Vision Zero suggests that the focus of attempts to improve road safety should be on reducing injuries and fatalities. Although reducing total crashes is a worthy cause, efforts to reduce the most dangerous crashes are the most efficient way to improve safety (*51*). Crash severity has been incorporated into crash models using various techniques (*3*). Common frequency models may be combined with injury severity models which are "conditional on the crash having occurred" (*49*), though these methods are data heavy, requiring detailed crash characteristics from each crash in order to predict severity. Multivariate Bayesian models (*64*) have also been explored for crash modelling, though large-scale estimation is time-intensive. By relying on ranking criteria derived from historical crash data (*59*), existing crash modelling techniques are intrinsically reactive (*4*), and crash-based network screening is often performed periodically (once every few years) so that crashes can accrue and databases can be updated, rather than continuously. There is an opportunity to evolve surrogate safety analysis from a site-specific to a network-wide approach, potentially using smartphone GPS data from instrumented bicycles as explored by Strauss et al. (*125*). Although such methods rely on crash data for calibration, the application of models to monitor safety depends only on probe data that is continuously available (if the relationship between the data and crashes remains constant).

Though Bayesian methods are popular in crash-based screening models, estimation by MCMC simulation can be computationally expensive. Though not typically a concern for safety models, as considered networks become larger, models contain more complex spatial and temporal correlations, and volumes of probe data grow to billions of observation, any method for reducing computation time will become extremely valuable. Despite recent advances in Bayesian inference, very few studies to date have applied the INLA approach to the field of road safety. The purpose of this chapter, published in two studies (*126*, *127*) is to propose a network screening approach based on SSMs derived from GPS data and using a mixed-multivariate model for crash frequency and severity. A Full Bayesian Spatial Latent Gaussian Model (LGM) of crash frequency is

calibrated using the R-INLA program, and an FMNL Model is estimated for crash severity. Site rankings developed using this model and a traditional crash-based approach are then compared.

4.2 Literature Review

Common regression techniques for crash modelling implement statistical count models such as Poisson models (*52*), NB models (*53*), and Zero-inflated Poisson models (*52*), which though theoretically appropriate for crash modelling, have assumptions that may cause misinterpretations of results (*53*). More advanced regression models incorporate random effects, multivariate outcomes, and hierarchical structures (*50*). In terms of crash severity, conditional models require detailed environment, roadway, user, and vehicle information for crashes that have already occurred. Early binary models with two levels of injury severity have evolved to include multiple discrete outcome models, whether unordered (multinomial and nested logit models) or ordered (ordered probit and logit models) (*52*). In contrast to conditional models which use detailed crash-level data, some have proposed severity models based on aggregate, site-level, geometric, traffic, and environmental data. Anastopoulos and Mannering (*49*) estimated the proportion of collisions at each severity level using an aggregate model that, although had a poorer goodness-of-fit, was comparable to a conditional model in terms of identifying hotspots.

In regression models, estimated coefficients take fixed values. In Bayesian models, coefficients are defined by a probability distribution (*56*). EB models, popular in the 1980s and 1990s, fix some parameters of the model based on observed crash data (*57*) instead of using hyper-priors. Hauer (*128*) described the EB process, noting that the safety of a site is described by both its characteristics and historical crash record, and presented applications of the model for estimating crashes in the US and Ontario, Canada. Mountain, Fawaz, and Jarret (*129*) applied the EB technique to a series of at grade crossings in the UK, showing an improvement over naïve regression models. FB techniques for complex problems (such as non-conjugate models including LGMs) typically determine the posterior distributions by first assuming a prior distribution and then iteratively computing and updating the posterior marginal using MCMC simulation, though this process is time and resource intensive (*58*). Examples of crash frequency modelling using an FB model are found in El-Basyouny and Sayed (*130*) and Aguero-Valverde and Jovanis (*131*). More complex FB models found in the existing literature include models which incorporate random effects (*57*, *130*) and/or spatial correlations (*132*, *64*) which significantly improve model

accuracy. Several examples of univariate FB models approximated using the INLA technique can be found in Hu et al. (*133*) and Serhiyenko et al. (*134*), both of which used Bayesian models to study the temporal trends of road traffic crashes. Crash severity can be integrated with Bayesian methods using either two-step or multivariate models. For example, Wang, Quddus, and Ison (*63*) developed a two-step model, using an FB spatial model to estimate frequency and an unordered nominal response model to determine the proportion by severity type. Multivariate Bayesian models were estimated by Miaou and Song (*64*), Aguero-Valverde and Jovanis (*5*), and Park and Lord (*65*) to simultaneously estimate the number of collisions at several injury severity levels. Multivariate FB models have also been approximated using INLA (*135*).

Despite their ever-increasing use in road safety management, SSMs have rarely been integrated into statistical models for network screening purposes. Most existing studies have only considered traffic conflicts and have required microsimulation to build network-wide data sets. For example, both Ariza (*136*) and Lorion (*137*) demonstrated that adding conflicts as an explanatory variable could improve estimates of crash frequency at intersections using a generalized linear regression model based on simulated data. Li et al. (*138*) used video-extracted conflicts to estimate a multivariate linear regression model at a single freeway interchange. Though several studies have demonstrated the potential for extracting from probe vehicle data both event-based surrogate measures, individual driver manoeuvres including steering, braking, or accelerating (*139*, *4*, *29*), and traffic flow SSMs, including speed or speed variation (*117*, *118*) and congestion (*94*), very few studies have incorporated such SSMs into screening models. At the time this thesis was written, only exploratory work by Kluger (*140*) could be found.

Shortcomings in the existing literature generally fall at the intersections of crash modelling, Bayesian inference, and SSMs. First, although FB techniques are the most accurate and well-accepted approach for crash modelling, current MCMC simulation is computationally expensive and time consuming. This study takes advantage of recent advances in Bayesian inference, namely INLA, as a state-of-the-art method to solve a complex problem in the field of road safety. Second, although crash models continue to improve, all existing approaches are crash-based. SSMs are under continuous development and can be extracted from probe vehicle data for the entire network. Yet, almost no studies to date have incorporated SSMs into statistical crash models. This chapter aims to address these gaps by developing a mixed multivariate model for crash frequency and severity by incorporating INLA and SSMs extracted from the smartphones of regular drivers.

4.3 Methodology

4.3.1 Modelling Crash Frequency with Latent Gaussian Models

LGMs are a subclass of structured additive models, in which the response variable y_i (in this case, crash frequency) for each subject i is assumed to follow a distribution from the exponential family (Normal, Poisson, or Binomial). This model can be written as a structured additive model, in which the mean of y_i, noted μ_i, at site i is related to the predictors through a link function $g()$ such that

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \epsilon_i \tag{4-1}$$

where β_0 is the intercept, β_k are the coefficients (up to the total number of coefficients, n_β) representing the linear effect of covariates z_{ki} (in this case, SSMs, trip counts, and roadway functional classification), $f^{(j)}$ are functions of covariates u_{ji} used to relax these linear relationships or introduce random effects (up to the total number of functions, n_f) , ϵ_i is the unstructured (i.e. containing no spatial or serial structure) error component, and η_i is the structured additive predictor (*58*). LGMs are extremely flexible because of the forms that $f^{(j)}$ can take, incorporating temporal dependences or spatial correlations that are critical for reducing biases in crash modelling (*141*, *56*).

The likelihood of an LGM can be represented as a three-stage hierarchical structure, beginning with the conditionally independent likelihood function:

$$\pi(y|x,\theta) = \prod_{i=1}^{n} \pi(y_i|\eta_i(x),\theta) \tag{4-2}$$

where y is the response vector, x is the latent field of covariates, θ is the vector of hyperparameters, and $\eta_i(x)$ is the ith additive predictor. Next, the latent Gaussian field is formally defined with a mean $m(\theta)$ and precision matrix $Q^{-1}(\theta)$ conditioned on hyperparameters θ:

$$x|\theta \sim \mathcal{N}(m(\theta), Q^{-1}(\theta)) \tag{4-3}$$

Finally, a prior distribution is assigned to the hyperparameters:

$$\theta \sim \pi(\theta) \tag{4-4}$$

Although the prior distribution of the latent field must be Gaussian by definition, the prior distributions of the hyperparameters are not subject to this constraint. For more details on LGMs, readers are referred to Rue, Martino, and Chopin (*58*).

4.3.2 Spatial Correlations using the Besag Proper Model

Previous research has highlighted the need to account for spatial correlations, which can be incorporated to account for similarities of adjacent links or intersections (*56*) whether the origins of these similarities are known or unknown (*141*). Failure to account for spatial dependence may lead to model biases (*141*, *56*). In this study, the spatial component is accounted for using a modified version of the Besag–York–Mollié (BYM) model presented by Besag et al. (*142*). Assuming the outcome y_i follows a Poisson distribution (note that the same model can be written using a Negative Binomial (NB) distribution), the model is written as:

$$y_i \sim Poisson(\mu_i) \tag{4-5}$$

$$\text{where} \quad log(\mu_i) = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k z_{ik} + u_i + v_i \tag{4-6}$$

where β_k are regression coefficients (up to the total number of coefficients, n_β) for fixed effects, z_{ik} are covariates, v_i is a site-specific random effect modeled using an exchangeable correlation structure across sites, and u_i is another site-specific random effect modeled as spatially structured. Several structures can be specified for $u = (u_1, \dots, u_n)$. Here, a proper version of the conditional autoregressive (CAR) structure is chosen:

$$u_i | u_{-i}, \tau, d \sim \mathcal{N}\left(\frac{1}{d + n_i}\sum_{i \sim j} u_j, \frac{1}{\tau(d + n_i)}\right) \tag{4-7}$$

where n_i is the number of neighbours for link or intersection i, u_{-i} represents all members of u excluding the ith, $i \sim j$ indicates if links or intersections i and j are neighbours, $d > 0$ is a weight

parameter, and $\tau > 0$ is a scaling parameter (*56*). Python scripts were used to generate graphs describing the network topology and define neighbours as links or intersections which are immediately adjacent to one another. For links to be considered neighbours, they must share a common intersection (the links must intersect), while intersections must share a common link (be connected by a link). This is illustrated further in Figure 4-1.

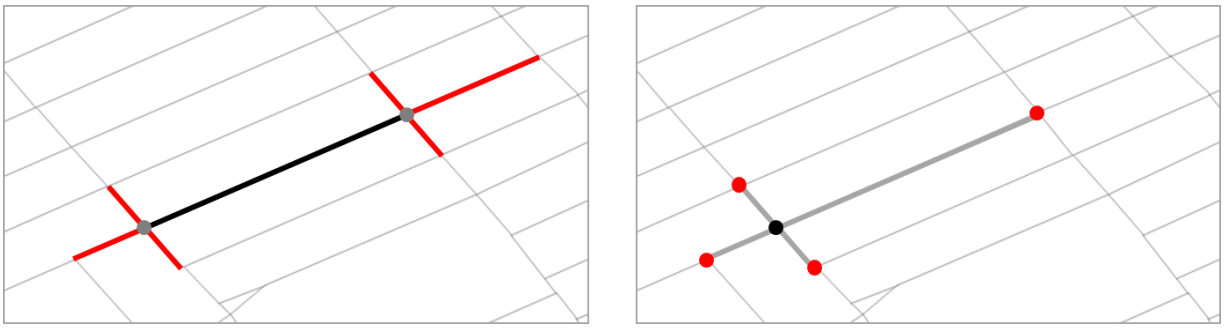


FIGURE 4-1 Illustration of a site (black) and its neighbours (red) for links (a) and intersections (b)

4.3.3 Integrated Nested Laplace Approximation

The greatest challenge with Bayesian models is estimating the posterior distributions of the latent field. Traditionally, MCMC simulations have been used to iteratively compute and update the posterior marginal of the latent field. However, this process is time and resource intensive, requiring tens of thousands of iterations (*58*). This shortcoming has led to both new techniques for MCMC simulation (including Hamiltonian Monte Carlo (*143*)) and new developments in Bayesian inference. To reduce the time required to estimate LGMs using MCMC techniques, this work proposes the implementation of the INLA technique for Bayesian approximation, which has been shown to produce accurate approximations with a significant reduction in computational time (*144*). The INLA approach was proposed by Rue, Martino, and Chopin (*58*) to perform Bayesian approximations on LGMs using a combination of Laplace approximations and numerical integration to estimate the posterior marginal of the latent field. Besides a significant reduction in computational time (*144*) from a range of days to a range of hours, the approximation error in the INLA approach is nearly equal to the estimation error in typical MCMC methods (*58*). The necessity of incorporating a spatial component and its associated complexity makes INLA

"particularly suitable in this context" (*56*). A package for the programming language R has been developed to easily deploy INLA (R-INLA) (*144*).

4.3.4 Modelling Crash Severity using Discrete Choice Models

To integrate crash severity modelling into the above approach, the method for estimating a mixed multivariate model presented by Wang et al. (*63*) is adopted. The mixed multivariate outcome is estimated using two models. First, crash counts, or frequency, are estimated using the presented Spatial LGM. Second, crash severity is integrated through a discrete choice model. At each site, the proportion of crashes at each severity level (fatal, major injury, and minor injury) are modelled using the fractional Multinomial Logit (FMNL). The probability for a crash at a given severity level m for link or intersection i is

$$P_i(m) = Pr(U_{mi} \geq U_{li} \; \forall l) \tag{4-8}$$

$$P_i(m) = \frac{exp(U_{mi})}{\sum_{\forall l} exp(U_{li})} \tag{4-9}$$

where U_{mi} is the utility associated with severity level m at site i. The utility for alternative m can be written as the sum of a deterministic component, V_{mi}, and a random error component, ϵ_{mi}

$$U_{mi} = V_{mi} + \epsilon_{mi} \tag{4-10}$$

where V_{mi} can be further decomposed as

$$V_{mi} = \alpha_{m0} + \sum_{k=1}^{n} \alpha_{mk} z_{ki} \tag{4-11}$$

where α_{m0} is the intercept, and α_{mk} are the coefficients of covariates z_{ki} (including SSMs and roadway functional classification) for severity level m. In the FMNL, rather than restricting the observed choice to 0 or 1, observations may vary as a fraction between 0 and 1, with the only constraint being that for any link and intersection, observations across all severity levels must sum to exactly 1. In this model, minor injury crash is taken as the reference level.

4.3.5 Site Ranking

Once both models are estimated, the results must be combined so that sites can be ranked based on the estimated number of crashes at each severity level. To compare objectively across sites, Wang et al. (*63*) propose the use of a decision parameter combining estimated crash counts systematically for each site under investigation. The authors note that the choice of the decision parameter is context dependent, and can take several forms, including expected crash frequency, rate, or economic cost. Herein, the cost of a crash at the various severity levels can be computed and used as a decision parameter. The total cost of crashes at each site, θ_i, is computed as

$$\theta_i = \sum_m \mu_i \cdot P_i(m) \cdot C(m) \tag{4-12}$$

where, as defined before, μ_i is the mean expected number of crashes and $P_i(m)$ is the probability of a crash at severity level m. Then, the product $\mu_i \cdot P_i(m)$ yields the crash counts at each severity level and is multiplied by $C(m)$, the relative cost of a crash at severity level m. In this case, $C(m)$ was chosen as the relative cost of a crash at each severity level according to Transport Canada (*145*). Minor injury crashes were assigned cost of 1, major injury crashes a cost of 10, and fatalities a cost of 160. A link or intersection with a higher value of θ_i has a higher crash cost, is considered more hazardous and should therefore be given a higher priority in the network screening process. Alternatively, sites can be ranked by the cost per vehicle-km, δ_i, computed as

$$\delta_i = \frac{\theta_i}{t_i \cdot l_i} \tag{4-13}$$

where t_i is the number of observed GPS traces at each site (used as a proxy of traffic exposure) and l_i is the length of the site (intersections are assigned a length of 1). Site rankings based on δ_i are determined first using the crash data and then using the outcome of the mixed multivariate model. In the case of sites with zero observed crashes, the ranking is determined using

$$\delta_i = \frac{1}{100 \cdot t_i \cdot l_i} \tag{4-14}$$

This prevents a high number of low duplicate rankings for the crash data, while ensuring sites with zero observed crashes are ranked below those with at least one observed crash. The crash-based and modelled rankings are compared using Spearman's rank correlation. Hotspots identified using the crash data and the safety model are compared using the percent deviation, calculated as

$$\%\ \text{deviation}_r = 100 \cdot\ (1 - {}^{\kappa}/_{r}) \tag{4-15}$$

where κ is the number of locations found in hotspot lists determined both by the crash-based and surrogate safety modelling methods and r is the number of hotspots considered (the list size) (*146*).

4.3.6 Model Calibration and Validation

While the calibrated models demonstrate the association of each SSM on crashes, their fit overstates predictive power. Calibration and validation of the crash prediction model was completed in several distinct steps. Using the INLA approach, Poisson and NB Bayesian models were first estimated at both the link and intersection levels for a subset of the road network. These models use the observed number of crashes as the outcome, y, the extracted SSMs, number of GPS trips (used as a proxy of exposure, or volume), and the roadway length and functional classification as independent variables. Spatial correlation was incorporated in the best performing Poisson or NB model based on Deviance Information Criteria (DIC) to create a third model for both links and intersections. The spatial models were then validated on a separate subset of the network with similar geometric and built environment characteristics. Due to the way solitary sites (sites without neighbours) are handled by R-INLA, the crash counts at these sites are predicted using only the fixed effects of the model to prevent prediction of extreme mean crash counts. The models are compared based on goodness-of-fit variables including mean square error (MSE) and Pearson's linear correlation coefficient (CORR). MSE is the averaged squared difference between the fitted (or predicted) crash count and the observed crash count. CORR is the covariance of fitted (or predicted) crashes with observed crashes divided by the product of their standard deviations.

Next, the final spatial models are calibrated based on the data from the entire road network. To evaluate model predictions, these models are validated using a 10-fold cross validation. Sites are randomly split into 10 groups (or folds). The models are then estimated 10 times, with a different fold set aside for validation each time. Each calibrated model is then used to predict the

number and severity of crashes at each site in the validation fold. To demonstrate the predictive power, the sites are ranked based on the predicted crash cost and compared to the crash-based site rankings using Spearman's rank correlation and percent deviation. Finally, maps are generated to observe the locations of the identified hotspots.

4.4 Sample Network Results

4.4.1 Data Exploration

A subset of the road network near Laval University, west of the city center as shown in Figure 4-2, was used for model calibration and validation. This area was selected for its road density, diversity of roadway functional class, and average trip volumes. An area containing approximately 1000 links and intersections was selected, from which approximately 50 % of the sites were selected randomly for calibration, with the remaining sites used for validation. At the link level, the calibration data set contained 71 motorway links, 237 arterial and collector links (grouped from the primary, secondary, and tertiary classifications within OSM), and 220 residential links, for 528 links in total. In terms of intersections, 82 were classified as motorway, 157 as arterial/collector, and 275 as residential, for 514 total intersections, classified by taking the highest classification for links adjacent to the intersection. The validation set contained 58 motorway links and 87 intersections, 288 arterial and collector links and 166 intersections, and 220 residential links and 291 intersections (566 links and 544 intersections total).

FIGURE 4-2 Map of study location

Descriptive statistics for all model variables are provided in Table 4-1. An average link experiences 0.58 crashes over the study period, through the maximum value is 35 crashes. 3.89 crashes occurred at an average intersection, up to 53 crashes at the most extreme site. The number of trips on each facility varies greatly, from a minimum of 2 to a maximum of several thousand. In this study, the number of GPS trips is used as a proxy for exposure (volume). It is assumed that the distribution of GPS trips is proportional to the total demand. Based on a small sample of freeways, arterials, and collectors (45 links) the correlation between average annual daily traffic (AADT) (*147*) and observed GPS trips was 0.75. This result supports the assumption that GPS trip counts can be used as a proxy of traffic exposure. On average, each trip along a link will experience 0.11 HBEs, while for intersections, the number is 0.23. Congestion tends to be higher along links, while CVS is higher at intersections. Average travel speed is approximately 42 km/h. Correlations for the model variables are provided in Table 4-2. Note that both the number of trips and link length were converted using the natural log. For intersections, multiple functional forms relating exposure to crash frequency, such as those proposed by Miaou and Lord (*61*), were tested. The natural log of the sum of all trips passing through the intersection was found to provide the best results.

TABLE 4-1 Variables and Descriptive Statistics for the Calibration Data Set

	Mean	Minimum	Maximum	Std. Dev.
Links (sample size = 528)				
Crashes	0.58	0.00	35.00	3.15
ln(Trips)	4.38	0.69	7.61	1.22
ln(Length)	4.69	1.50	6.95	0.74
HBEs/Trip	0.11	0.00	0.82	0.13
Congestion Index	0.11	0.00	0.78	0.15
CVS	0.33	0.02	1.40	0.19
Average Speed (km/h)	41.26	7.13	107.42	18.83
Intersections (sample size = 514)				
Crashes	3.89	0.00	53.00	7.55
ln(Trips)	5.31	1.79	8.06	1.10
HBEs/Trip	0.23	0.00	4.66	0.38
Congestion Index	0.12	0.00	0.77	0.14
CVS	0.32	0.02	0.97	0.15
Average Speed (km/h)	42.91	8.32	107.42	18.9

TABLE 4-2 Correlations between Model Variables at the Link and Intersection Level

	Trips	HBEs / Trip	Congestion Index	CVS	Average Speed
Links					
ln(Trips)	1.00	0.36	0.29	-0.13	0.56
HBEs / Trip	-	1.00	0.20	0.24	-0.10
Congestion Index	-	-	1.00	-0.19	0.30
CVS	-	-	-	1.00	-0.62
Average Speed	-	-	-	-	1.00
Intersections					
ln(Trips)	1.00	-0.29	0.23	-0.12	0.50
HBEs / Trip	-	1.00	-0.08	0.07	-0.14
Congestion Index	-	-	1.00	-0.20	0.29
CVS	-	-	-	1.00	-0.64
Average Speed	-	-	-	-	1.00

4.4.2 Model Calibration

At both the link and intersections levels, three separate models were estimated, resulting in six models in total. First, the classic Poisson and NB Bayesian models are estimated using INLA and compared. Next, the Besag Spatial model is fitted using the best performing distribution (either the Poisson or NB) and compared by DIC to determine superiority. DIC was selected because it is less sensitive to extreme values. Goodness of fit measures, including MSE and CORR, are potentially sensitive to extreme values estimated by R-INLA for sites without neighbours.

Poisson and Negative Binomial Models: Link level results for the Poisson and NB models are presented in Table 4-3, where variables statistically significant at 95 % confidence are bolded in red. The NB model outperformed the Poisson by DIC, though it had a poorer goodness-of-fit by mean-square-error (MSE). Goodness-of-fit is illustrated further in Figure 4-3a and Figure 4-3c, where fitted values are plotted against the number of observed crashes. Both models perform well, resulting in MSE below 9.0 and CORR above 0.40. Considering the posterior means of the covariates, results generally supported the relationships between SSMs and crash frequency established previously. In both models, the posterior mean for the number of trips is positive. This result mirrors expectations that increasing exposure (increasing traffic volumes) should increase the likelihood of a crash. Therefore, it is logical that sites with higher traffic volumes also have a higher number of crashes. One departure from expectation is the negative posterior mean

associated with HBEs in both models. In Chapter 3, HBEs were found to be positively correlated with crash frequency, while the posterior mean for HBE/Trip in both the Poisson and NB models is negative. Initially, it is unclear why the coefficient is negative in these models.

For the traffic flow SSMs, the posterior means for both CI and CVS are positive (increasing congestion and speed variation are generally associated with increasing crash frequency). Again, it is logical and expected that sites with higher levels of congested and higher variation in travel speeds would also have a higher frequency of crashes. Not only does this result mirror intuition, but it supports results from Chapter 3. For average speed, the mean is also positive (all else being equal, facilities with higher average speeds tend to have more crashes), although this result contradicts results from Chapter 3. Longer links experience more crashes, and all else being equal, motorways experience fewer crashes compared to residential streets (negative posterior mean), while and arterials/collectors experience more (positive posterior mean). In both models, most of the proposed covariates were statistically significant at 95 % confidence.

TABLE 4-3 Link Model Results for Poisson and Negative Binomial Models

	Poisson				NB			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	**-15.87**	**0.72**	**-17.30**	**-14.48**	**-17.12**	**2.21**	**-21.77**	**-13.07**
ln(Trips)	**0.724**	**0.09**	**0.559**	**0.897**	**0.259**	**0.23**	**-0.200**	**0.710**
HBEs/Trip	**-4.287**	**0.72**	**-5.724**	**-2.904**	-3.120	1.84	-6.862	0.379
Congestion Index	0.613	0.50	-0.385	1.577	0.500	1.36	-2.118	3.249
CVS	**4.201**	**0.50**	**3.209**	**5.163**	**6.979**	**2.14**	**2.956**	**11.395**
Average Speed	0.004	0.03	-0.060	0.067	0.122	0.10	-0.079	0.334
ln(Length)	**1.882**	**0.11**	**1.660**	**2.106**	**2.190**	**0.39**	**1.464**	**3.004**
Motorway	**-1.001**	**0.47**	**-1.937**	**-0.098**	**-1.866**	**1.26**	**-4.440**	**0.509**
Arterial/Collector	**1.692**	**0.21**	**1.297**	**2.102**	**1.134**	**0.49**	**0.167**	**2.094**
Number of cases	528				528			
DIC	1183.4				586.0			
MSE	6.7				8.7			
CORR	0.58				0.43			

Note: Variables significant at 95 % confidence are bolded in red

Table 4-4 contains the results for the intersection level Poisson and NB models. Model fit is illustrated in Figure 4-3b and Figure 4-3d. Although the CORR values are comparable to the link-level models, the intersection level models have a higher MSE (approximately 40 compared to less than 10 for link-level models). However, this is expected simply because crash counts at

intersections are much higher on average. As with the link level model, the NB was the superior by DIC though it had a slightly higher MSE. Although this result may seem counterintuitive, DIC is relatively insensitive to extreme values that may be produced by the model, while MSE is quite sensitive. Therefore, it is possible for a model to be superior by DIC, even if it produces one or more extreme values which are captured by the MSE.

TABLE 4-4 Intersection Model Results for Poisson and Negative Binomial Models

	Poisson				NB			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	**-2.701**	**0.23**	**-3.146**	**-2.259**	**-3.627**	**0.64**	**-4.891**	**-2.378**
ln(Trips)	**0.890**	**0.04**	**0.820**	**0.961**	**0.840**	**0.09**	**0.664**	**1.020**
HBEs/Trip	**0.602**	**0.06**	**0.478**	**0.719**	**0.839**	**0.28**	**0.318**	**1.424**
Congestion Index	**0.477**	**0.21**	**0.063**	**0.885**	**1.217**	**0.58**	**0.108**	**2.380**
CVS	0.171	0.25	-0.319	0.656	1.086	0.77	-0.412	2.607
Average Speed	**-0.133**	**0.01**	**-0.153**	**-0.112**	**-0.064**	**0.03**	**-0.120**	**-0.009**
Motorway	**-0.346**	**0.12**	**-0.576**	**-0.121**	**-0.709**	**0.34**	**-1.357**	**-0.034**
Arterial/Collector	**0.616**	**0.05**	**0.511**	**0.722**	**0.656**	**0.17**	**0.323**	**0.991**
Number of cases	514				514			
DIC	3798.2				2126.4			
MSE	38.6				42.5			
CORR	0.57				0.51			

Note: Variables significant at 95 % confidence are bolded in red

The posterior mean for trips was, again, positive and statistically significant. However, at the intersection level, the posterior mean for HBEs was also positive, in accordance with Chapter 3. It is expected that intersections with more braking events would also have more crashes (as HBEs are considered evasive manoeuvres), a result that is observed in the intersection-level modelling results. For CI and CVS, results were also similar, with both positive posterior means. Unlike the link level model, the posterior mean of average speed is negative, which matches expectations from Chapter 3. This result may seem counterintuitive, with Imprialou et al. (*148*) attributing this negative relationship to link-aggregation bias. However, in this context, it is suggested that that FFS captures other variables not considered, including geometry. In these models, the posterior mean for the motorway variable is negative (motorway intersections, including on- and off-ramps, are less likely than residential intersections to experience crashes), while for arterials and collectors, it is positive. Again, most of the proposed covariates were statistically significant at 95 % confidence.

Spatial Models: The NB model was the best performing model by DIC for both links and intersections. Therefore, the spatial component was incorporated into the NB model formulation to yield spatial NB models, which are summarized in Table 4-5.

TABLE 4-5 Negative Binomial Spatial Model Results for Links and Intersections

	Links				Intersections			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	**-17.66**	**2.14**	**-22.10**	**-13.72**	**-3.627**	**0.64**	**-4.890**	**-2.378**
ln(Trips)	0.130	0.23	-0.325	0.581	**0.840**	**0.09**	**0.664**	**1.020**
HBEs/Trip	-3.136	1.70	-6.623	0.075	**0.839**	**0.28**	**0.318**	**1.424**
Congestion Index	1.648	1.47	-1.243	4.527	**1.217**	**0.58**	**0.108**	**2.380**
CVS	**4.592**	**1.86**	**1.112**	**8.470**	1.086	0.77	-0.412	2.606
Average Speed	0.066	0.10	-0.122	0.264	**-0.064**	**0.03**	**-0.120**	**-0.009**
ln(Length)	**2.506**	**0.38**	**1.783**	**3.296**	*N/A*	*N/A*	*N/A*	*N/A*
Motorway	**-2.095**	**1.17**	**-4.496**	**0.110**	**-0.709**	**0.34**	**-1.357**	**-0.034**
Arterial/Collector	**1.234**	**0.49**	**0.276**	**2.219**	**0.656**	**0.17**	**0.324**	**0.991**
Number of cases	528				514			
DIC	556.4				2126.3			
MSE	3.4				42.5			
CORR	0.83				0.51			

Note: Variables significant at 95 % confidence are bolded in red

By DIC, the spatial NB models were the best performing of all considered model types, and improved the goodness-of-fit compared to the non-spatial NB models, as shown in Figure 4-3e and Figure 4-3f. The spatial model greatly improved the goodness-of-fit at the link level, while having virtually no effect for intersections. The reason for this may be that adjacent links truly are adjacent (are physically connected at intersections) and have much stronger spatial correlations than intersections, where neighbouring sites may be separated by several hundred metres or more. Not only were these models statistically superior while improving goodness-of-fit on the calibration data set, supporting previous work (*57*) indicating the importance of including spatial correlations in crash frequency models, but the signs of the posterior means for covariates generally matched expectations from Chapter 3. Means for trips were consistently positive. Although HBEs/Trip had a negative posterior mean in the link-level model, the mean was positive at the intersection level. CI and CVS were similarly positive, while average speed had a positive mean for links and a negative mean for intersections. However, these models also contain more non-significant variables than the non-spatial models (at 95 % confidence).

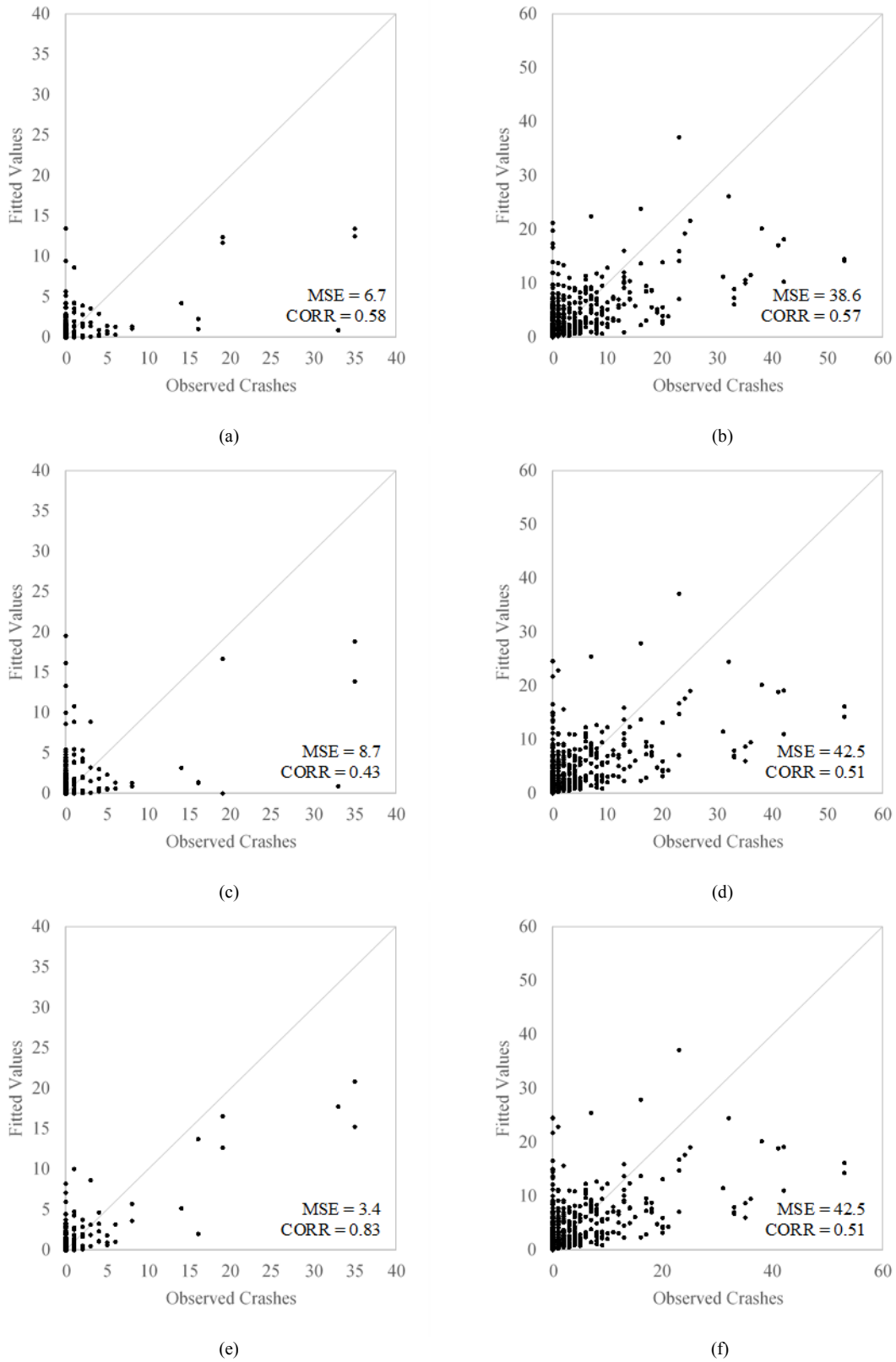
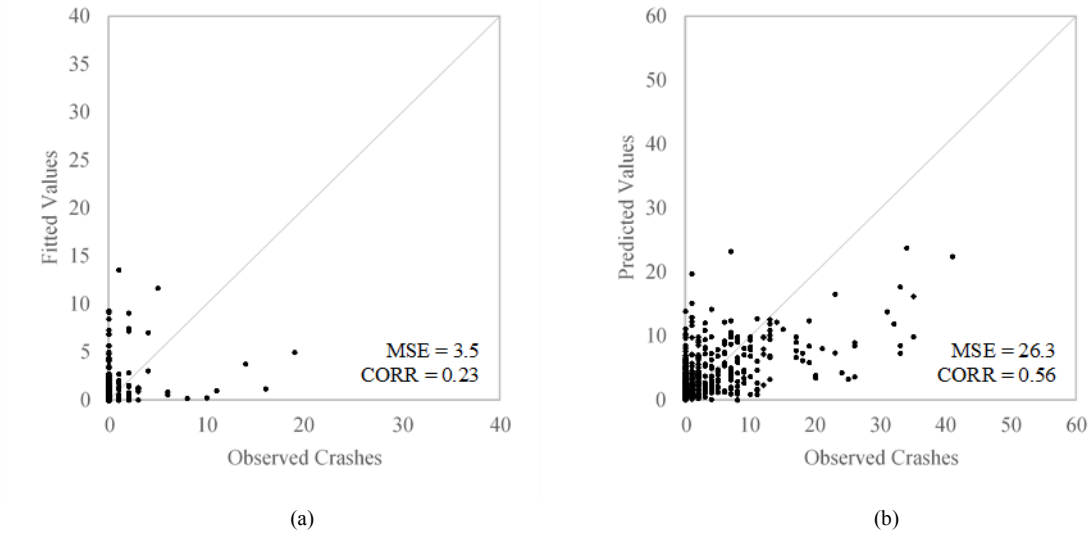


FIGURE 4-3 Fitted values versus observed crashes for Poisson models for links (a) and intersections (b), NB models for links (c) and intersections (d) and NB Spatial models for links (e) and intersections (f)

For example, trips, HBEs, congestion, and speed were non-significant in the link-level model, while HBEs, congestion, speed variation, and average speed are non-significant in the intersection-level model. This is logical, as the spatial component provides information that may have been correlated with these independent variables It is also important to note that these models are estimated on a relatively small sample of the whole road network

4.4.3 Model Validation

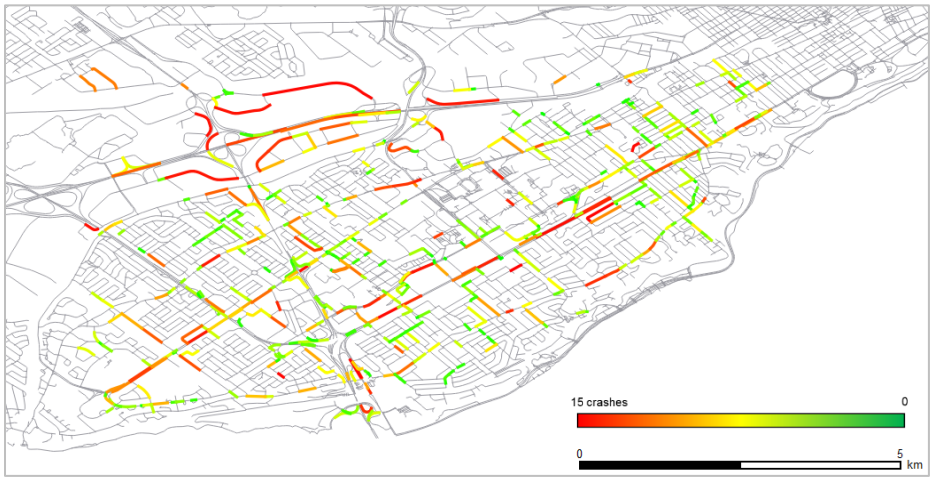
The NB Spatial models developed above were validated on a separate data set randomly selected from the data sample near Laval University. The results of this validation procedure are summarized in Figure 4-4a for the link-level model and Figure 4-4b at the intersection level. As can been seen in Figure 4-4b, the model makes accurate predictions for the intersection-level data set, and a CORR of 0.56 is quite promising. It is believed that the sample size used in this study was not sufficient to yield accurate link-level predictions. Still, prediction at the link level produces an MSE below 4.0. At the time of publication, R-INLA has no specific module for model validation. Prediction must be accomplished by setting null values for the validation sites during model calibration. Although both models can describe the relationships between SSMs and crash frequency, the intersection-level model shows more promise of predicting crash occurrence.



(a) (b)

FIGURE 4-4 Predicted values versus observed crashes for links (a) and intersections (b)

The posterior expected number of crashes as predicted by the model are illustrated in Figure 4-5. These maps clearly show the locations of dangerous sites (sites with a high expected number of crashes) as predicted by the model, shown in red. For links, the most extreme sites are expected to experience 15 crashes in an 11-year period, while for intersections, 25 crashes is the expected maximum. Furthermore, the most dangerous sites tend to be along the major arterials within the study area or on freeway ramps and interchanges, as would be expected for a major city.



(a)

(b)

FIGURE 4-5 Sites ranked by posterior expected number of crashes for links (a) and intersections (b)

4.5 Full Network Results

4.5.1 Data Exploration

There were sufficient collected data (at least two trips with two observations per trip) to calculate the proposed SSMs for 4623 links, of which 494 were classified as motorways, 2166 were classified as arterials and collectors (grouped from the primary, secondary, and tertiary classifications within OSM), and 1963 were classified as residential. In terms of intersections, 4429 had sufficient data for modelling, of which 605 were motorways (mainly at the locations of access ramps), 1223 were arterials/collectors, and 2601 were residential. Descriptive statistics for model variables are provided in Table 4-6. On average, all sites experience slightly more than one crash every two years, with the vast majority resulting in minor injuries. The most crash prone links experience up to 43 crashes over the observation period, while up to 57 crashes occurred at the most extreme intersections. For an average link, about one in 12 trips experience an HBE, while for intersections, the number is one in five. Average CI was 0.08 and 0.09 for links and intersections respectively, average CVS was 0.31, and average speed was approximately 42 km/h.

TABLE 4-6 Variables and Descriptive Statistics

	Mean	Minimum	Maximum	Std. Dev.
Links				
Crashes				
Minor Injury	0.31	0.00	40.00	1.76
Major Injury	0.02	0.00	2.00	0.15
Fatal	$2 \cdot 10^{-3}$	0.00	1.00	0.05
Trips	128.71	2.00	2140.00	218.60
HBEs/Trip	0.08	0.00	2.58	0.14
Congestion Index	0.08	0.00	0.78	0.12
CVS	0.31	0.01	1.67	0.19
Average Speed (m/s)	11.32	1.05	30.46	5.26
Intersections				
Crashes				
Minor Injury	2.70	0.00	51.00	5.60
Major Injury	0.15	0.00	4.00	0.48
Fatal	0.02	0.00	2.00	0.15
Trips	271.34	4.00	4282.00	417.87
HBEs/Trip	0.22	0.00	7.50	0.39
Congestion Index	0.09	0.00	0.78	0.11
CVS	0.31	0.01	1.40	0.16
Average Speed (m/s)	11.72	1.38	30.40	5.34

4.5.2 Modelling Crash Frequency

The results of the Spatial NB model of crash frequency are presented in Table 4-7, and two promising observations are evident. First, most of the proposed SSMs are statistically significant at 95 % confidence in both the link- and intersection-level models, highlighted in red. Second, the direction of the effect of all variables (whether the posterior mean is positive or negative) is generally consistent with expectation and results from previous work. The posterior mean for the natural log of GPS trips, a proxy for exposure, is positive in both models. As exposure increases, crash frequency also increases. Previously, HBEs were shown to be positively correlated with crash frequency, and the posterior mean of HBEs/Trips is positive and consistent in the intersection model, providing more evidence for this positive correlation. The signs for CVS and CI are also consistent with previous results. Increased congestion increases crash frequency, as does CVS. Sites with a higher average speed tend to have fewer crashes overall. Increasing link length is also associated with crash frequency. All else being equal, motorways have fewer crashes than residential streets, while arterials and collectors tend to have a greater number.

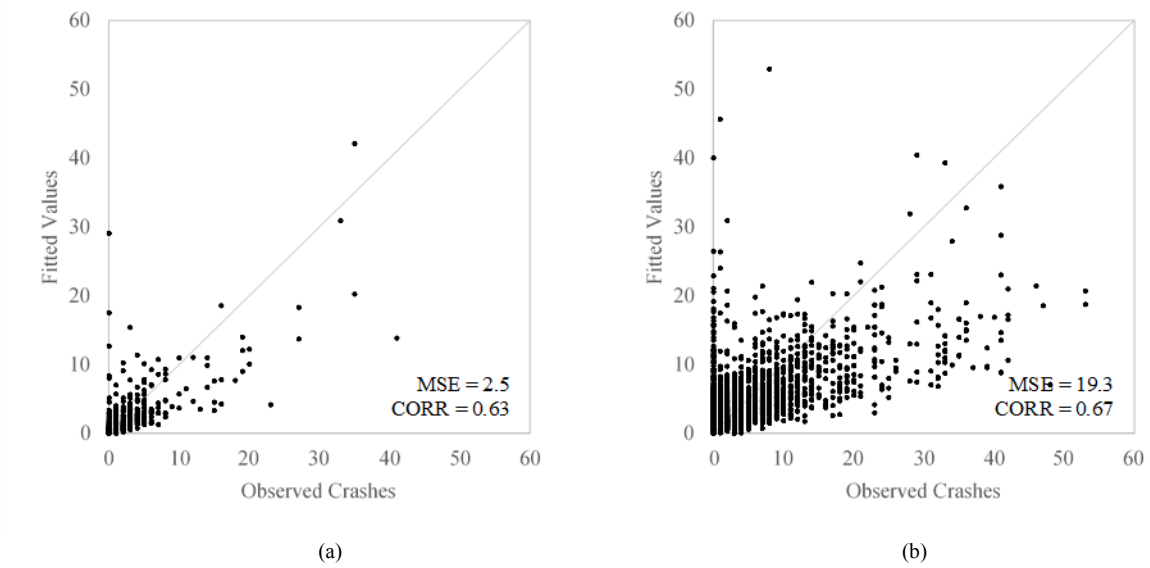
TABLE 4-7 Results of the Negative Binomial Spatial Model for Links and Intersections

	Links				**Intersections**			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	**-15.83**	**0.87**	**-17.60**	**-14.22**	**-2.423**	**0.20**	**-2.811**	**-2.039**
ln(Trips)	**0.495**	**0.09**	**0.320**	**0.672**	**3.826**	**0.17**	**3.502**	**4.153**
HBEs/Trip	-0.438	0.65	-1.735	0.806	**1.131**	**0.10**	**0.940**	**1.328**
Congestion Index	**1.436**	**0.56**	**0.339**	**2.521**	**0.892**	**0.28**	**0.346**	**1.440**
CVS	0.972	0.51	-0.044	1.978	**0.905**	**0.26**	**0.402**	**1.408**
Average Speed	-0.046	0.03	-0.106	0.014	**-0.064**	**0.01**	**-0.085**	**-0.043**
ln(Length)	**2.110**	**0.12**	**1.881**	**2.354**	*N/A*	*N/A*	*N/A*	*N/A*
Motorway	**-3.037**	**0.47**	**-3.997**	**-2.135**	**-0.670**	**0.13**	**-0.921**	**-0.420**
Arterial/Collector	**0.848**	**0.19**	**0.485**	**1.220**	**0.449**	**0.06**	**0.323**	**0.576**
Number of cases	4623				4429			
DIC	3756.5				15920.7			
MSE	2.5				19.3			
CORR	0.63				0.67			

Note: Variables significant at 95 % confidence are bolded in red

Model fit is further demonstrated in Figure 4-6. Model performance, as measured by MSE and CORR, are comparable between the models. Although the MSE is higher in the intersection level model, this is simply because the crash counts at intersections are higher overall (mean crash

counts are 2.87 at intersections and only 0.33 on links). Correlations between observed crashes and fitted values are relatively high, which is a promising result. Based on the calibration data set, the selected covariates can explain 63 % and 67 % of the variation in the crash counts for links and intersections respectively. The model tends to underestimate the number of expected crashes, as most of the data points fall below the diagonal line which represents ideal performance.



(a) (b)

FIGURE 4-6 Fitted values versus observed crashes for links (a) and intersections (b)

4.5.3 Modelling Crash Severity

Results of the FMNL model of crash severity are presented in Table 4-8. The trip and length variables are omitted as the volume/exposure and length are not typically associated with crash severity. Additionally, the arterial/collector variable had to be omitted from the link-level model to achieve convergence of the log-likelihood. Minor injury crashes were selected as the base case, so coefficients represent the change in utility for major injury and fatal crashes. Compared to the frequency model, fewer variables are significant at 95 % confidence. However, considering those variables that are statistically significant, results are generally consistent with expectations based on results from Chapter 3. For example, intersections with a higher number of HBEs/Trip are more likely to experience fatal collisions, confirming a positive correlation between braking and crash severity shown previously. In contrast, a higher CI is associated with a significant reduction in the chance of a fatal crash at links. Although congestion is likely to increase the frequency of crashes

due to increased exposure, it is also likely to reduce crash severity as a result of speed reduction. Both average speed and variation in speed are positively linked to fatal crashes.

TABLE 4-8 Results of the Fractional MNL for Links and Intersections

	Links				Intersections			
Explanatory variables	**coeff**	**std err**	**95% CI**		**coeff**	**std err**	**95% CI**	
Major Injury								
Intercept	**-4.022**	**0.98**	**-5.939**	**-2.105**	**-2.835**	**0.42**	**-3.656**	**-2.014**
HBEs/Trip	-0.504	2.23	-4.872	3.863	-0.234	0.18	-0.592	0.125
Congestion Index	-0.500	1.76	-3.941	2.942	-0.450	0.55	-1.531	0.631
CVS	0.124	1.17	-2.164	2.412	-0.322	0.59	-1.475	0.832
Average Speed	**0.114**	**0.05**	**0.007**	**0.221**	0.031	0.02	-0.017	0.079
Motorway	-1.543	0.84	-3.198	0.111	**-0.890**	**0.36**	**-1.603**	**-0.178**
Arterial/Collector	*N/A*	*N/A*	*N/A*	*N/A*	-0.054	0.14	-0.320	0.211
Fatal								
Intercept	**-11.17**	**2.81**	**-16.67**	**-5.662**	**-7.483**	**1.07**	**-9.573**	**-5.393**
HBEs/Trip	-1.91	5.66	-13.00	9.18	**0.512**	**0.14**	**0.245**	**0.779**
Congestion Index	**-9.393**	**4.15**	**-17.53**	**-1.253**	-2.164	1.65	-5.408	1.080
CVS	**5.729**	**2.38**	**1.072**	**10.385**	**1.938**	**0.88**	**0.213**	**3.663**
Average Speed	**0.366**	**0.16**	**0.047**	**0.685**	**0.197**	**0.06**	**0.079**	**0.314**
Motorway	-1.723	2.43	-6.487	3.040	**-1.555**	**0.76**	**-3.051**	**-0.058**
Arterial/Collector	*N/A*	*N/A*	*N/A*	*N/A*	**-0.862**	**0.34**	**-1.522**	**-0.201**
Number of cases	453				2204			
Log likelihood	-117.7				-574.4			

Note: Variables significant at 95 % confidence are bolded in red

4.5.4 Site Ranking

Next, the results of both models are combined to rank sites based on the decision parameter, δ_i, calculated using both crash data and the modelling results. Results are compared using Spearman's rho and percent deviation. Figure 4-7 compares the percent deviation for the crash-based and modelling approaches for hotspot lists of 100-1000, increasing in 100 site increments. Note, a lower percent deviation indicates superior performance. Percent deviation for links is around 50 % regardless of list size, while for intersections, percent deviation decreases with increasing list size, from about 70 % down to 40 %. This means, based on the calibration data set, there is about 50 % agreement between site rankings determined using the model and the crash data when considering the top 500 dangerous sites. When considering the results for all links, Spearman's rho between the ranks based on crashes and the model is 0.47, while for intersections Spearman's rho was 0.63.

The slightly better performance at the intersection level is again attributed to larger sample size and overall higher observed crash counts which aid the calibration of the models.

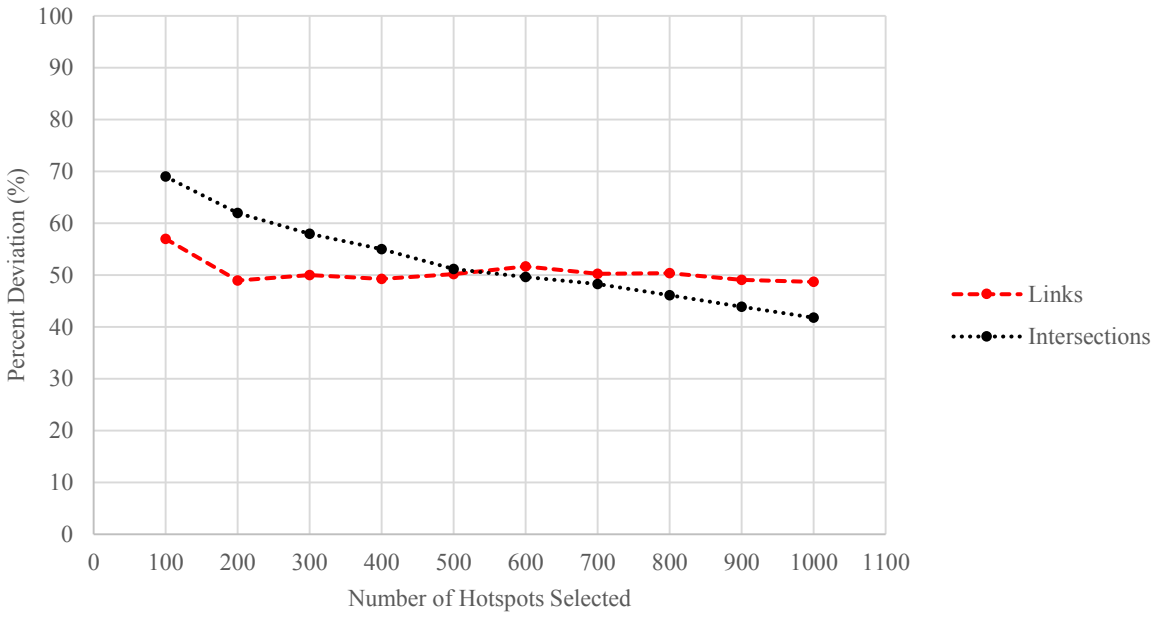


FIGURE 4-7 Percent deviation for hotspots generated by crash data and modelled using the calibration data

The disagreement between the model and observed data can be ascribed to two sources. First, sites may be ranked higher by the crash data than by the model. This is generally not an issue for network screening if the site rankings are quite high (they would not be selected for remediation by either crash data or the developed model). However, there are several sites in both models with a high history of crashes which are not ranked highly by the surrogate model. At these sites, the selected covariates are simply unable to capture the source of historical crashes. Additional variables or increased model flexibility may be able to capture the cause of crashes at these few sites. Conversely, there are many sites which are expected to have a high crash cost based on the modelling results, which have zero (or very few) observed crashes (they are ranked much higher by the surrogate safety model than by the crash data). However, this does not necessarily mean that the model is performing poorly. Crashes are not perfect predictors of safety. These sites have the features of high-risk sites as determined by the model, even though they have no observed crashes. Therefore, these sites would be of interest because of their potential for future crashes.

4.5.5 Model Validation

The final objective of this paper is to evaluate the model's prediction power using a 10-fold cross validation. The crash counts from the 10 folds predicted using the NB spatial model are compared to observed crashes in Figure 4-8. For the link-level model, CORR decreased from 0.63 to 0.25 (a reduction of 0.39) while CORR for the intersection-level model decreased from 0.67 to 0.46 (a reduction of 0.21). Based on Figure 4-8, the prediction power is much stronger at the intersection level. Although the correlation at the link-level is still positive, the model struggles to predict crash counts. As before, the better performance for intersections is attributed to sample size. For sites without neighbours, crash counts are predicted using only the fixed effects of the model. This effectively reduces the MSE by eliminating extreme values otherwise predicted by R-INLA for solitary sites. Next, the predicted crash counts and severity proportions are again combined to calculate the decision parameter and rank the sites.

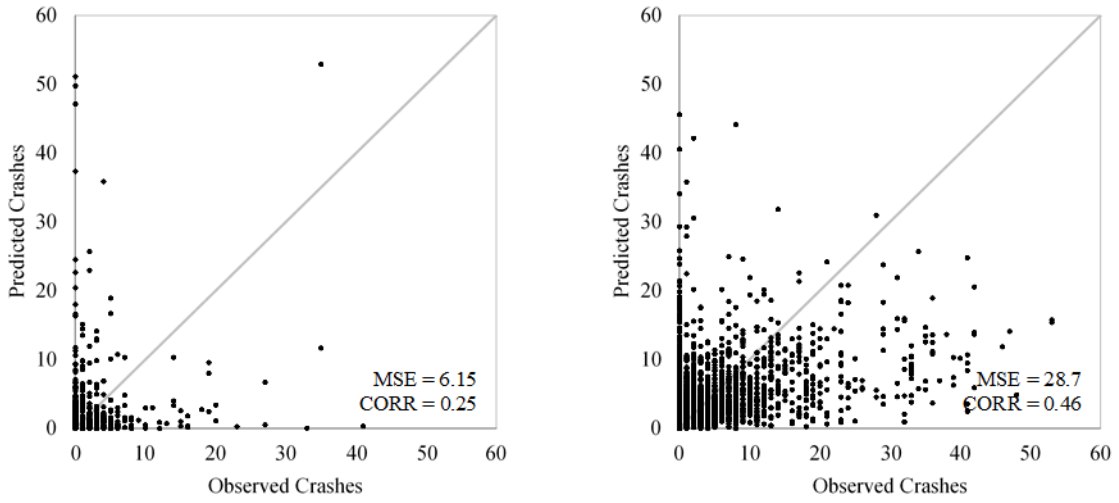


FIGURE 4-8 Predicted values versus observed crashes for links (a) and intersections (b)

Figure 4-9 shows the percent deviation between the list of hotspots generated by the crash data and safety model for various list sizes. When considering a small number of hotspots, the two methods deviate significantly from each other (approximately 85 %). However, as the number of considered hotspots inceases, the deviation between the lists decreases (deviation is 55 % for intersections and 65 % for links when considering 1000 sites). Again, it should be noted that while some of this deviation is due to deficiencies in the model (some truly dangerous sites are not

captured by the model), some of this deviation is attributed to sites with potential for future crashes (sites that have not yet experienced a high number of crashes, but which are likely to in the future). A decrease in the goodness-of-fit is expected and observed for the predicted data. Spearman's rho for the link-level rankings decreased from 0.47 to 0.32 (a reduction of 0.15), while rho for the intersection-level rankings decreased from 0.63 to 0.45 (a reduction of 0.18).

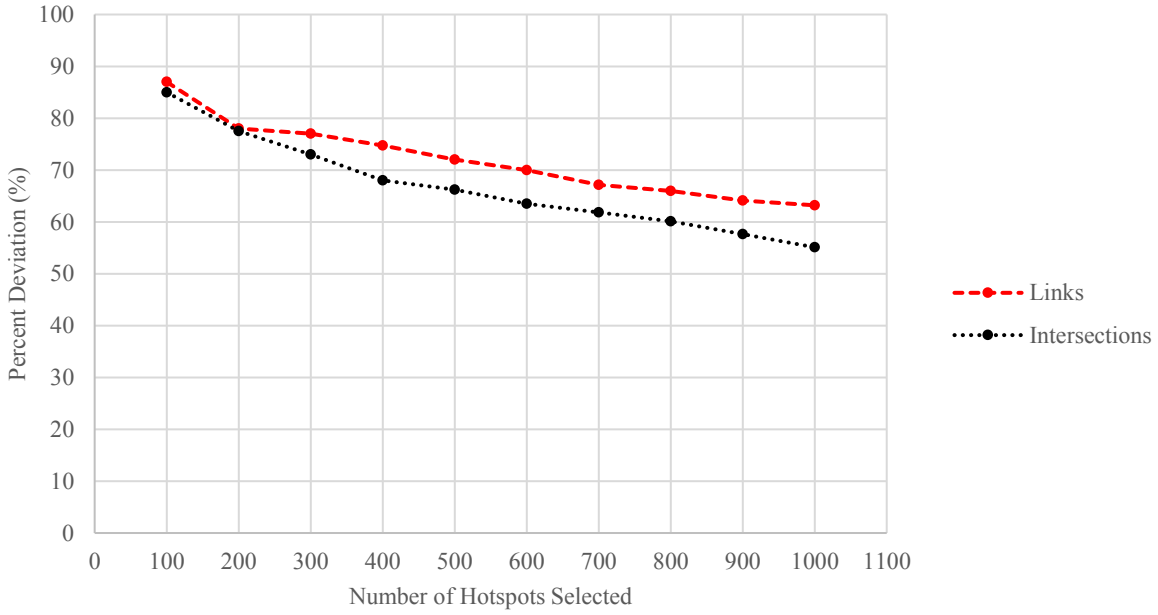


FIGURE 4-9 Percent deviation for hotspots generated by crash data and modelled using the validation data

Finally, the spatial distribution of the predicted rankings, along with the difference between the model and crash based rankings, is illustrated in Figure 4-10. Considering the better performing intersection-level model, there is a concentration of sites with a high predicted ranking near the downtown core of the city. This mirrors expectations of more dangerous sites being located in denser areas with higher traffic volumes. Second, the freeways leading to and from the city centre are generally receive a low ranking. This demonstrates the effect of controlling for exposure using the number of observed trips. Although freeways are expected to have a high number of crashes, because they also experience the highest volumes of all considered sites, the crash cost per vehicle-km (or the crash rate) is expected to be lower than other sites. However, considering the link-level results, the dangerous sites are dispersed throughout the network, demonstrating the poorer

prediction power of the link-level model. Observing the differences between model and crash-based rankings, there is no observable spatial trend. Sites where the model overpredicts (or conversely, underpredicts) the rrelatively site ranking does not appear to be location dependent. Once such a model is calibrated for a given network, the results (the posterior number of crashes) could be used to estimate crashes based on surrogate indicators (when crash data is not available) or to supplement historical crash data in the network screening process (when crash data is available). These models can continuously monitor safety by updating the expected crash cost at each site as more probe data becomes available.

4.6 Conclusions

The purpose of this chapter was to develop mixed-multivariate model capable of estimating the posterior of the expected crash frequency and severity at the link and intersection level across a large urban road network based on smartphone GPS data. This thesis uses two models, combining a Bayesian Spatial LGM to model crash frequency and an FMNL model to estimate crash severity. From this, crash frequency, severity, and cost measures can be derived for use in network screening. Based on a sample network, NB models outperformed Poisson models at both the link and intersection level. In general, the relationships between SSMs and crash frequency established earlier were supported by the modelling results. The results indicate that crashes are more likely at links and intersections with more trips (as a proxy of exposure). Positive means for CI, CVS, and link length also indicate a positive relationship with crash frequency. The only disagreement between link- and intersection-level models was for HBEs and average speed. At the link-level, HBEs had a negative mean (positive at the intersection-level) and speed had a positive mean (negative at the intersection-level). The greatest improvement in model fit was achieved by adding a spatial component to the models. By DIC, the NB Spatial models were the best performing of all considered model types and improved the goodness-of-fit. This result supports previous work indicating the importance of including spatial correlations in crash frequency models. Some of the SSMs were observed to be non-significant in the spatial models. Despite this, the sign of the posterior means of the significant variables generally supported the results of previous research (the relationships between SSMs and crash frequency were confirmed). When considering the predictive power of the model based on a separate validation data set, the intersection-level model performed well, providing a relatively high correlation between observed and predicted crashes.

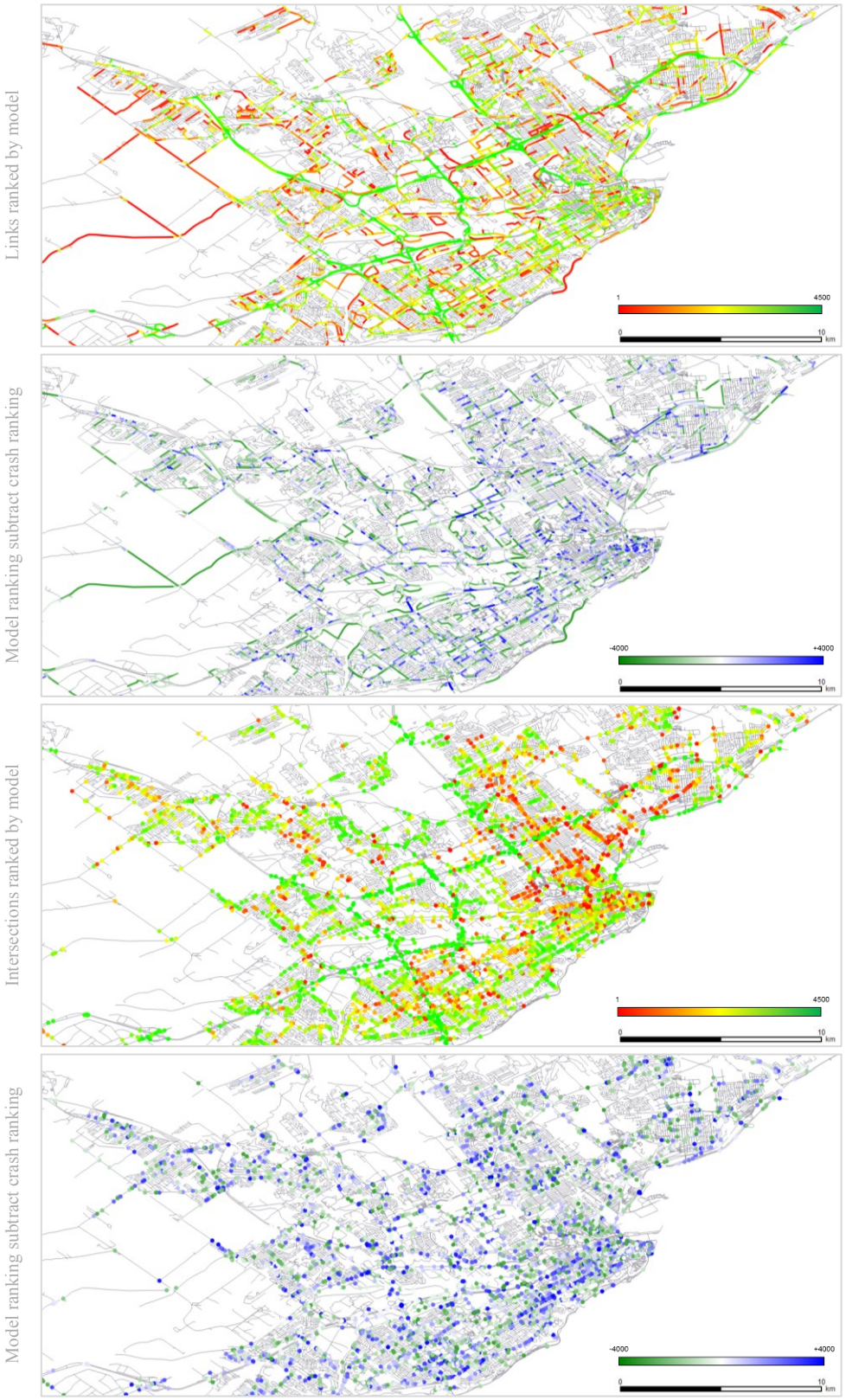


FIGURE 4-10 Sites ranked by model from highest (red) to lowest (green) and difference between model and crash rankings, ranked higher by the model (blue) and higher by crashes (green)

The Bayesian Spatial NB model was shown to model crash frequency well at the network scale. Critically, most of the proposed SSMs were observed to be statistically significant at 95 % confidence in both the link- and intersection-level models. Not only were these variables significant, but the direction of their effect was consistent with previous results. Namely, HBEs, congestion, and speed variation were all positively correlated with crash frequency, while average speed was negatively correlated. This result supports the fundamental relationship between crash frequency and the proposed SSMs. Model fit, as measured by the MSE and correlation coefficient, was comparable for both link- and intersection level models, with correlation exceeding 0.60 using the calibration data set. In the second stage, crash severity was accounted for using a discrete choice framework. The probability for crashes at three distinct severity levels was estimated using an FMNL model. In this model, fewer variables were significant compared to the frequency model. Yet, the direction of the effect of all significant variables was again consistent with previous research or known causal relationships. Namely, congestion tends to reduce the likelihood of fatal crashes, while increases in HBEs, speed, and speed variation are linked with an increase in crash severity. Each link and intersection was ranked according to the cash cost per vehicle-km, calculated first using the historical crash data and second using the modelling results. The ranked lists generated by the mixed multivariate model and the ranked lists based on crash data had a correlation of 0.47 for links and 0.63 for intersections. Percent deviation was 55 % when considering 500 sites. The model was finally validated using a 10-fold cross validation approach, and site rankings were again generated based on the predicted crash counts and severity proportions. For the validation data set, Spearman's rho was observed to decrease by 0.15 for links and 0.18 for intersections. This result shows that the predicted site rankings differ from the fitted site rankings by less than 20 %. Although the intersection level model provided promising results, identifying many hotspots near the city centre, the link level model had poorer prediction power. Still, the selected covariates can explain 32 % of the variation in crash cost for links and 45 % of the variation for intersections. Again, not all the discrepancy is considered a negative, as this approach is able to identify sites with a high potential for crashes (based on the selected SSMs) even if those sites have not historically experienced crashes, or if crashes have been unreported or incorrectly geo-referenced.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

"It's more fun to arrive a conclusion than to justify it."

Malcom Forbes, Entrepreneur

5.1 Summary of Results and Contributions

The purpose of this thesis was to develop statistical models capable of predicting crash frequency and severity at the network scale based on the GPS travel data of regular drivers and minimal, publicly available geometric information. This purpose was achieved through three specific objectives. First, methods for automated data processing and integration were proposed and implemented. The raw GPS smartphone data was processed using map matching and speed filtering algorithms to reduce the noise present in the signal. Techniques for managing associated mapping and crash data were revised and implemented. Finally, a data visualization exercise was undertaken to observe network-wide congestion patterns. Second, several SSMs were proposed, extracted, and validated based on a large amount of historical crash data at links and intersections in an urban environment. Importantly, SSMs were evaluated with respect to both crash frequency and severity, which has rarely been the case in existing surrogate safety studies. Third, statistical models were developed to estimate and predict crash risk for dangerous locations. These models utilized the extracted SSMs as covariates for screening the network using GPS travel data. As has been recommended in existing work, spatial autocorrelations were included to account for unobserved homogeneity between neighbouring sites. The results and contributions related to each of these objectives is summarized in the following sections.

5.1.1 Data Collection, Processing, and Visualization

In Chapter 2, the details of data collection and implementation of a methodology for data processing and visualization are provided. The GPS data used in this study were collected during a three-week period in Quebec City, Canada. In total, nearly 22,000 trips were collected from 4000 drivers. Processing was required to remove noise from the raw GPS data required, both in terms of position and reported speeds. The TrackMatching algorithm proved to be highly successful in accurately linking the GPS observations to the OSM road network. Furthermore, various parameters for the Savitzky-Golay filter, recommended by earlier studies for filtering and differentiating kinematic vehicle data, were tested and their effects quantified. In general, the filter successfully removed noise in the reported GPS speeds while sufficiently maintaining the original signal for analysis. Furthermore, the filter can automatically provide the acceleration rate (by taking the derivative of the polynomials successively fitted to the observed speeds), which is further useful for studying SSMs based on acceleration or deceleration, as in this thesis.

Once processed, measures of interest were extracted at the link level using developed Python scripts, and an application of measuring and visualizing network-wide congestion levels was presented. This work defines a network link as a road segment which connects adjacent intersections. The OSM road network was modified using GIS software to ensure this definition was consistently met. Next, CI was computed for every link during each hour of the AM and PM peak periods (defined as 6:00 AM to 10:00 AM and 3:00 PM to 7:00 PM respectively). CI proved easy to compute, requiring only several in-peak and off-peak observations (with the off-peak observations used to defined FFS). The analysis and visualization of the congestion data provided results that were consistent with expectations for both microscopic and macroscopic travel patterns. The general rise and fall of congestion over an average weekday was observed, and peak periods were observed to have both on onset and dissipation period (in which CI was consistently lower than during the middle hours) lasting approximately one hour each. The presented analysis also identified the relative contributions of different functional classes on overall congestion (with motorways contributing the most) and identified chronically congested links by quantifying the number of hours each link spent in a highly congested state.

The primary contribution presented in the chapter is the integration of tools and methods into a framework to automatically collect, process, and analyze smartphone GPS data. In general, although opportunities for collecting naturalistic driving data with instrumented vehicles have grown thanks to the emergence and development of in-vehicle sensors and GPS smartphones, few studies have attempted to collect network-wide data using the smartphone GPS of regular drivers alone. This research combines both pre-existing and newly-developed algorithms to enable network-wide analysis of GPS data, which has been rare in existing work. Both TrackMatching and the Savitzky-Golay filter are practical and accessible, making them transferable and easy to integrate into other systems for similar purposes. Additionally, different methods for data visualization are presented. Although this chapter presents one application for congestion monitoring and analysis, the tools developed for extracting traffic information and visualizing the data could be easily applied to other measures of interest. The methodology presented in this chapter make it clear that large-scale analysis of smartphone data is not only possible but extremely practical, and the techniques adopted and developed should assist others in analyzing network-wide traffic patterns using probe vehicle data.

5.1.2 Extracting Surrogate Safety Measures from GPS Data

Chapter 3 covers the extraction and validation of various SSMs with respect to a large quantity of historical crash data. SSMs were first extracted from the processed GPS data using developed Python scripts. Then, Spearman's rank correlation coefficient was used to determine relationships with crash frequency, and pairwise K-S tests were used to determine relationships with crash severity. This analysis was completed separately for the five major functional classes (motorway, primary, secondary, tertiary, and residential). First, event-based measures related to vehicle manoeuvres were considered. Results for both HBEs and HAEs were substantially similar, with correlations of 0.53 to 0.64 between braking/accelerating and crash frequency at the intersection level depending on functional classification. Correlations at the link level were still positive, though their strength was significantly lower, indicating that HBEs and HAEs are better predictors of crashes at the intersection level. In terms of crash severity, the distributions of the number of vehicle manoeuvres were shifted to significantly higher values for links and intersections that had experienced at least one fatal crash (compared to sites that had experienced, at worst, major injuries or minor injuries). It can be concluded that, in addition to a positive relationship with crash frequency, HBEs and HAEs also have a positive relationship with crash severity. In other words, sites that experience a high number of vehicle manoeuvres are likely to experience more crashes that are more severe.

Next, the previously extracted measure of congestion was similarly validated, along with additional traffic flow measures of FFS and speed variation. CI was shown to be positively correlated with crash frequency, although the strength of this relationship was weaker compared to vehicle manoeuvres, with a maximum Spearman's rho of 0.21. This result supports findings in existing literature, that increasing congestion increases the likelihood of crashes. The relationship between CI and crash severity was non-monotonous. Links experiencing, at worst, a major injury crash were more likely to have higher levels of congestion. As congestion increases, the likelihood of a site experiencing a fatal crash decreased, as did the likelihood of experiencing only minor injury crashes. As discussed earlier, this complex relationship can be explained considering the relationship between flow, density, and speed. FFS was found to be negatively correlated with crash frequency, demonstrating a maximum Spearman's rho of -0.45, and generally had no conclusive statistical relationship with crash severity. The negative relationship between FFS and crash occurrence may seem counterintuitive. However, this analysis does not control for other

factors except for roadway functional class. The considered speed variable likely captures other factors, such as road geometry, that effectively reduce the likelihood of crashes. Additionally, the study considers the effect of speed at the network scale, rather than for a single link or corridor. Speed uniformity, measured with CVS, was positively correlated with both crash frequency and severity. Spearman's rho varied between 0.13 and 0.38, and relationships with crash severity were found to be significant in half of the test cases, with CVS distributions on links experiencing major and fatal crashes shifted to higher values.

This chapter contains three main contributions. First, the use of instrumented vehicles to extract traffic flow SSMs is likely the first of its kind. Although event-based measures related to acceleration and braking have been studied using both conventionally-instrumented vehicles and smartphones, traffic flow SSMs have not been studied using probe vehicles. This work demonstrates how both event-based and traffic flow measures can be extracted from smartphone GPS data and how their relationships to historical crash frequency and severity can be quantified. In fact, the methods used to compare SSMs to the observed crash data are robust when compared to existing literature. Related to this, the second contribution presented in this chapter is the comparison of SSMs to large volumes of historical crash data. The majority of surrogate safety studies have focussed on video-based conflict analysis for site diagnosis. This means that, even when compared to historical data, the comparison can only be carried out at sites where the video cameras have been installed. Additional studies considering SSMs of hard braking have compared results to self-reported safety data or to near misses. By definition, SSMs must be physically and predictably related to crashes. Therefore, more effort was needed to compare any proposed SSM with a reasonable amount of historical crash data to demonstrate that such a relationship exists. This study utilizes 11 years of crash data across the entire Quebec City road network, resulting in a substantial amount of data for comparison. The final significant contribution of this chapter is the validation of SSMs with respect to both collision frequency and severity. Frequency and severity are independent dimensions of road safety that have been studied extensively using crash-based techniques. Yet few surrogate safety studies have explicitly considered injury severity. Spearman's correlation coefficient, a widely used and proven technique, was adopted to determine relationships with crash frequency, but pairwise K-S tests were further proposed to determine relationships with crash severity. The inclusion of the independent dimension of crash severity in the validation of SSMs is original, having been ignored in existing surrogate safety studies.

5.1.3 Modelling Crash Frequency and Severity

Crash modelling results are covered in Chapter 4. First, several models for crash frequency were estimated and compared based on a subset of data. All crash frequency models were estimated using the INLA technique for Bayesian inference. At both the link and intersection levels, NB models outperformed Poisson models by DIC (DIC was reduced by 45-50 % in the NB models). Next, spatial autocorrelations were introduced using the Besag proper model. By DIC, the NB Spatial models were the best performing of all considered model types and improved the goodness-of-fit (although improvements were very minor for the intersection-level model). This result supports previous work indicating the importance of including spatial correlations in crash frequency models. Second, the relationships between SSMs and crash frequency and severity were corroborated. In general, the relationships established in previous studies and earlier in this thesis were supported by the modelling results. Critically, most of the proposed SSMs were observed to be statistically significant at 95 % confidence in both the link- and intersection-level models. Not only were these variables significant, but the direction of their effect was consistent with previous research. Namely, HBEs, congestion, and speed variation were all positively correlated with crash frequency, while average speed was negatively correlated. The probability for crashes at three distinct severity levels was estimated using an FMNL model. In this model, fewer variables were significant compared to the frequency model. Yet, the direction of the association of all significant variables was again consistent with previous research or intuition. Namely, congestion is associated with a reduction of the likelihood of fatal crashes, while increases in HBEs, speed, and speed variation are linked with an increase in crash severity.

Third, the fit and prediction power of the models were quantified. The fit of the NB Spatial models, as measured by MSE and CORR, was comparable for both link- and intersection level models, with correlation exceeding 0.60 using the calibration data set. When sites were ranked using the decision parameter of crash cost per vehicle km, the list of hotspots determined using the models and determined using the crash data were approximately 50 % similar. When it comes to prediction, the intersection-level model performed well, achieving a correlation of 0.46 between observed and predicted crashes, and hotspot lists were between 20 % and 45 % similar measured on the validation data set, depending on the number of hotspots considered. Link level prediction was worse, with a correlation of 0.25 between predicted and observed crash counts. If a different source of map data, with more information on the geometric or environmental attributes of

segments, becomes available and as volumes of sensor data increase with the growth of connected vehicles or probe data from private industry sources (insurance or technology companies) prediction power is expected to improve significantly. Ensuring time periods for crash data and GPS data overlap should also improve model predictions.

This chapter contains two major contributions to the existing literature. First, despite the rich history of crash modelling and the growing popularity of SSMs, few studies have incorporated SSMs into statistical models of crash frequency, and none have attempted to add SSMs to models of crash severity. In fact, most, if not all, existing models for predicting crashes are based on historical data crashes. The best method for reducing dependence on crash data in network screening is to develop models capable of predicting crash counts and injury levels using SSMs as predictive variables. The presented crash model, which predominantly relies on measures derived from GPS data as the covariates, markedly reduces the necessity for crash data in the network screening process and facilitates a major shift from reactive crash-based models to proactive models based on SSMs. Second, this work is one of the earliest applications of the INLA technique for Bayesian inference to the field of road safety and crash modelling. Though Bayesian techniques are well accepted and highly accurate, most existing estimation techniques (MCMC simulation) are computationally expensive and time consuming. Despite recent advances in Bayesian inference, few studies to date have applied INLA to the field of road safety. Importantly, this model represents one of the first examples of a proactive method for network screening. Rather than relying on crash data (typically available once a year), the developed techniques will allow for screening to be carried out continuously as probe data is continuously collected.

5.2 Limitations

5.2.1 Data Collection, Processing, and Visualization

There were three limitations related to the collection and processing of the GPS data. First, this study assumed that the studied smartphone users are representative of all drivers. While not an exact representation, collecting data from the smartphones of regular drivers represents the least biased method currently feasible for collecting large volumes of GPS travel data, particularly when compared to methods using fleet vehicles or taxis which are inherently biased towards a specific segment of the population. Second, although the three weeks of data contained nearly 22,000 trips and 20 million observations, only part of the road network was covered. Data was not available

for many residential streets, and there were several gaps in more major streets which were also missing data. Overall, data was available on the major freeways, arterials, and collectors. Sites with missing data are unlikely to have a major impact on overall traffic flow patterns or crash models, and so their absence from the data set is unlikely to skew the results. Third, the start and end times of the peak periods used to compute congestion and other traffic flow measures were selected arbitrarily. Obviously, the peak period (and the observed onset and dissipation periods) are not required to start on the hour or last exactly an hour, as the peak period arises naturally from random travel decisions made at the individual level. More advanced analysis is needed if detailed information on the peak periods is required, however even the rather naïve method used here is reasonable for visualizing the effects of congestion

5.2.2 Extracting Surrogate Safety Measures from GPS Data

Additional limitations were related to the extraction and validation of the SSMs. First, preliminary analysis was used to determine the time periods which resulted in the strongest correlation between each SSM and crash frequency. This may not be ideal for correlations with crash severity. Perhaps the greatest limitation of this work, and in fact most surrogate safety studies, is the fact that the temporal coverage of the data used for the SSMs and the crash data do not overlap. However, the assumption underlying the validity of surrogate safety methods is that the relationship between SSMs and safety should remain stable, though more research is needed in this area. Ideally, the temporal coverage would overlap, though there is a significant trade-off between volumes of crash data and GPS data. Several years would be needed to accumulate enough crash data for analysis, and several years of probe vehicle data would also need to be collected. This would result in an extremely large volume of data for analysis. Conversely, using a reasonable amount of probe data would result in too few crashes for analysis. Additionally, direction of travel was not considered when computing the SSMs. Although this is clearly possible, as demonstrated in Chapter 2 with the calculation of congestion, direction of travel was not supported by the crash dataset, and therefore link-level data is not direction dependent. Incorporating direction of travel is theoretically possible and relatively simple, if available in the crash dataset. Cumulative measures, such as HBEs should be relatively insensitive to direction, as should the off-peak measures. CI is likely the only measure affected and could require further consideration. This study used CI measured during the PM peak, while some directions of travel are obviously more congested in the morning.

HBEs (and HAEs) were defined as the minimum (or maximum) acceleration in a serious of consecutively negative (or positive) accelerations. If a constant deceleration bridged over several links, mis-locating the HBE on adjacent links is a possibility. However, because even long periods of constant acceleration or deceleration rarely last more than a few seconds, the impact is likely minimal. Lastly, all traffic flow measures are calculated at the link level and must be aggregated for intersection-level analysis. Despite this, the fit of the intersection models is higher than their link-level counterparts, even if more traffic flow measures are insignificant.

5.2.3 Modelling Crash Frequency and Severity

A final three limitations were noted in during the crash modelling procedure. First, large discrepancies were observed between fitted and predicted values for certain sites, with some sites having more than a hundred crashes predicted where none were observed. R-INLA has a particular difficulty in handling solitary sites (sites without neighbors in the spatial structure) during validation (prediction). However, not all this discrepancy is necessarily a negative. One of the benefits of the surrogate safety model is that it can identify sites with a potential for collisions to occur even if collisions have not occurred there historically. Some of these sites are likely dangerous, according to the identified relationships between risk and SSMs, even if they have not had an observed crash. Second, the neighbourhood structure of 4600 links was comprised of 323 discrete connected components, making the spatial models hardly numerically tractable. One benefit of the Besag proper model is that it is a special case of Markov random fields, implying sparsity in the covariance matrix and making numerical estimation tractable. However, this is negated in a case with so many disparate islands. Third, the neighbourhood structure was defined strictly in terms of connected links, which is likely suboptimal. A distance-weighted neighbourhood structure (measured along the network) would be superior, though this is not supported in the model formulation used.

5.3 Future Work

5.3.1 Data Collection, Processing, and Visualization

With regards to data collection and processing, opportunities for future work revolve around completing the mapping data and completing the GPS data. Although the presented methodology and proof of concept were shown to be successful, some key areas within the OSM data were

incomplete or missing entirely. This is partially due to the collaborative nature of the OSM data, but largely due to issues with downloading and importing the data into a GIS environment. Although the missing data did not affect any of the analysis or results presented, a complete map is certainly necessary for network screening in practice. Alternatively, other map-matching techniques could be explored that are compatible with more stable and complete mapping sources, such as those owned and maintained by the province or municipality.

In addition to some missing map data, there are several isolated links in the network despite data being present on adjacent links. Although it is acceptable for some links to be without data, namely minor residential streets which are unlikely to be prioritized during network screening, methods for imputing data on more major roadways (potentially based on spatial correlation with other links and temporal correlation with other time periods) would provide a great benefit for work in this area. Inverse distance weighting (IDW) has already been considered, though ordinary kriging or other techniques could also be implemented. Importantly, network characteristics should be considered when imputing traffic data in order to avoid, for example, predicted speeds on urban arterials being influenced by slower speeds on nearby residential streets. Additionally, smartphone applications with higher penetration rates (Waze, for example) could be used to ensure more complete data coverage. In the short term, combining multiple source of data, along with reliable imputation methods, is likely to provide the best results. In the long-term, it is anticipated that smartphone data will be complemented or replaced by data collected from smart vehicles that, in addition to GPS, will generate other sensor data including accelerometer or gyroscopic information. As automakers increasingly integrate data collection and communications technologies into their new models, this type of data, along with the measures and methods developed herein, will become abundantly available and easily implemented. Lastly, an integrated system, such as a software platform, for automating the processing and imputation of data could be developed to make this an accessible and practical tool for safety analysts.

5.3.2 Extracting Surrogate Safety Measures from GPS Data

Although this work rigorously compares SSMs to large volumes of historical crash data (in order to meet the requirement that SSMs be predictably related to crashes) more validation of the data and measures is required to prove that SSMs capture the desired events, behaviours, and flow characteristics (in order to meet the requirement that SSMs be physically related to crashes). More

data, especially data collected in multiple cities, would be valuable for confirming the observed relationships. GPS data with a higher frequency of observations, or data from other mobile sensors such as accelerometer or gyroscope, could aid in validating the measures extracted from the smartphone GPS. This study has only scratched the surface regarding the types of safety measures that can be extracted from mobile sensors. In the future, many more measures could be extracted, validated, and integrated into the proposed crash models. Again, other sensors, whether integrated into smartphones or standalone, could be used to provide new and more accurate SSMs. For example, jerk measured using smartphone accelerometers may be an improvement over decelerations measured using GPS. More detailed travel data could yield vehicle manoeuvres related to steering or provide more precise measures of traffic flow.

5.3.3 Modelling Crash Frequency and Severity

The developed crash models are sensitive to the crash assignment method, as discussed in Chapter 4. One of the most critical areas for future work is the consideration of the impact of the crash assignment method on modelling results. Geo-locating crashes on the road network based only on text from police reports is a well-known problem. Although several assignment algorithms were tested, more work is needed to determine the impact of assignment including model sensitivity to crash mis-location. It has been suggested that MCMC simulations could be used to randomly assign crashes in cases of ambiguity to determine the sensitivity of the models to crash assignment and reduce the issues associated with crash mis-location.

In terms of the models themselves, additional flexibility is desired to improve the fit and predictive power. Although spatial correlations were considered, improved methods for defining neighbours, such as a distance-weighted scheme, would be beneficial in improving model accuracy, especially for sites with zero-observed crashes. Temporal correlations could also be considered, allowing for the inclusion of SSMs calculated for different time periods and capturing latent effects throughout the day or seasonal effects throughout the year. Increased flexibility through interaction variables or new model formulations could help to further improve the accuracy of the model. A Poisson point process model for a linear network was recently developed, where crash intensity is modelled using a linear function whose domain corresponds to the edges of the road network. The primary challenge with this model is that GPS data is only available for a subset of the road network. Excluding crashes based on GPS data availability results in a sizeable

reduction in observations for model fitting and invalidates the assumed likelihood, as the likelihood function includes a term for the entire length of the network. Imputation schemes discussed earlier, such as IDW, ordinary kriging, or regression models, would be required. Lastly, site rankings were based only on the posterior mean of expected crashes. More complex ranking criteria, such the posterior probability of excess or the posterior of ranks could be considered in future work.

5.4 Final Remarks

This thesis was successful in demonstrating how mobile sensor data, namely collected from the GPS-enabled smartphones of regular drivers, can be integrated into the network screening process. Not only is processing network wide GPS data practical and accessible, but the types of measures that can be extracted are observed to be consistent with expectation. Measures extracted from the data, whether related to traffic flow or individual events, share statistically significant relationships with both crash frequency and severity, making them viable as SSMs. Moreover, when used as covariates in statistical crash models, these measures are useful in estimating and predicting the frequency and severity of crashes on the network scale. Not only does the proposed screening model demonstrate the practical application of SSMs derived from smartphone GPS data, but, as it controls for other factors of geometry and exposure, it contributes to a better understanding of the complex relationship between the proposed SSMs and crash frequency and severity.

Safety analysis using SSMs provides opportunities to guide improvements of facilities and reduce safety issues. Crashes themselves are not perfect predictors of safety, and surrogate measures would allow practitioners to identify sites with the potential for collisions to occur, regardless if collisions have occurred or have been reported there in the past. The greatest strength of this proposed approach is that, as GPS data is continually collected within a city, site rankings can be continuously updated. Changes to site rankings can occur before additional crashes occur at the most dangerous sites, and this proactive approach has the potential to reduce road traffic crashes, injuries, and fatalities. Regardless of the specific SSMs or modelling techniques, prioritizing sites based on GPS data and SSMs rather than historical crash data represents a substantial contribution to the field of road safety which will allow for a proactive approach to network screening and importantly works towards the minimization of crash data as a necessity in safety evaluation and monitoring.

REFERENCES

1. Abdel-Aty, M. Analysis of driver injury severity levels at multiple locations using. *Journal of Safety Research*, Vol. 34, 2003, pp. 597-603.
2. Transport Canada. Canadian Motor Vehicle Traffic Collision Statistics. *Transport Canada*, 2017. https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2016.html. Accessed June 27, 2018.
3. Hauer, E., J. Kononov, B. Allery, and M. S. Griffith. Screening the Road Network for Sites with Promise. *Transportation Research Record*, no. 1784, 2002, pp. 27-32.
4. Agerholm, N., and H. Lahrmann. Identification of Hazardous Road Locations on the basis of Floating Car Data. *Road safety in a globalised and more sustainable world*, 2012.
5. Aguero-Valverde, J., and P. P. Jovanis. Bayesian Multivariate Poisson Lognormal Models for Crash Severity Modeling and Site Ranking. *Transportation Research Record*, no. 2136, 2009, pp. 82-91.
6. Lu, M. Modelling the effects of road traffic safety measures. *Accident Analysis and Prevention*, no. 38, 2007, pp. 507-517.
7. Abdel-Aty, M., and A. Pande. Identifying crash propensity using specific traffic speed conditions. *Journal of Safety Research*, no. 36, 2005, pp. 97-108.
8. Kockelman, K. M., and Y.-J. Kweon. Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 313-321.
9. Lee, C., B. Hellinga, and K. Ozbay. Quantifying effects of ramp metering on freeway safety. *Accident Anaysis and Prevention*, no. 38, 2006, pp. 279-288.
10. Tarko, A., G. Davis, N. Saunier, T. Sayed, and S. Washington. Surrogate Measures of Safety. Transportation Research Board, 2009.
11. Laureshyn, A., C. Jonsson, T. De Ceunynck, A. Svensson, M. de Goede, N. Saunier, P. Wlodarek, R. van der Horst, and S. Daniels. Review of current study methods for VRU safety. InDeV, Warsaw, Poland, 2016.
12. El Faouzi, N.-E., H. Leung, and A. Kurian. Data fusion in intelligent transportation systems: Progress and challenges – A survey. *Information Fusion*, no. 12, 2011, pp. 4-19.
13. Li, Y., and M. McDonald. Link Travel Time Estimation Using Single GPS Equipped Probe Vehicle. in *The IEEE 5m international Conference on Intelligent Transportation Systems*, Singapore, 2002, pp. 932-937.
14. FHWA. Travel Time Data Collection Handbook. Federal Highway Administration, 1998.
15. Dunlop, M. D., M. Roper, M. Elliot, M. Rebecca, and M. Bruce. Using smartphones in cities to crowdsource dangerous road sections and give effective in-car warnings. in *Proceedings of SEACHI 2016*, 2016, pp. 14-18.
16. Wright, C. A theoretical analysis of the moving observer method. *Transportation Research*, no. 7, 1973, pp. 293-311.
17. Chen, M., and S. I. Chien. Determining the Number of Probe Vehicles for Freeway Travel Time Estimation by Microscopic Simulation. *Transportation Research Record*, no. 1719, 2000, pp. 61-68.
18. Yang, J.-S. Travel Time Prediction Using the GPS Test Vehicle and Kalman Filtering Techniques. in *2005 American Control Conference*, Portland, 2005, pp. 2128-2133.

19. Quiroga, C. A., and D. Bullock. Travel time studies with global positioning and geographic information systems: an integrated methodology. *Transportation Research Part C*, no. 6, 1998, pp. 101-127.
20. D'Este, G. M., R. Zito, and M. A. Tayler. Using GPS to Measure Traffic System Performance. *Computer-Aided Civil and Infrastructure Engineering*, no. 14, 1999, pp. 255-265.
21. Skabardonis, A., K. Petty, H. Noeimi, D. Rydzewski, and P. P. Varaiya. I-880 Field Experiment: Data-Base Development and Incident Delay Estimation Procedures. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1554, 1996, pp. 204-212.
22. Kerner, B. S., C. Demir, R. G. Herrtwich, S. L. Klenov, H. Rehborn, M. Aleksic, and A. Haug. Traffic State Detection with Floating Car Data in Road Networks. in *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*, Vienna, AU, 2005, pp. 700-705.
23. Balan , R. K., N. X. Khoa, and L. Jiang. Real-Time Trip Information Service for a Large Taxi Fleet. in *MobiSys'11*, Bethesda, MD, 2011, pp. 99-112.
24. Moreira-Matias, L., J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas. Predicting Taxi–Passenger Demand Using Streaming Data. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, no. 3, 2013, pp. 1393-1402.
25. Kluger, R., B. L. Smith, H. Park, and D. J. Dailey. Identification of safety-critical events using kinematic vehicle data and the discrete fourier transform. *Accident Analysis and Prevention*, no. 96, 2016, pp. 162-168.
26. Herrera, J. C., D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C*, no. 18, 2010, pp. 568-583.
27. Li, M., J. Dai, S. Sahu, and M. Naphade. Trip Analyzer through Smartphone Apps. in *ACM SIGSPATIAL GIS '11*, Chicago, IL, 2011, pp. 537 - 540.
28. Laureshyn, A. Application of automated video analysis to road user behaviour. Lund University, Lund, PhD Thesis 2010.
29. Bagdadi, O. Assessing safety critical braking events in naturalistic driving studies. *Transportation Research Part F*, no. 16, 2013, pp. 117-126.
30. Saunier, N., T. Sayed, and K. Ismail. An Object Assignment Algorithm for Tracking Performance Evaluation. in *11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Miami, 2002, pp. 9-17.
31. Gettman, D., and L. Head. Surrogate Safety Measures from Traffic Simulation Models. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1840, 2003, pp. 104-115.
32. Saunier, N., and T. Sayed. Automated Analysis of Road Safety. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2019, 2007, pp. 57-64.
33. Bahler, S. J., J. M. Kranig, and E. D. Minge. Field Test of Nonintrusive Traffic Detection Technologies. *Transportation Research Record*, no. 1643, 1998, pp. 161-170.
34. Laureshyn, A., K. Astrom, and K. Brundell-Freij. From Speed Profile Data to Analysis of Behaviour. *IATSS Research*, Vol. 33, no. 2, 2009, pp. 88-98.

35. Anderson-Trocme, P., J. Stipancic, and L. Miranda-Moreno. Performance Evaluation and Error Segregation of Video-Collected Traffic Speed Data. in *Transportation Association Annual Conference Proceedings*, Montreal, 2014.
36. Dingus, T. A., S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. R. Knipling. The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment. NHTSA, Washington, DC, DOT HS 810 593, 2006.
37. Fazeen, M., B. Gozick, R. Dantu, M. Bhukhiya, and M. C. Gonzalez. Safe Driving Using Mobile Phones. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, no. 3, 2012, pp. 1462-1468.
38. Jun, J., J. Ogle, and R. Guensler. Relationships between Crash Involvement and Temporal-Spatial Driving Behavior Activity Patterns Using GPS Instrumented Vehicle Data. in *Transportation Research Board Annual Meeting*, Washington, DC, 2007.
39. Yan, X., M. Abdel-Aty, E. Radwan, X. Wang, and P. Chilakapati. Validating a driving simuator using surrogate safety measures. *Accident Analysis and Prevention*, no. 40, 2008, pp. 274-288.
40. Oh, C., J.-s. Oh, and S. G. Ritchie. Real-time estimation of Freeway Accident Likelihood. in *Transportation Research Board Annual Meeting*, Washington, D.C., 2001.
41. Golob, T. F., W. W. Recker, and V. M. Alvarez. Freeway safety as a function of traffic flow. *Accident Analysis and Prevention*, no. 36, 2004, pp. 933-946.
42. Lee, C., F. Saccomanno, and B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1784, 2002, pp. 1-8.
43. Bhatti, J. A., A. Constant, L.-R. Salmi, M. Chiron, S. Lafont, M. Zins, and E. Lagarde. Impact of retirement on risky driving behaviour and attitudes towards road safety among a large cohort of French drivers (the GAZEL cohort). *Scandanavian Journal of Work, Environment & Health*, Vol. 34, no. 4, 2008, pp. 307-315.
44. Schroeder, B. J., and N. M. Rouphail. Event-Based Modeling of Driver Yielding Behavior at Unsignalized Crosswalks. *Journal of Transportation Engineering*, Vol. 137, 2011, pp. 455-465.
45. Turner, S., K. Fitzpatrick, M. Brewer, and E.S. Park. Motorist Yielding to Pedestrians at Unsignalized Intersections: Findings from a National Study on Improving Pedestrian Safety. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1982, 2006, pp. 1-12.
46. Parker, D., R. West, S. Stradling, and A. S.R. Manstead. Behavioural characteristics and involvement in different types of traffic accident. *Accident Analysis and Prevention*, Vol. 27, no. 4, 1995, pp. 571-581.
47. Ulleberg, P., and T. Rundmo. Personality, attitudes and risk perception as predictors of risky driving behaviour among yong drivers. *Safety Science*, no. 41, 2003, pp. 427-443.
48. Jonah, B. A. Accident Risk and Risk-Taking Behaviour Among Young Drivers. *Accident Analysis and Prevention*, Vol. 18, no. 4, 1986, pp. 255-271.

49. Anastasopoulos, P. C., and F. L. Mannering. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. *Accident Analysis and Prevention*, no. 43, 2011, pp. 1140-1147.
50. Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A*, Vol. 44, no. 5, 2010, pp. 291-305.
51. Chang, L.-Y., and H.-W. Wang. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, no. 38, 2006, pp. 1019-1027.
52. Mannering, F. L., and C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 2014, pp. 1-22.
53. Lord, D., S. P. Washington, and J. H. Ivan. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*, no. 37, 2005, pp. 35-46.
54. Ye, F., and D. Lord. Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models. *Analytic Methods in Accident Research*, no. 1, 2014, pp. 72-85.
55. Yasmin, S., and N. Eluru. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis and Prevention*, no. 59, 2013, pp. 506-521.
56. Biangiardo, M., and M. Cameletti. *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons, Ltd, 2015.
57. Jiang, X., M. Abdel-Aty, and S. Alamili. Application of Poisson random effect models for highway network screening. *Accident Analysis and Prevention*, no. 63, 2014, pp. 74-82.
58. Rue, H., S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society*, no. 71, 2009, pp. 319-392.
59. Huang, H., H. C. Chin, and M.M. Haque. Empirical Evaluation of Alternative Approaches in Identifying Crash Hot Spots: Naive Ranking, Empirical Bayes, and Full Bayes Methods. *Transportation Research Record*, no. 2103, 2009, pp. 32-41.
60. Persuad, B., B. Lan, C. Lyon, and R. Bhim. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. *Accident Analysis and Prevention*, no. 32, 2010, pp. 38-43.
61. Miaou, S.-P., and D. Lord. Modeling Traffic Crash–Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1840, 2003, pp. 31-40.
62. Quddus, M. A. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis and Prevention*, no. 40, 2008, pp. 1486-1497.
63. Wang, C., M. A. Quddus, and S. G. Ison. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention*, no. 43, 2011, pp. 1979-1990.

64. Miaou, S.-P., and J. J. Song. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention*, no. 37, 2005, pp. 699-720.
65. Park, E.S., and D. Lord. Multivariate Poisson–Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2019, 2007, pp. 1-6.
66. Xie, Y., D. Lord, and Y. Zhang. Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis and Prevention*, no. 39, 2007, pp. 922-933.
67. Li, X., D. Lord, Y. Zhang, and Y. Xie. Predicting motor vehicle crashes using Support Vector Machine models. *Accident Analysis and Prevention*, no. 40, 2008, pp. 1611-1618.
68. de Ona, J., R. O. Mujalli, and F. J. Calvo. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis and Prevention*, no. 43, 2011, pp. 402-411.
69. Chong, M., A. Abraham, and M. Paprzycki. Traffic Accident Analysis Using Machine Learning Paradigms. *Informatica*, no. 29, 2005, pp. 89-98.
70. City of Quebec. Mon Trajet. *City of Quebec*, 2015. http://www.ville.quebec.qc.ca/citoyens/deplacements/mon_trajet.aspx. Accessed May 13, 2015.
71. Liu, H. X., and W. Ma. A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials. *Transportation Research Part C*, no. 17, 2009, pp. 11-26.
72. Zhang, H. M. Link-Journey-Speed Model for Arterial Traffic. *Transportation Research Record*, no. 1676, 1999, pp. 109-115.
73. Taylor, M. A.P. Travel through time: the story of research on travel time reliability. *Transportmetrica B: Transport Dynamics*, Vol. 1, no. 3, 2013, pp. 174-194.
74. Taylor, M. A.P., E. Woolley, and Zi. Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transportation Research Part C*, Vol. 8, 2000, pp. 257-285.
75. Sioui, L., and C. Morency. Building congestion indexes from GPS data : Demonstration. in *13th WCTR*, Rio de Janeiro, 2013.
76. Saeedmanesh, M., and N. Geroliminis. Clustering of heterogeneous networks with directional flows based on "Snake"similarities. *Transportation Research Part B*, no. 91, 2016, pp. 250-269.
77. Stipancic, J., L. Miranda-Moreno, A. Labbe, and N. Saunier. Measuring and Visualizing Space-Time Congestion Patters in an Urban Road Network Using Large-Scale Smartphone-Collected GPS Data. *Transportation Letters*, 2017.
78. Dailey, D.J. Travel-time estimation using cross-correlation techniques. *Transportation Research Part B*, Vol. 27B, no. 2, 1993, pp. 97-107.
79. Ostrand, M., K. F. Petty, P. Bickel, J. Jiang, J. Rice, Y. Ritov, and F. Schoenberg. Simple Travel Time Estimation from Single-Trap Loop Detectors. *Intellimotion*, Vol. 6, no. 2, 1997, pp. 4-5, 11.
80. Coifman, B. Evaluating travel times and vehicle trajectories on freeways using dual loop detectors. *Transportation Research Part A*, no. 36, 2002, pp. 351-364.

81. van Lint, J., and N. van der Zijpp. Improving a Travel-Time Estimation Algorithm by Using Dual Loop Detectors. *Transportation Research Record*, no. 1855, 2007, pp. 41-48.
82. Sun, C. C., G. S. Arr, R. P. Ramachandran, and S. G. Ritchie. Vehicle Reidentification Using Multidetector Fusion. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 5, no. 3, 2004, pp. 155-164.
83. Coifman, B., and S. Krishnamurthy. Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. *Transportation Research Part C*, no. 15, 2007, pp. 135-153.
84. Anagnostopoulos, C., T. Alexandropoulos, V. Loumos, and E. Kayafas. Intelligent traffic management through MPEG-7 vehicle flow surveillance. in *EEE John Vincent Atanasoff 2006 International Symposium on Modern Computing* , 2006.
85. Haseman, R. J., J. S. Wasson, and D. M. Bullock. Real-Time Measurement of Travel Time Delay in Work Zones and Evaluation Metrics Using Bluetooth Probe Tracking. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2169, 2010, pp. 40-53.
86. Coifman, B., and M. Cassidy. Vehicle reidentification and travel time measurement on congested freeways. *Transportation Research Part A*, Vol. 36, 2002, pp. 899-917.
87. Kwong, K., R. Kavaler, R. Rajagopal, and P. Varaiya. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C*, no. 17, 2009, pp. 586-606.
88. Stipancic, J., L. Miranda-Moreno, and N. Saunier. The Who and Where of Road Safety: Extracting Surrogate Indicators From Smartphone Collected GPS Data in Urban Envrionments. in *Transportation Research Board Annual Meeting 2016*, Washington, DC, 2016.
89. Shi, Q., and M. Abdel-Aty. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C*, Vol. 58, 2015, pp. 380-394.
90. Palen, J. The Need for Surveillance in Intelligent Transportation Systems. *Intellimotion*, Vol. 6, no. 1, 1997, pp. 1-3.
91. Falcocchio, J. C., and H. S. Levinson. Measuring Traffic Congestion, in *Road Traffic Congestion: A Concise Guide*.: Springer International Publishing, 2015, pp. 93-110.
92. Skabardonis, A., P. Varaiya, and K. F. Petty. Measuring Recurrent and Nonrecurrent Traffic Congestion. *Transportation Research Record*, no. 1856, 2003, pp. 118-124.
93. ADEC. Évaluation des coûts de la congestion routière dans la région de Montréal pour les conditions de référence de 2008. Ministère des Transports du Québec, Quebec City, 2014.
94. Dias, C., M. Miska, M. Kuwahara, and H. Warita. Relationship between congestion and traffic accidents on expressways: an investigation with Bayesian belief networks. in *Proceedings of 40th Annual Meeting of Infrastructure Planning (JSCE)*, Japan, 2009.
95. Aftabuzzaman, M. Measuring Traffic Congestion - A Critical Review. in *30th Australasian Transport Research Forum*, 2007.
96. Bachman, C. Multi-Sensor Data Fusion for Traffic Speed and Travel. University of Toronto, Toronto, Masters Thesis 2011.
97. Marchal, F. TrackMatching. 2015. https://mapmatching.3scale.net/. Accessed May 1, 2015.

98. OpenStreetMap. About. *OpenStreetMap*, 2015. http://www.openstreetmap.org/about. Accessed May 11, 2015.
99. Marchal, F., J. Hackney, and K. W. Axhausen. Efficient Map Matching of Large Global Positioning System Data Sets. *Transportation Research Record*, no. 1935, 2005, pp. 93-100.
100. Zaki, M. H., T. Sayed, and K. Shaaban. Use of Drivers' Jerk Profiles in Computer Vision–Based Traffic Safety Evaluations. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2434, 2014, pp. 103-112.
101. Bagdadi, O., and A. Varhelyi. Development of a method for detecting jerks in safety critical events. *Accident Analysis and Prevention*, no. 50, 2013, pp. 83-91.
102. Burns, S., L. Miranda-Moreno, J. Stipancic, N. Saunier, and K. Ismail. Accessible and Practical Geocoding Method for Traffic Collision Record Mapping. *Transportation Research Record*, no. 2460, 2014, pp. 39-46.
103. Kartika, C. S. Visual Exploration of Spatial-Temporal Traffic Congestion Patterns Using Floating Car Data. Technische Universität München, Munich, Masters Thesis 2015.
104. Brisk Synergies. *Brisk Synergies*, 2015. http://www.brisksynergies.com/. Accessed July 22, 2015.
105. Wu, K.-F., and P. P. Jovanis. Defining and screening crash surrogate events using naturalistic driving data. *Accident Analysis and Prevention*, no. 61, 2013, pp. 10-22.
106. Woodward, D., G. Nogin, P. Koch, D. Racz, M. Goldszmidt, and E. Horvits. Predicting Travel Time Reliability using Mobile Phone GPS Data. Microsoft Research, Redmond, WA, 2015.
107. Stipancic, J., L. Miranda-Moreno, and N. Saunier. Vehicle Manoeuvres as Surrogate Safety Measures: Extracting Data From the GPS-Enables Smartphones of Regular Drivers. *Accident Analysis and Prevention*, Vol. 115, 2018, pp. 160-169.
108. Stipancic, J., L. Miranda-Moreno, and N. Saunier. The Impact of Congestion and Traffic Flow on Crash Frequency and Severity: An Application of Smartphone-Collected GPS Travel Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2659, 2017, pp. 43-54.
109. Johnson, D. A., and M. M. Trivedi. Driving Style Recognition Using a Smartphone as a Sensor Platform. in *2011 14th International IEEE Conference on Intelligent Transportation Systems*, Washington, DC, 2011, pp. 1609-1615.
110. Eren, H., S. Makinist, E. Akin, and a. Yilmaz. Estimating Driving Behavior by a Smartphone. in *2012 Intelligent Vehicles Symposium*, Alcalá de Henares, Spain, 2012, pp. 234-239.
111. Guido, G., A. Vitale, V. Astarita, F. Saccomanno, V. P. Giofre, and V. Gallelli. Estimation of safety performance measures from smartphone SENSORS. *Procedia - Social and Behavioral Sciences*, no. 54, 2012, pp. 1095-1103.
112. Noland, R. B., and M. M. Quddus. Congestion and safety: a spatial analysis of London. *Transportation Research Part A: Policy and Practice*, Vol. 39, 2005, pp. 737-754.
113. Quddus, M. A., C. Wang, and S. G. Ison. Road Traffic Congestion and Crash Severity: An Econometric Analysis Using Ordered Response Models. *Journal of Transportation Engineering*, Vol. 136, no. 5, 2010, pp. 424-435.

114. Martin, J.-L. Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis and Prevention*, Vol. 34, no. 619-229, 2002.
115. Zhou, M., and V. P. Sisiopiku. Relationship Between Volume-to-Capacity Ratios and Accident Rates. *Transportation Research Record*, Vol. 1581, 1997, pp. 47-52.
116. Wang, C., M. A. Quddus, and S. G. Ison. Impact of tra c congestion on road accidents: a spatial analysis of the M25 motorway in England. *Accident Analysis and Prevention* , Vol. 41, no. 4, 2009, pp. 798-808.
117. Moreno, A.T., and A. Garcia. Use of speed profile as surrogate measure: Effect of traffic calming devices oncrosstown road safety performance. *Accident Analysis and Prevention*, no. 61, 2013, pp. 23-32.
118. Boonsiripant, S. Speed profile variation as a surrogate measure of road safety based on GPS-equipped vehicle data. Georgia Institute of Technology, PhD Thesis 2009.
119. Ko, J., R. Guensler, M. Hunter, and H. Li. Instrumented Vehicle Measured Speed Variation and Freeway Traffic Congestion. in *Applications of Advanced Technology in Transportation*, Chicago, 2006, pp. 356-361.
120. Federal Highway Administration. Surrogate Safety Assessment Model and Validation: Final Report. U.S. Department of Transportation, McLean, VA, FHWA-HRT-08-051 2008.
121. Elvik, R. The power model of the relationship between speed and road safety: update and new analyses. Institute of Transport Economics, Oslo, Norway, 2009.
122. Park, P. Y., and R. Sahaji. Safety network screening for municipalities with incomplete traffic volume data. *Accident Analysis and Prevention*, no. 50, 2013, pp. 1062-1072.
123. Miranda-Moreno, L., A. Labbe, and L. Fu. Bayesian multiple testing procedures for hotspot identification. *Accident Analysis and Prevention*, Vol. 39, no. 6, 2007, pp. 1192-1201.
124. Heydecker, B. G., and J. Wu. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. *Advances in Engineering Software*, Vol. 32, no. 10-11, 2001, pp. 859-869.
125. Strauss, J., L. F. Miranda-Moreno, and P. Morency. Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. *Accident Analysis and Prevention*, no. 83, 2015, pp. 132-142.
126. Stipancic, J., L. Miranda-Moreno, N. Saunier, and A. Labbe. Surrogate Safety and Network Screening: Modelling Crash Frequency Using GPS Data and Latent Gaussian Models. *Accident Analysis and Prevention*, Vol. 120, 2018, pp. 174-187.
127. Stipancic, J., L. Miranda-Moreno, N. Saunier, and A. Labbe. Network Screening for Large Urban Road Networks: Using GPS Data and Surrogate Measures to Model Crash Frequency and Severity. *Accident Analysis and Prevention*, Vol. in revision, 2018.
128. Hauer, E. Empircal bayes approach to the estimation of "unsafety": the multivariate regression method. *Accident Analysis and Prevention*, Vol. 24, no. 5, 1992, pp. 457-477.
129. Mountain, L., B. Fawaz, and D. Jarret. Accident prediction models for roads with minor junctions. *Accident Analysis and Prevention*, Vol. 28, no. 6, 1996, pp. 695-707.
130. El-Basyouny, K., and T. Sayed. Accident prediction models with random corridor parameters. *Accident Analysis and Prevention*, Vol. 41, no. 5, 2009, pp. 1118–1123.

131. Aguero-Valverde, J., and P. P. Jovanis. Analysis of Road Crash Frequency with Spatial Models. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2061, 2008, pp. 55-63.
132. Song, J., M. Ghosh, S. Milaou, and B. Mallick. Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, Vol. 97, no. 1, 2006, pp. 246-273.
133. Hu, S., J. N. Ivan, N. Raishanker, and J. Mooradian. Temporal modeling of highway crash counts for senior and non-senior drivers. *Accident Analysis and Prevention*, Vol. 50, 2013, pp. 1003-1013.
134. Serhiyenko, V., J. N. Ivan, N. Ravishanker, and M.S. Islam. Dynamic compositional modeling of pedestrian crash counts on urban roads in Connecticut. *Accident Analysis and Prevention*, Vol. 64, 2014, pp. 78-85.
135. Serhiyenko, V., S. A. Mamun, J. N. Ivan, and N. Ravishanker. Fast Bayesian inference for modeling multivariate crash counts. *Analytic Methods in Accident Research*, Vol. 9, 2016, pp. 44-53.
136. Ariza, A. Validation of Road Safety Surrogate Measures as a Predictor of Crash Frequency Rates on a Large-Scale Microsimulation Network. University of Toronto, Toronto, Masters Thesis 2011.
137. Lorion, A. Investiagation of Surrogate Measures for Safety Assessment of Two-Way Stop Controlled Intersections. Ryerson University, Toronto, Masters Thesis 2014.
138. Li, S., Q. Xiang, M. Yongfen, X. Gu, and H. Li. Crash Risk Prediction Modeling Based on the Traffic Conflict Technique and a Microscopic Simulation for Freeway Interchange Merging Areas. *International Journal of Environmental Research and Public Health*, Vol. 13, no. 1157, 2016.
139. Ellison, A. B., S. Greaves, and M. Bliemer. Examining Heterogeneity of Driver Behavior with Temporal and Spatial Factors. *Transportation Research Record*, no. 2386, 2013, pp. 158-157.
140. Kluger, R. Identifying Crashes and Near-Crashes Using Vehicle Trajectory Data to Support Roadway Safety Analysis in Connected Vehicle Environments. in *Transportation Research Board Annual Meeting*, Washington, DC, 2017.
141. Latouche, A., C. Guihenneuc-Jouyaux, C. Girard, and D. Hémon. Robustness of the BYM model in absence of spatial variation in the residuals. *International Journal of Health Geographics*, Vol. 6, no. 39, 2007.
142. Besag, J., J. York, and A. Mollie. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, no. 43, 1991, pp. 1-59.
143. Radford, N. M. MCMC using Hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo*, Brooks, S., A. Gelman, G. L. Jones, and X.-L. Meng.: CRC Press, ch. 5, 2011, pp. 113-162.
144. Schrodle, B., and L. Held. A primer on disease mapping and ecological regression using INLA. *Computation Statistics*, no. 26, 2011, pp. 241-258.
145. Economic Analysis Directorate of Transport Canada. Estimates of the Full Cost of Transportation in Canada. Transport Canada, TP 14819E, 2008.

146. Miranda-Moreno, L. F., and L. Fu. A Comparative Study of Alternative Model Structures and Criteria for Ranking Locations for Safety Improvements. *Networks and Spatial Economics*, no. 6, 2006, pp. 97-110.
147. Québec. Débit de Circulation. *Données Québec*, https://www.donneesquebec.ca/recherche/fr/dataset/debits-de-circulation-transports-quebec/resource/9de14998-2e3b-4936-a587-2da4f3ddd3af. Accessed July 10, 2018.
148. Imprialou, M.-I. M., M. Quddus, D. E. Pitfield, and D. Lord. Re-visiting crash–speed relationships: A new perspective in crash modelling. *Accident Analysis and Prevention*, no. 86, 2016, pp. 173-185.

APPENDIX A: SELECTION OF TIME PERIODS FOR SSM VARIABLES

Chapter 3 describes the extraction of the considered SSMs from the GPS data. Although each SSM can be calculated for various time periods, a single period was consistently used for each traffic flow SSM throughout this thesis (values used for HBEs and HAEs were simply the sum of all events observed across the entire data collection period. In Chapter 3, it is stated that preliminary analysis was used to determine the time periods which resulted in the strongest correlation between each SSM and crash frequency. This preliminary analysis is summarized in the following tables. Table A-1 provides the correlation analysis at the link level, and Table A-2 provides the analysis at the intersection level. The colouring visualizes the correlation strengths from high positive correlation (green) to high negative correlation (red). The considered traffic flow SSMs were calculated for five different time periods of AM peak, PM peak, both peak periods combined, off-peak, and peak and off-peak combined. The strengths were tested considering the buffer sizes used in Chapter 3. Note that CI could not be calculated for two periods as it requires both peak and off-peak measured speeds.

Starting with Table A-1, it is clear that CI in the PM peak period shares stronger correlations with crash frequency than either the AM peak or peak periods combined. This pattern is subsequently observed at the intersection level in Table A-2. Therefore, the PM peak was the selected period for congestion. Although average speed during the off-peak period generally had the weakest correlations with crashes, the off-peak period was selected to avoid high correlations with CI (in general, high congestion results in low speeds). Additionally, the correlation strengths during the off-peak period were still moderately strong. Finally, for CVS, the correlation strengths observed in the off-peak period were, on average, slightly higher than for other time periods. Therefore, the off-peak period was chosen for the computation of CVS. In general, the analysis at both the link and intersection level demonstrate that correlations between CVS and crash frequency are moderately strong regardless of the period considered.

Although CVS was ultimately chosen as the measure to capture speed variation, initial tests were performed at the link level on several possible measures found in the existing literature. Table A-3 summarizes a similar correlation analysis on five other measures of speed variation. The primary observation from this table is that CVS had much stronger correlations than other potential measures of speed variation and was therefore selected as an SSM for this work. Although other measures contained in this table could also be considered SSMs, their high correlation with CVS limits the additional explanatory power of these variables.

TABLE A-1 Correlation analysis between SSMs and crash frequency at the link level

Measure	Functional Class	50 m AM	50 m PM	50 m PEAK	50 m OFF	50 m ALL	100 m AM	100 m PM	100 m PEAK	100 m OFF	100 m ALL
Congestion Index	Motorway	-0.037	0.054	0.006	*N/A*	*N/A*	-0.017	0.055	-0.016	*N/A*	*N/A*
	Primary	0.001	0.211	0.078	*N/A*	*N/A*	0.016	0.237	0.101	*N/A*	*N/A*
	Secondary	0.074	0.115	0.055	*N/A*	*N/A*	0.101	0.162	0.111	*N/A*	*N/A*
	Tertiary	0.084	0.119	0.085	*N/A*	*N/A*	0.077	0.106	0.073	*N/A*	*N/A*
	Residential	0.066	0.084	0.062	*N/A*	*N/A*	0.059	0.072	0.050	*N/A*	*N/A*
Average Speed	Motorway	-0.300	-0.298	-0.329	-0.267	-0.330	-0.284	-0.290	-0.313	-0.258	-0.311
	Primary	-0.356	-0.375	-0.403	-0.352	-0.411	-0.414	-0.396	-0.439	-0.387	-0.449
	Secondary	-0.479	-0.488	-0.492	-0.415	-0.497	-0.510	-0.529	-0.527	-0.446	-0.530
	Tertiary	-0.355	-0.383	-0.412	-0.225	-0.416	-0.375	-0.396	-0.425	-0.232	-0.428
	Residential	0.001	-0.034	-0.068	0.047	-0.085	-0.042	-0.078	-0.105	0.003	-0.119
Coefficient of Variation	Motorway	0.090	0.125	0.104	0.167	0.107	0.082	0.085	0.059	0.162	0.061
	Primary	0.120	0.163	0.122	0.158	0.137	0.101	0.152	0.104	0.122	0.121
	Secondary	0.083	0.095	0.085	0.104	0.095	0.030	0.052	0.039	0.027	0.052
	Tertiary	0.148	0.143	0.183	0.161	0.179	0.101	0.098	0.134	0.124	0.132
	Residential	0.172	0.175	0.191	0.153	0.182	0.123	0.130	0.154	0.106	0.143

TABLE A-2 Correlation analysis between SSMs and crash frequency at the intersection level

Measure	Functional Class	100 m AM	100 m PM	100 m PEAK	100 m OFF	100 m ALL	200 m AM	200 m PM	200 m PEAK	200 m OFF	200 m ALL
Congestion Index	Motorway	-0.007	0.022	-0.026	*N/A*	*N/A*	0.032	0.131	0.047	*N/A*	*N/A*
	Primary	0.093	0.183	0.126	*N/A*	*N/A*	0.091	0.291	0.165	*N/A*	*N/A*
	Secondary	0.078	0.110	0.077	*N/A*	*N/A*	0.140	0.195	0.154	*N/A*	*N/A*
	Tertiary	0.125	0.149	0.142	*N/A*	*N/A*	0.134	0.169	0.136	*N/A*	*N/A*
	Residential	0.033	0.093	0.058	*N/A*	*N/A*	0.061	0.140	0.087	*N/A*	*N/A*
Average Speed	Motorway	-0.167	-0.142	-0.162	-0.136	-0.160	-0.227	-0.252	-0.232	-0.224	-0.234
	Primary	-0.435	-0.419	-0.425	-0.454	-0.436	-0.444	-0.480	-0.458	-0.473	-0.469
	Secondary	-0.372	-0.365	-0.371	-0.375	-0.367	-0.473	-0.471	-0.476	-0.440	-0.481
	Tertiary	-0.255	-0.274	-0.277	-0.184	-0.270	-0.343	-0.373	-0.370	-0.282	-0.361
	Residential	-0.014	-0.052	-0.034	-0.002	-0.033	-0.133	-0.150	-0.143	-0.079	-0.138
Coefficient of Variation	Motorway	0.158	0.131	0.128	0.195	0.127	0.211	0.253	0.219	0.323	0.229
	Primary	0.321	0.301	0.296	0.376	0.310	0.329	0.372	0.308	0.334	0.314
	Secondary	0.345	0.327	0.323	0.360	0.333	0.404	0.392	0.396	0.403	0.401
	Tertiary	0.272	0.272	0.312	0.196	0.309	0.286	0.342	0.360	0.205	0.354
	Residential	0.106	0.141	0.130	0.128	0.122	0.161	0.214	0.200	0.180	0.188

TABLE A-3 Correlation analysis between measures of speed variation and crash frequency

Measure	Functional Class	50 m					100 m				
		AM	PM	PEAK	OFF	ALL	AM	PM	PEAK	OFF	ALL
Acceleration Noise	Motorway	0.163	0.162	0.179	0.167	0.173	0.149	0.126	0.133	0.161	0.128
	Primary	0.112	0.127	0.115	0.174	0.118	0.060	0.095	0.087	0.117	0.089
	Secondary	0.047	0.081	0.057	0.088	0.066	-0.017	0.015	-0.004	0.011	0.006
	Tertiary	0.017	0.009	-0.016	0.108	-0.017	-0.049	-0.057	-0.090	0.063	-0.088
	Residential	0.132	0.152	0.132	0.139	0.127	0.074	0.094	0.077	0.088	0.072
Standard Deviation	Motorway	-0.038	-0.012	-0.068	0.043	-0.063	-0.042	-0.051	-0.099	0.035	-0.094
	Primary	-0.054	-0.060	-0.094	0.002	-0.081	-0.115	-0.110	-0.159	-0.059	-0.145
	Secondary	-0.064	-0.096	-0.113	0.000	-0.128	-0.127	-0.140	-0.164	-0.072	-0.177
	Tertiary	-0.008	-0.025	-0.016	0.051	-0.024	-0.070	-0.084	-0.075	0.010	-0.082
	Residential	0.116	0.106	0.102	0.126	0.087	0.050	0.049	0.045	0.070	0.029
Difference	Motorway	-0.052	0.023	-0.051	0.011	-0.051	-0.042	-0.016	-0.082	0.043	-0.079
	Primary	0.022	0.062	0.062	0.064	0.079	-0.020	0.048	0.036	0.022	0.050
	Secondary	0.056	0.038	0.063	0.024	0.076	-0.008	0.004	0.012	-0.039	0.029
	Tertiary	0.079	0.080	0.094	0.107	0.086	0.032	0.037	0.046	0.078	0.041
	Residential	0.150	0.141	0.143	0.135	0.131	0.090	0.083	0.090	0.090	0.076
Accumulated Speed Uniformity	Motorway	-0.042	-0.003	-0.061	0.040	-0.058	-0.049	-0.044	-0.094	0.036	-0.091
	Primary	-0.021	-0.027	-0.054	0.033	-0.040	-0.075	-0.074	-0.109	-0.026	-0.095
	Secondary	-0.045	-0.072	-0.080	0.007	-0.093	-0.109	-0.116	-0.133	-0.071	-0.144
	Tertiary	-0.001	-0.017	-0.005	0.049	-0.009	-0.069	-0.076	-0.069	0.007	-0.071
	Residential	0.118	0.107	0.103	0.127	0.088	0.052	0.051	0.046	0.074	0.030
Accumulated Speeding	Motorway	-0.015	0.029	0.012	0.018	0.014	0.005	0.011	-0.001	0.029	0.003
	Primary	-0.024	0.044	0.012	0.018	0.007	-0.051	0.046	-0.011	-0.014	-0.016
	Secondary	0.035	0.062	0.041	-0.018	0.051	-0.014	0.028	-0.003	-0.047	0.013
	Tertiary	0.084	0.086	0.119	0.098	0.117	0.064	0.053	0.088	0.081	0.087
	Residential	0.158	0.153	0.171	0.141	0.162	0.109	0.100	0.126	0.097	0.120

APPENDIX B: ADDITIONAL CDFS FOR CRASH SEVERITY TESTING

The methods for testing relationships between the proposed SSMs and crash frequency and severity are outlined in Chapter 3. A series of pairwise K-S tests are used to test the CDFs of sites with minor, major, and fatal crashes in order to determine if the SSMs share any statistically signifcant relationships with crash severity. The results of the K-S tests are summarized in Table 3-4 and Table 3-7. In addition to these tables, Figure 3-3 and Figure 3-5 provide graphical examples of the CDFs being tested. Within the chapter, only two plots are provided for each SSM for brevity (one plot for all data, and another for a representative functional classification). For completeness, the remaining CDFs are presented in Figure B-1 through Figure B-10. For HBEs and HAEs, the figures provided CDFs considering acceleration thresholds of -2 m/s^2 and 2 m/s^2 respectively.

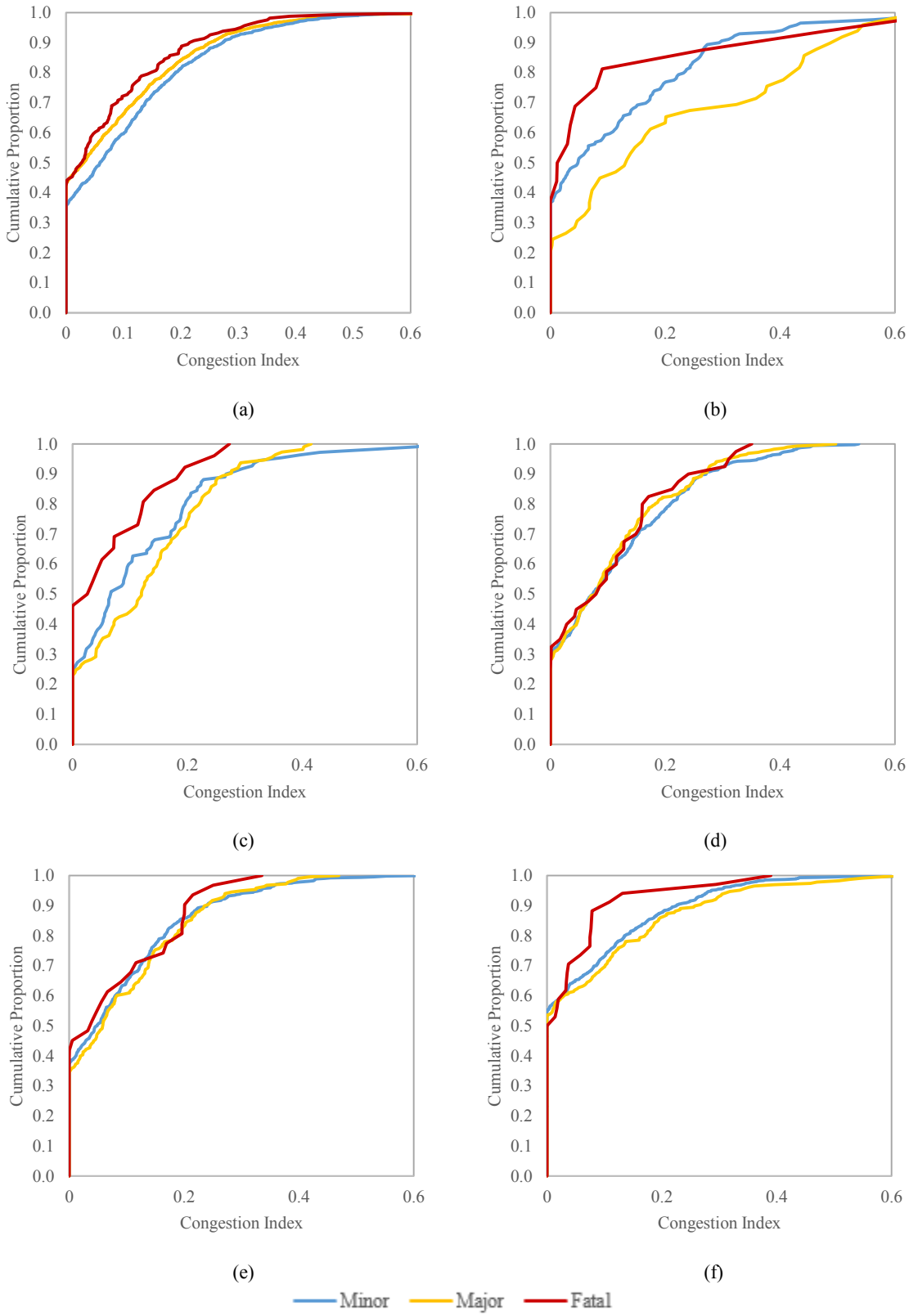


FIGURE B-1 Cumulative distributions for CI on links, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

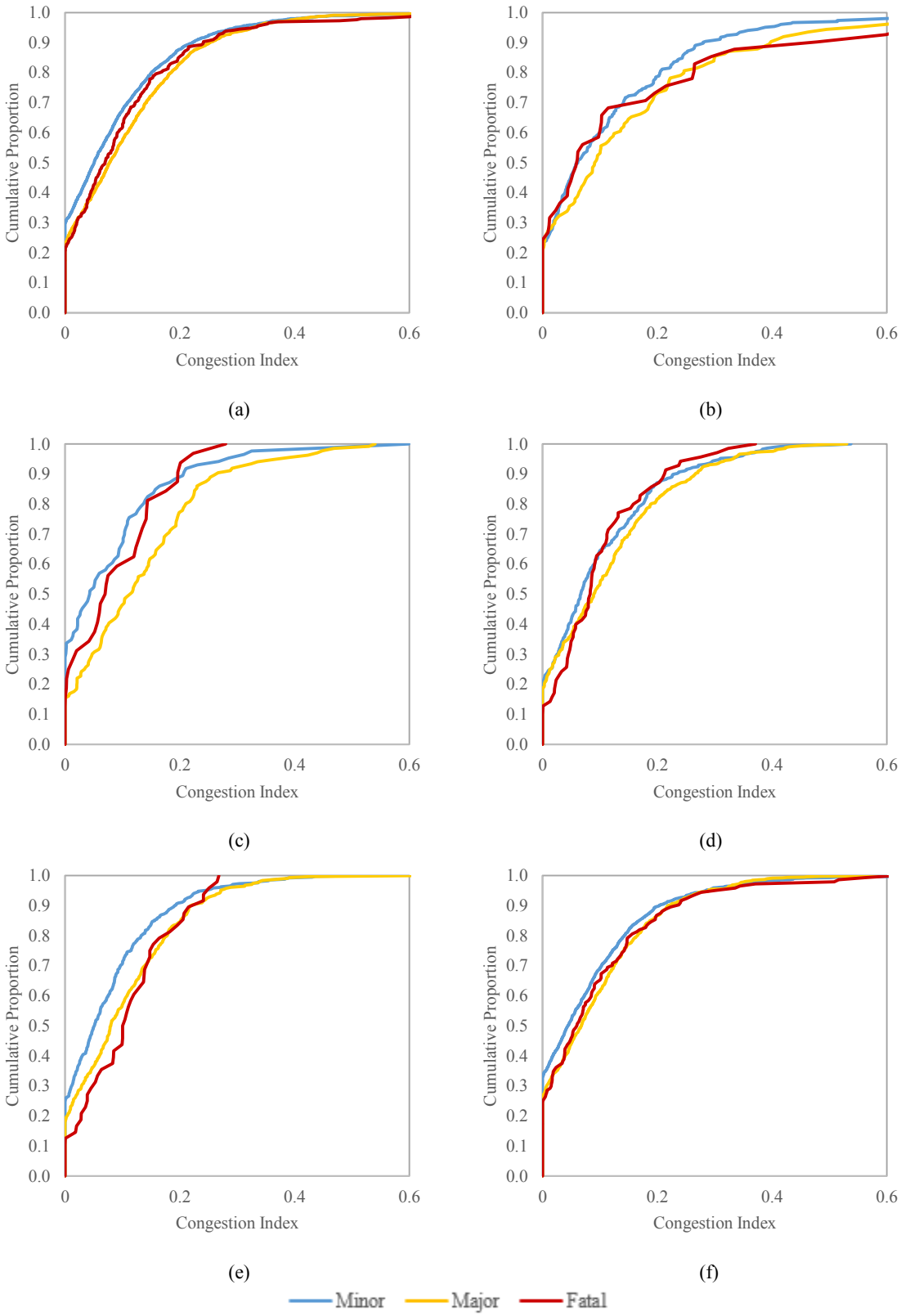


FIGURE B-2 Cumulative distributions for CI at intersections, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

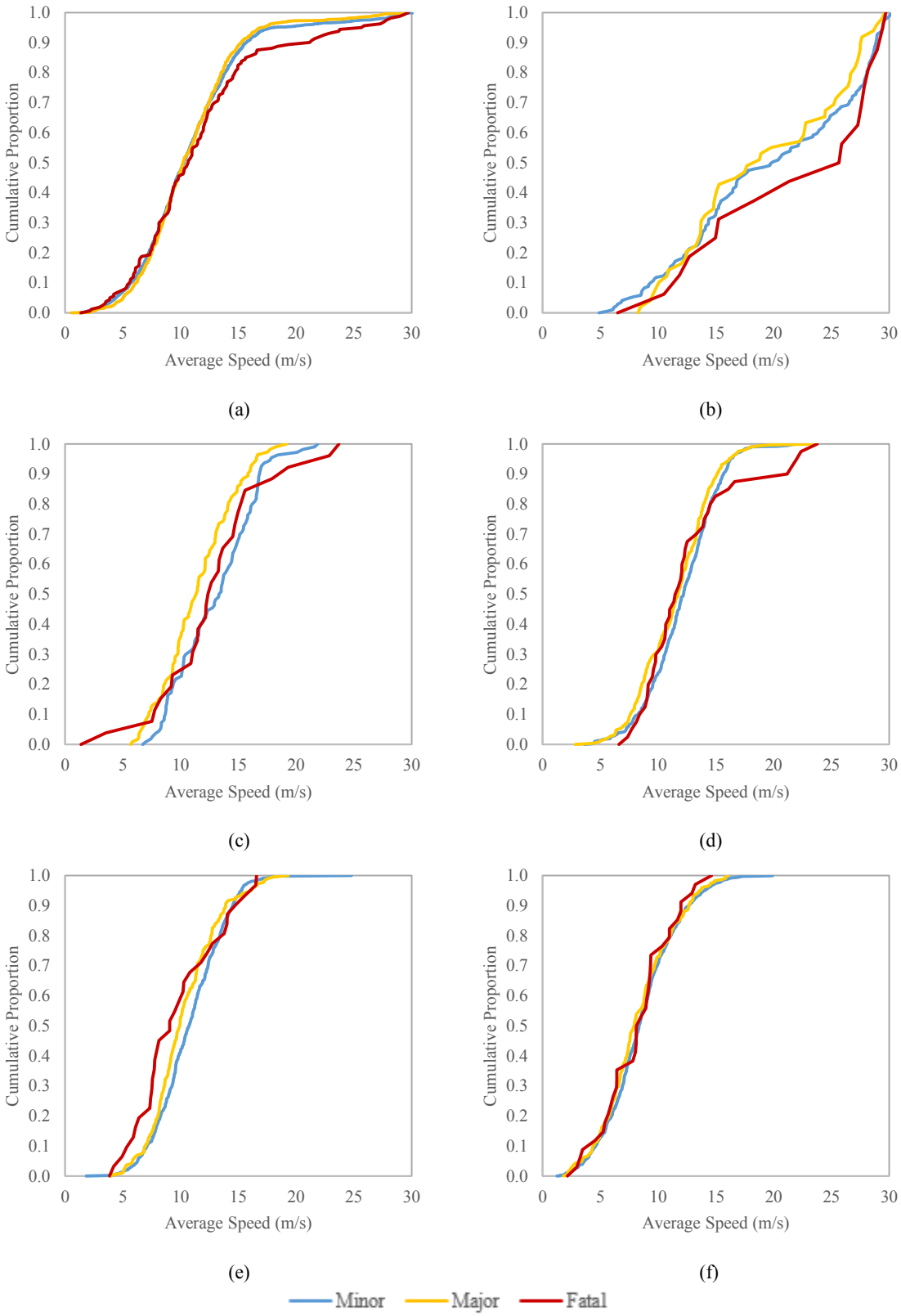


FIGURE B-3 Cumulative distributions for $\bar{V}$ on links, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

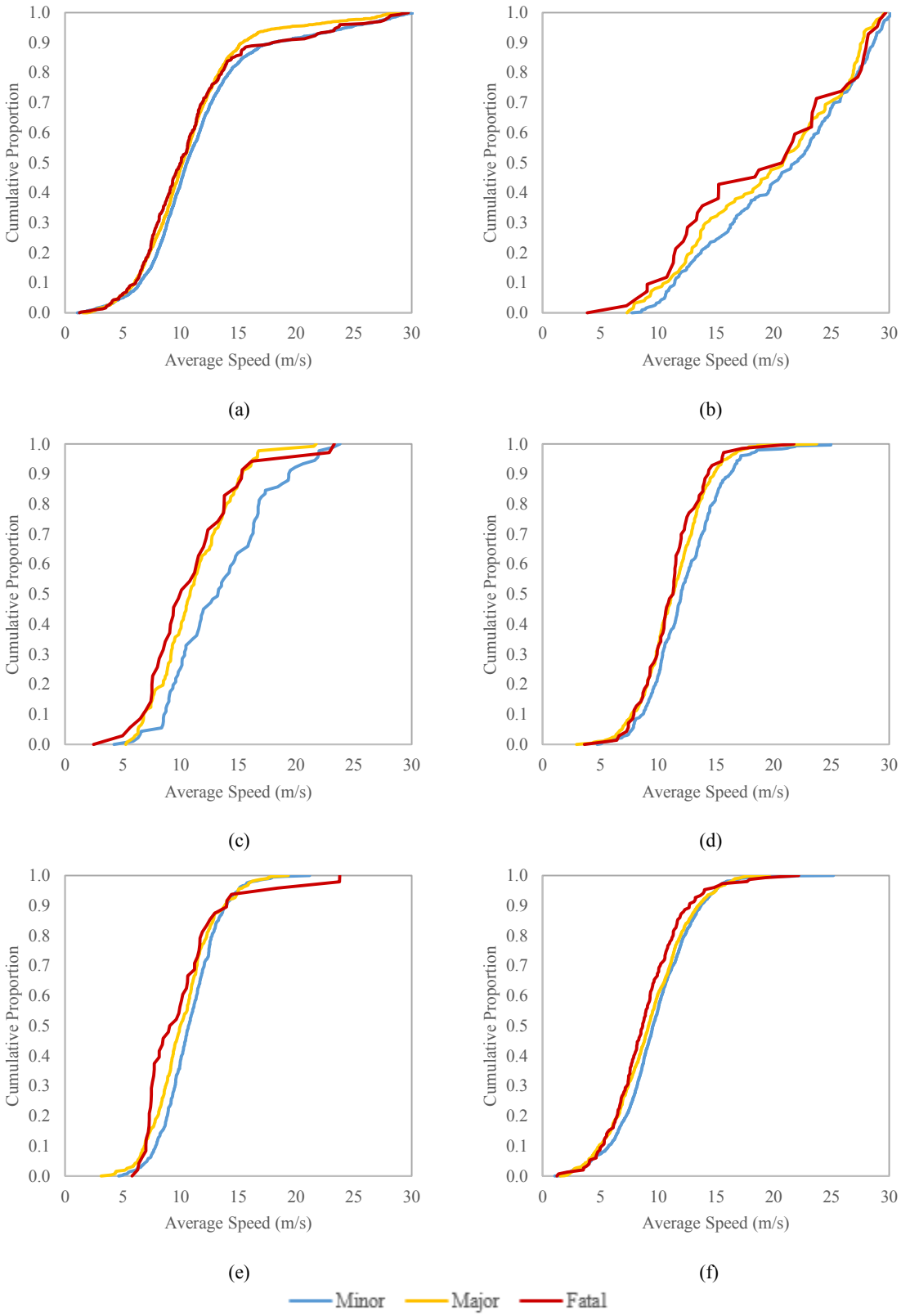


FIGURE B-4 Cumulative distributions for $\bar{V}$ at intersections, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

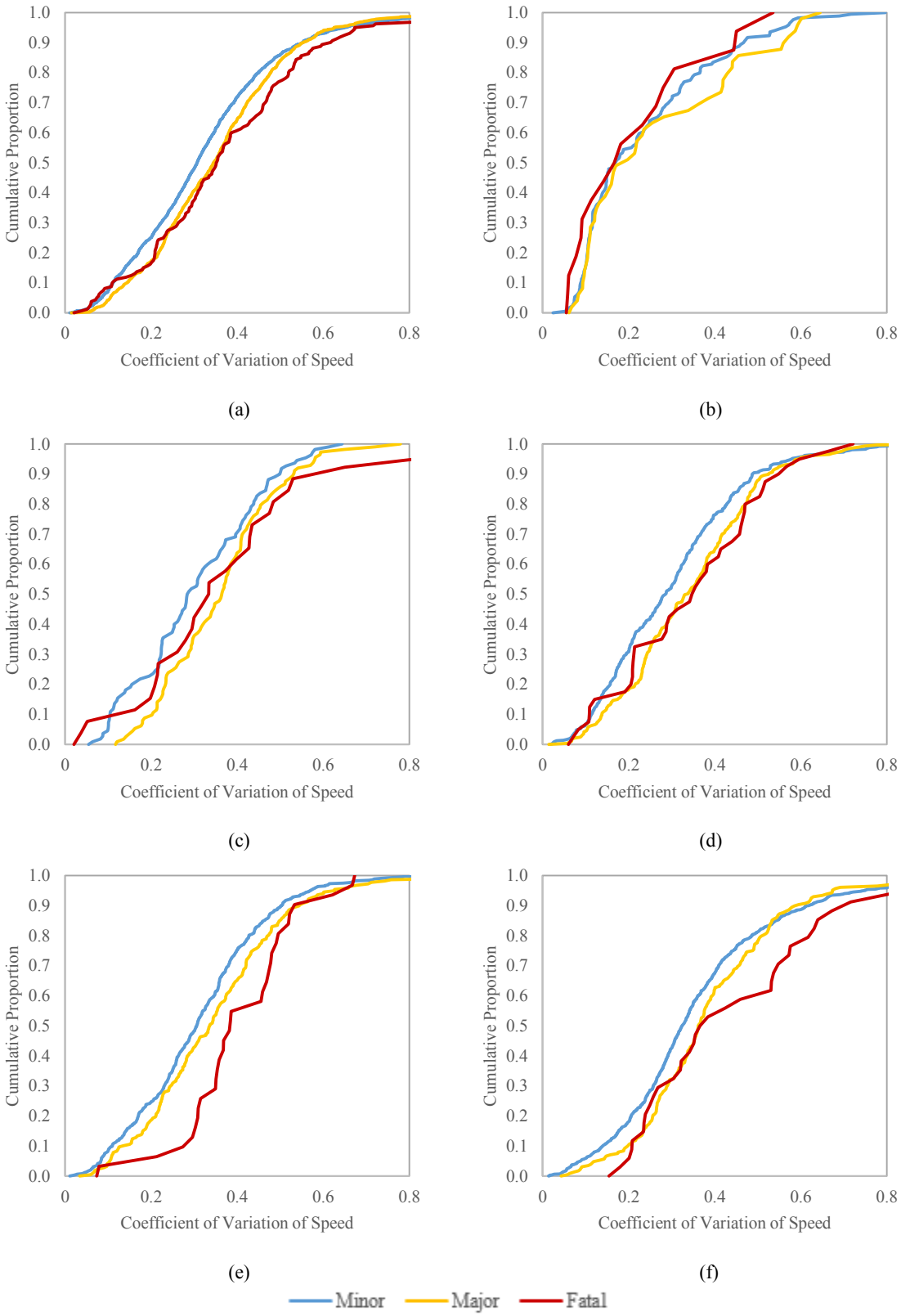


FIGURE B-5 Cumulative distributions for CVS on links, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

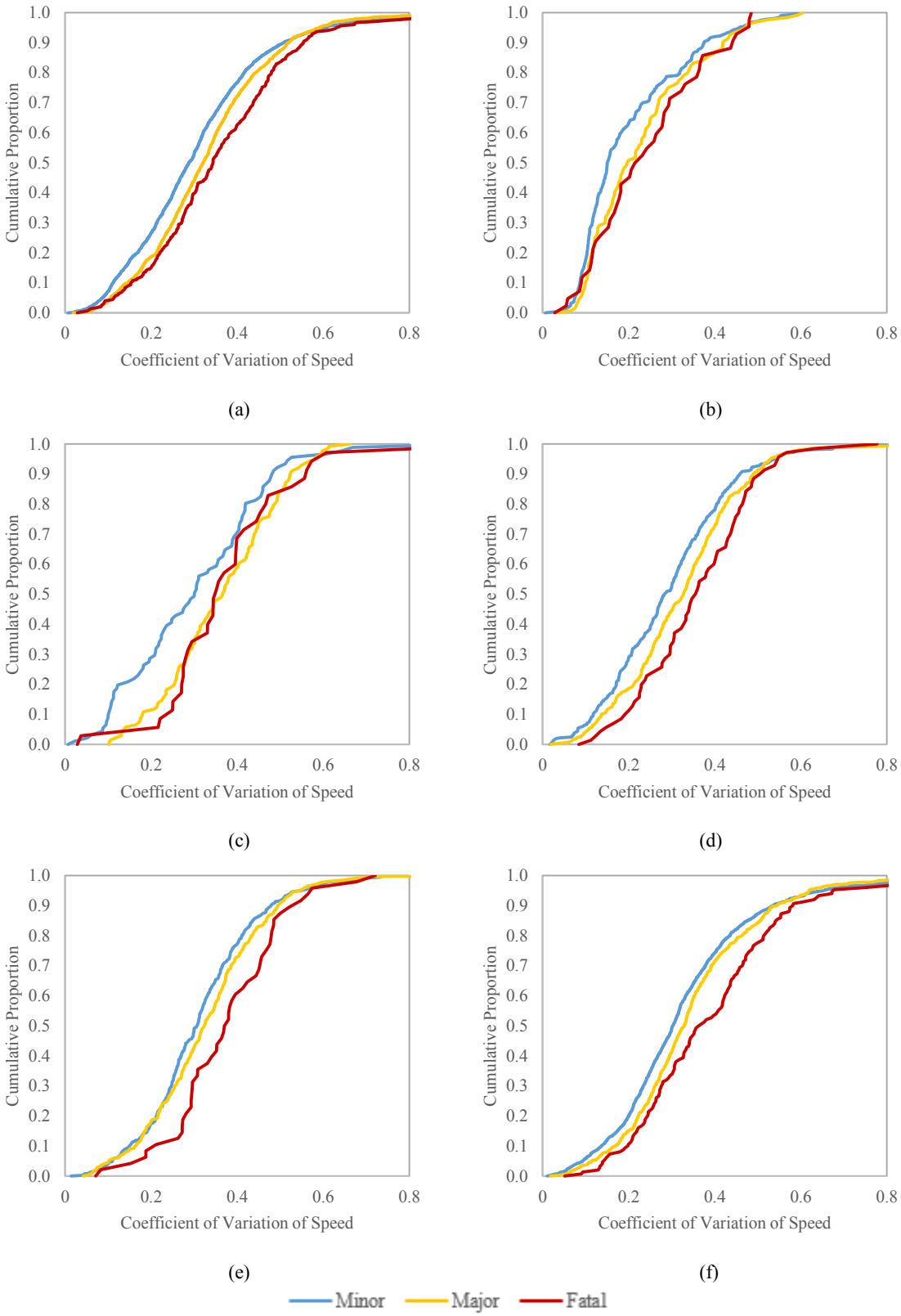


FIGURE B-6 Cumulative distributions for CVS at intersections, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

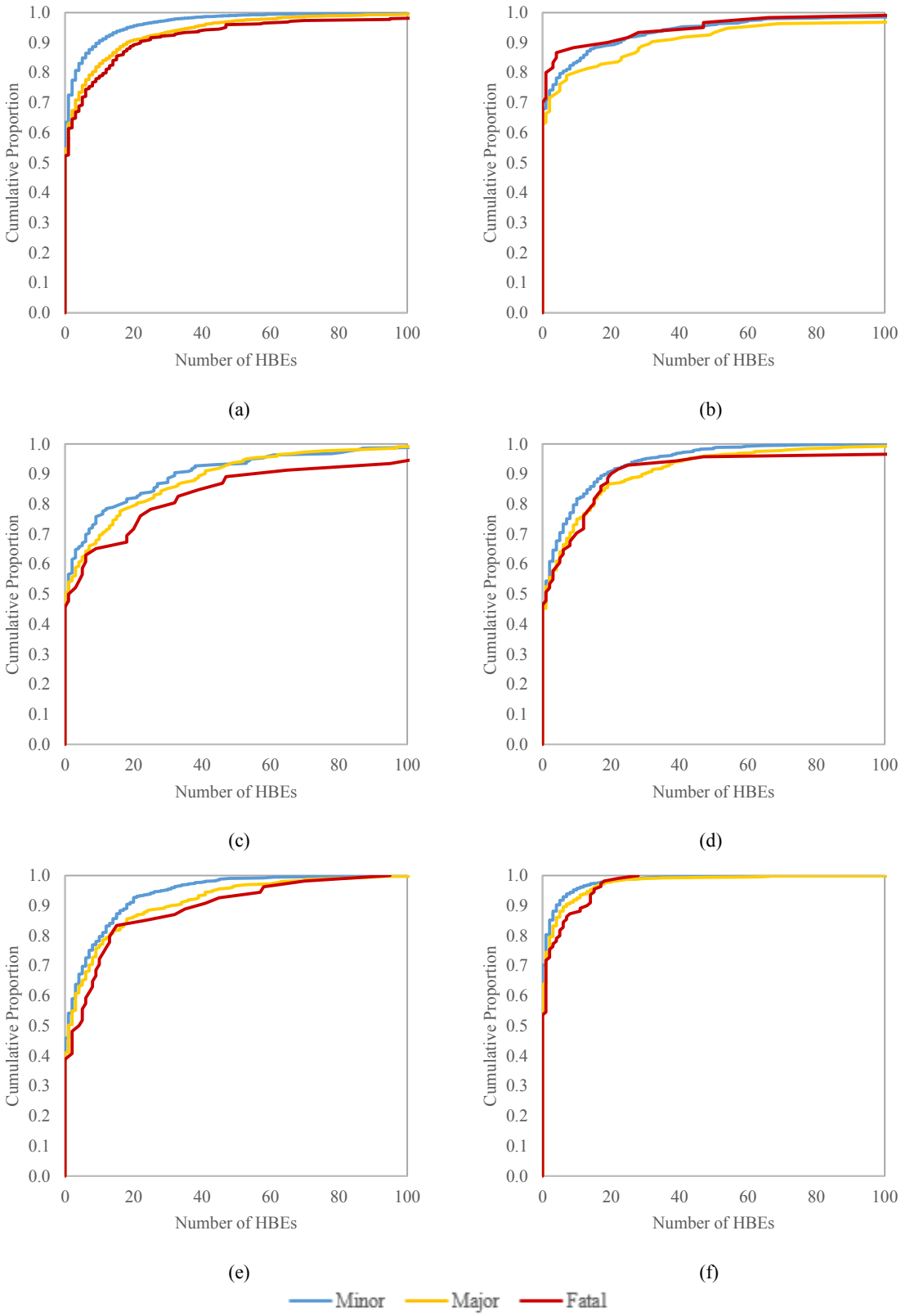


FIGURE B-7 Cumulative distributions for HBEs on links, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

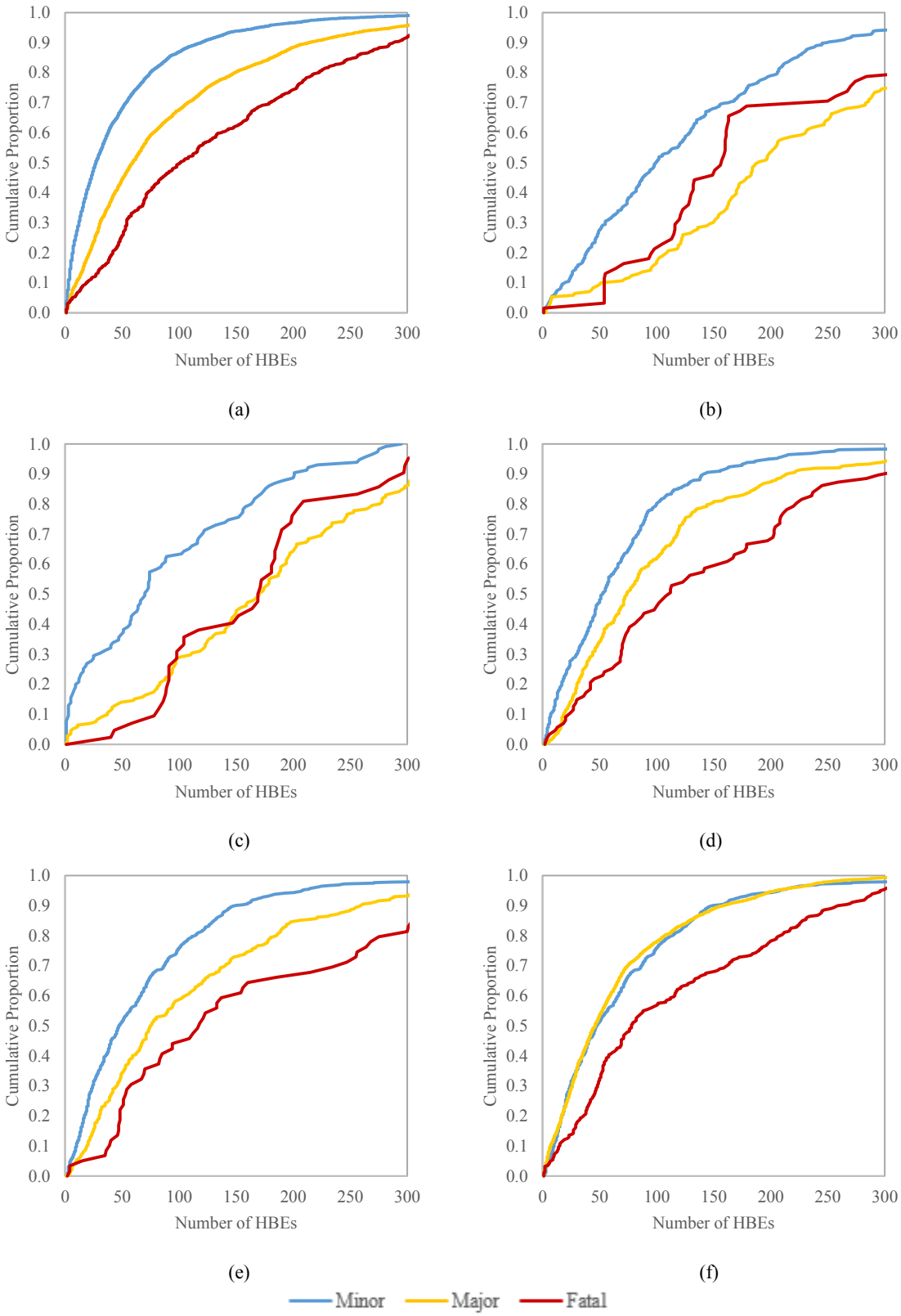


FIGURE B-8 Cumulative distributions for HBEs at intersections, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

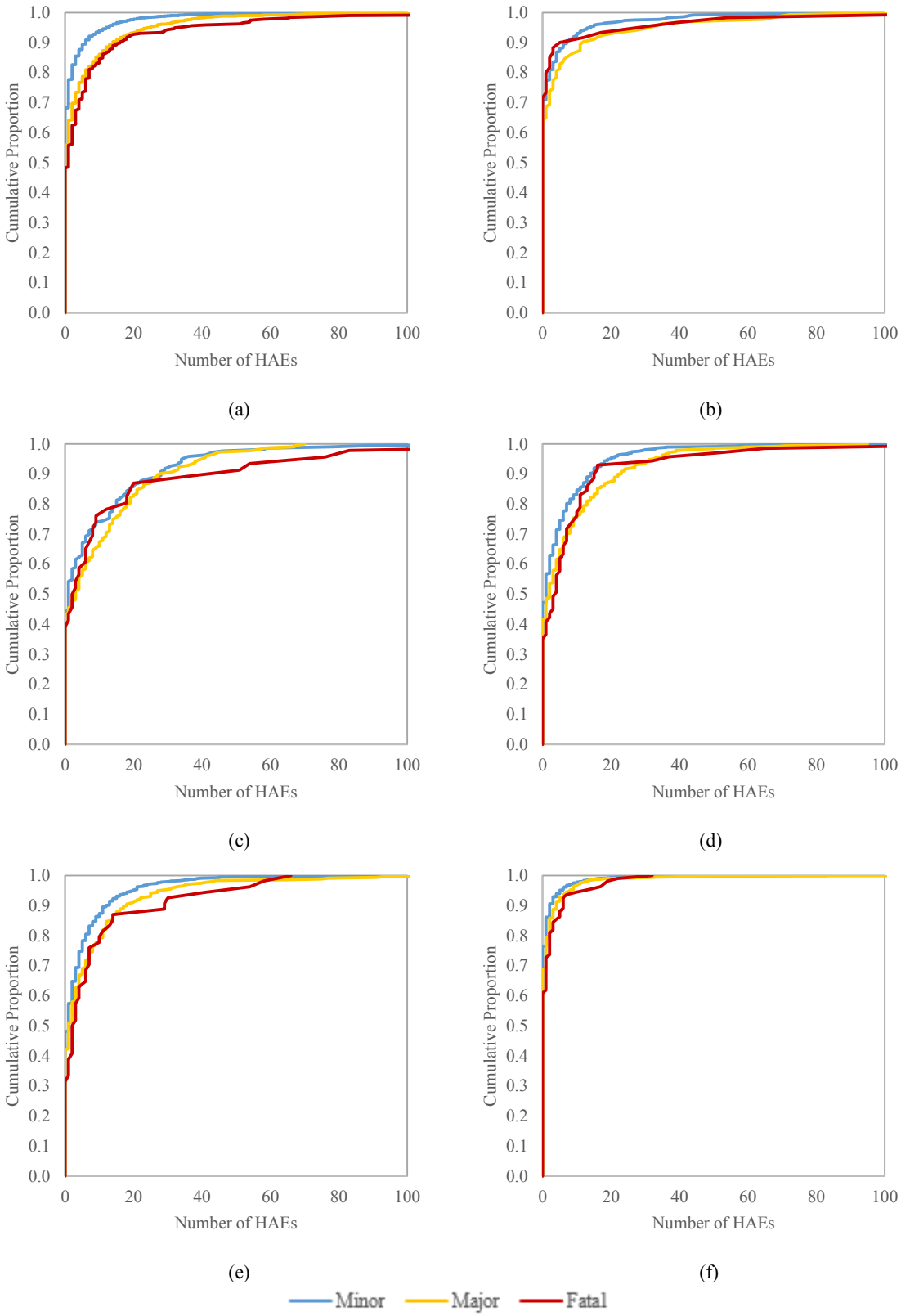


FIGURE B-9 Cumulative distributions for HAEs on links, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

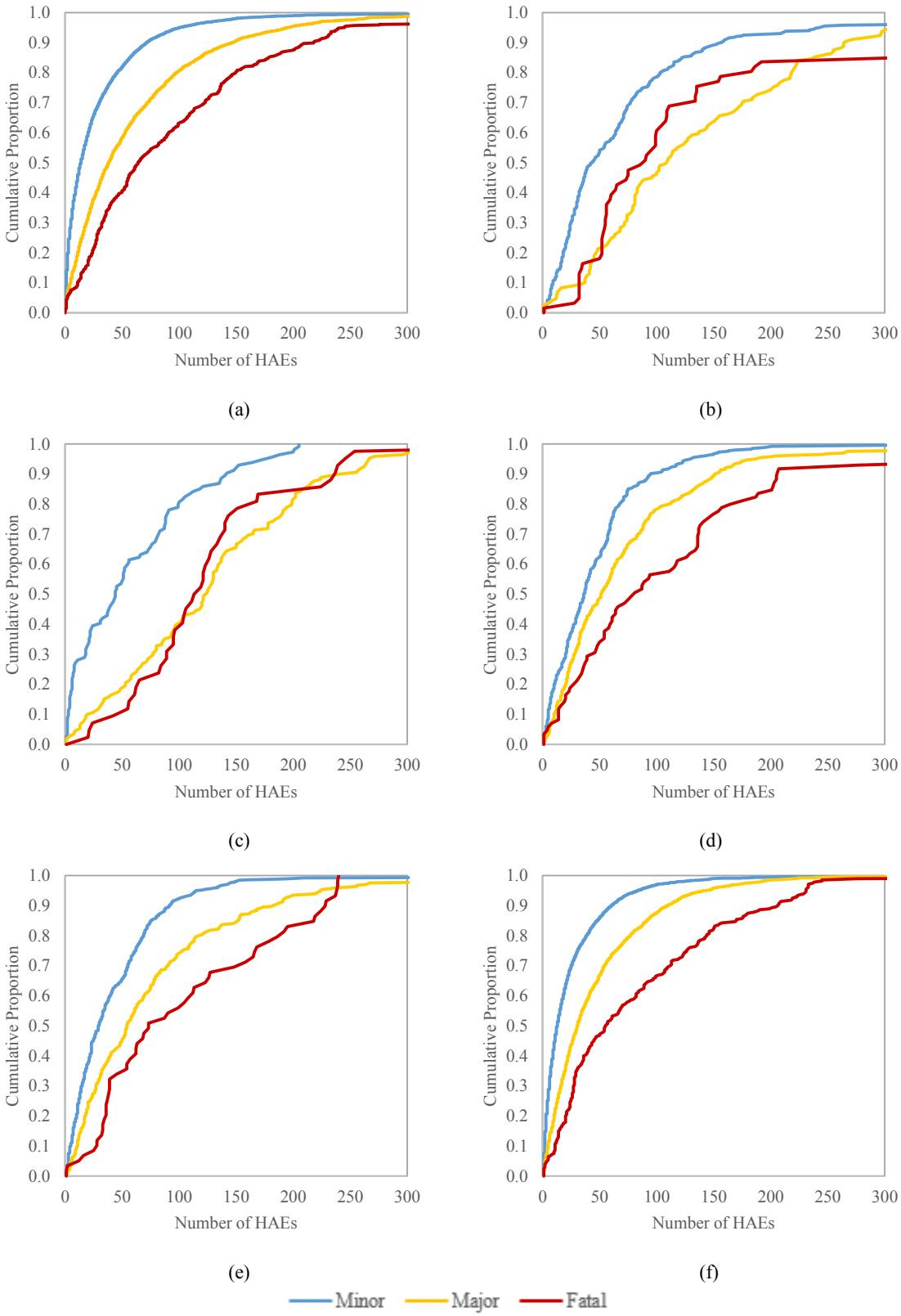


FIGURE B-10 Cumulative distributions for HAEs at intersections, all (a), motorways (b), primaries (c), secondaries (d), tertiaries (e), and residential streets (f)

APPENDIX C: IMPACT OF CRASH ASSIGNMENT ON MODELLING RESULTS

In Chapter 3, crashes were assignment to sites in the road network using a simple buffer technique. To generate crash counts at each link and intersection, the total number of crashes within a certain buffer around that site were counted. Although this did not significantly affect the results of frequency and severity testing presented in Chapter 3, the implications for the crash models presented in Chapter 4 were significantly higher. The model results attributed to using a simple buffer crash assignment method are demonstrated in Table C-1 and Figure C-1. Using this method, the over-counting of crashes (in overlapping buffers) creates artificial spatial autocorrelations which are observed as overfitting in the crash frequency models. This is especially apparent in the link level model, where correlation between the observed and fitted values was 0.99. Though the statistically significant variables, bolded in red, in both the link- and intersection-level models were consistent with expectations, a solution was clearly required to solve the issue of crash overcounting.

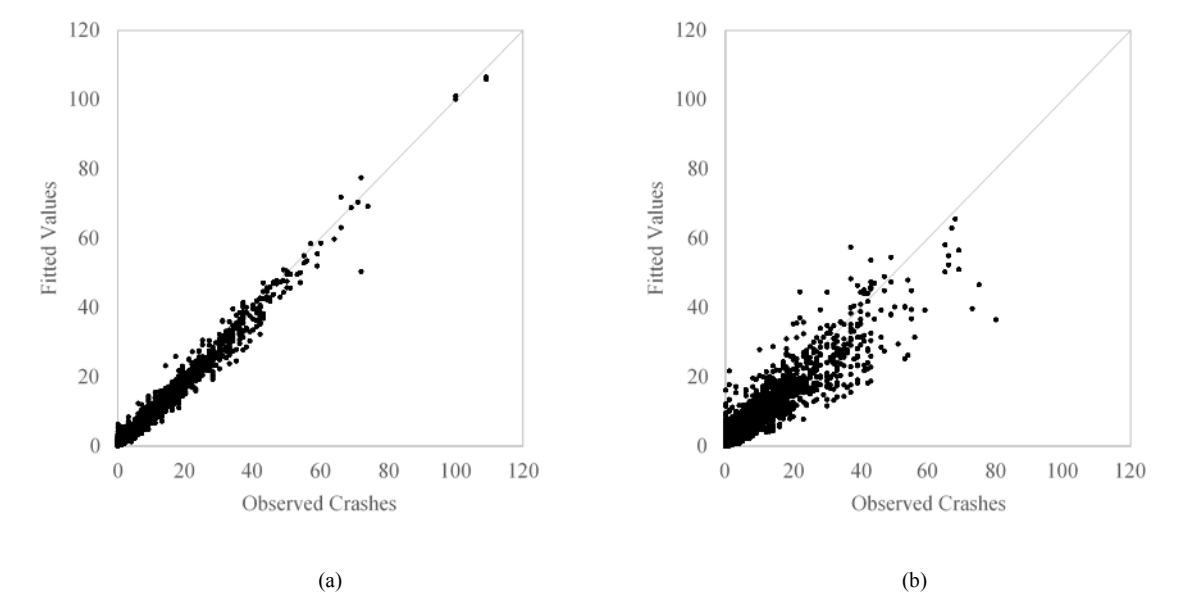
Two additional crash assignment methods were considered. First, crashes were assigned to the nearest site (link or intersection) in the network. The modelling results for this nearest neighbour assignment method are presented in Table C-2 and Figure C-2. Even under a new crash assignment method, the effects of the significant variables remain consistent with expectations based on results from Chapter 3. Clearly, this method solves the issues of overcounting, as the correlation between observed and fitted values significantly decreases. In fact, because each crash can only be assigned to one site, overcounting is eliminated. However, the greatest limitation of this method is that crashes may be easily mislocated, and there is no easy method determining the proportion of mislocated crashes. For example, crashes that occurred along a link can be matched to a parallel link if they are geolocated closer to that parallel link.

Instead, Chapter 4 presents the models based on crash assignment performed using non-overlapping buffers. 50 m buffers are used for both links and intersections, though link and intersection buffers do not overlap. This effectively eliminates 80 % or more of the overcounting at the link level. The results for these models are presented in Table C-3 and Figure C-3, and are discussed at length in Chapter 4. The main observation is that the means of the covariates are similar when compared to the results determined using the nearest neighbour crash assignment. Quantifying the impact of the crash assignment method on modelling results and developing a superior assignment method are critical areas for future work.

TABLE C-1 Spatial NB Model Results for Simple Buffer Crash Assignment

	Links				Intersections			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	**-1.64**	**0.15**	**-1.94**	**-1.35**	**-1.020**	**0.16**	**-1.333**	**-0.708**
ln(Trips)	**0.309**	**0.03**	**0.256**	**0.363**	**3.095**	**0.13**	**2.836**	**3.355**
HBEs/Trip	-0.025	0.13	-0.279	0.227	**1.043**	**0.07**	**0.907**	**1.184**
Congestion Index	**0.615**	**0.16**	**0.299**	**0.930**	**0.853**	**0.23**	**0.400**	**1.306**
CVS	**0.290**	**0.14**	**0.019**	**0.560**	0.243	0.21	-0.164	0.650
Average Speed	**-0.098**	**0.01**	**-0.114**	**-0.082**	**-0.072**	**0.01**	**-0.090**	**-0.055**
ln(Length)	**0.363**	**0.03**	**0.313**	**0.413**	*N/A*	*N/A*	*N/A*	*N/A*
Motorway	-0.107	0.10	-0.311	0.096	**-0.480**	**0.102**	**-0.681**	**-0.280**
Arterial/Collector	**0.583**	**0.06**	**0.472**	**0.695**	**0.292**	**0.05**	**0.197**	**0.388**
Number of cases	4623				4429			
DIC	19138.4				19992.2			
MSE	2.0				644.4			
CORR	0.99				0.30			

Note: Variables significant at 95 % confidence are bolded in red



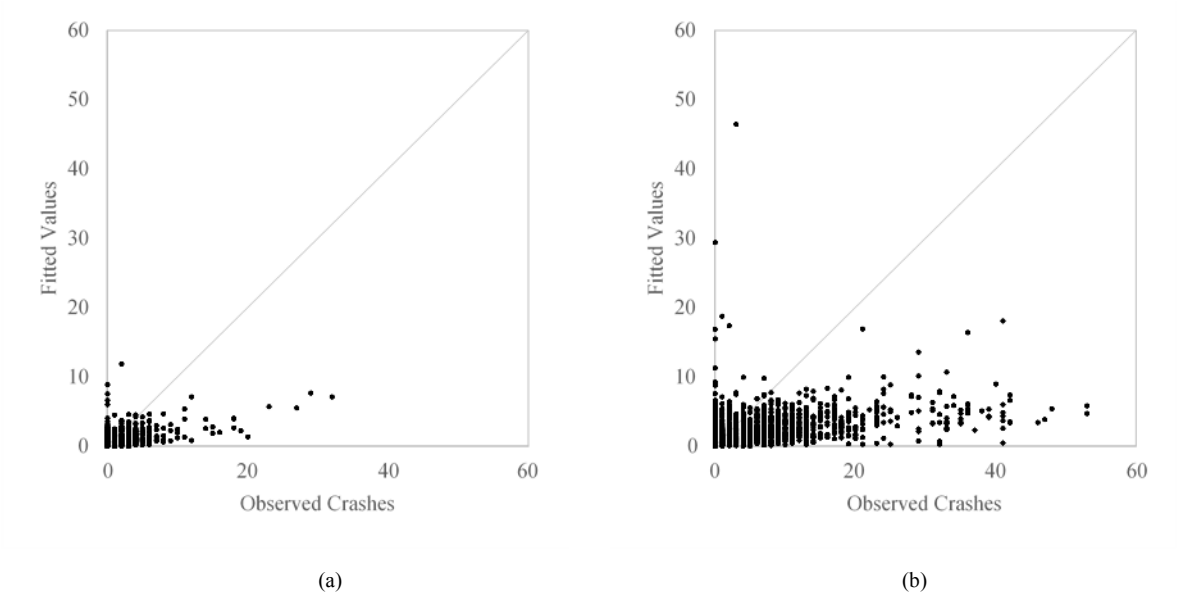
(a) (b)

FIGURE C-1 Fitted values versus observed crashes for links (a) and intersections (b) for simple buffer crash assignment

TABLE C-2 Spatial NB Model Results for Nearest Neighbour Crash Assignment

	Links				Intersections			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	**-7.37**	**0.37**	**-8.10**	**-6.66**	**-2.760**	**0.27**	**-3.288**	**-2.237**
ln(Trips)	**0.474**	**0.06**	**0.362**	**0.588**	**3.415**	**0.22**	**2.987**	**3.850**
HBEs/Trip	**-1.298**	**0.42**	**-2.122**	**-0.486**	**0.673**	**0.13**	**0.437**	**0.931**
Congestion Index	**0.821**	**0.38**	**0.073**	**1.567**	**1.099**	**0.38**	**0.356**	**1.847**
CVS	0.645	0.34	-0.017	1.305	**1.395**	**0.36**	**0.691**	**2.103**
Average Speed	**-0.094**	**0.02**	**-0.135**	**-0.054**	**-0.057**	**0.01**	**-0.086**	**-0.028**
ln(Length)	**0.925**	**0.06**	**0.800**	**1.053**	*N/A*	*N/A*	*N/A*	*N/A*
Motorway	**-2.418**	**0.34**	**-3.108**	**-1.772**	**-1.333**	**0.183**	**-1.694**	**-0.975**
Arterial/Collector	**0.931**	**0.11**	**0.709**	**1.155**	**0.297**	**0.09**	**0.122**	**0.472**
Number of cases	4623				4429			
DIC	6124.4				11100.2			
MSE	1.7				30.3			
CORR	0.56				0.47			

Note: Variables significant at 95 % confidence are bolded in red



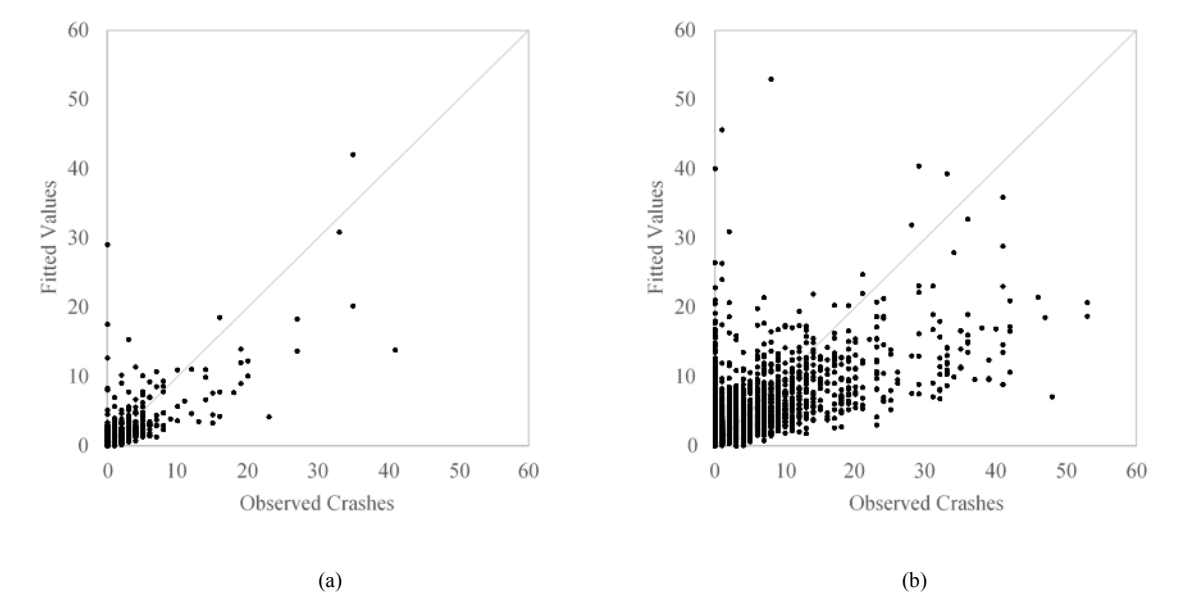
(a) (b)

FIGURE C-2 Fitted values versus observed crashes for links (a) and intersections (b) for nearest neighbour crash assignment

TABLE C-3 Spatial NB Model Results for Non-overlapping Buffer Crash Assignment

	Links				Intersections			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	**-15.83**	**0.87**	**-17.60**	**-14.22**	**-2.423**	**0.20**	**-2.811**	**-2.039**
ln(Trips)	**0.495**	**0.09**	**0.320**	**0.672**	**3.826**	**0.17**	**3.502**	**4.153**
HBEs/Trip	-0.438	0.65	-1.735	0.806	**1.131**	**0.10**	**0.940**	**1.328**
Congestion Index	**1.436**	**0.56**	**0.339**	**2.521**	**0.892**	**0.28**	**0.346**	**1.440**
CVS	0.972	0.51	-0.044	1.978	**0.905**	**0.26**	**0.402**	**1.408**
Average Speed	-0.046	0.03	-0.106	0.014	**-0.064**	**0.01**	**-0.085**	**-0.043**
ln(Length)	**2.110**	**0.12**	**1.881**	**2.354**	*N/A*	*N/A*	*N/A*	*N/A*
Motorway	**-3.037**	**0.47**	**-3.997**	**-2.135**	**-0.670**	**0.13**	**-0.921**	**-0.420**
Arterial/Collector	**0.848**	**0.19**	**0.485**	**1.220**	**0.449**	**0.06**	**0.323**	**0.576**
Number of cases	4623				4429			
DIC	3756.5				15920.7			
MSE	2.5				19.3			
CORR	0.63				0.67			

Note: Variables significant at 95 % confidence are bolded in red



(a) (b)

FIGURE C-3 Fitted values versus observed crashes for links (a) and intersections (b) for non-overlapping buffer crash assignment