August 2018

Advanced Quantitative Methods for Imminent Detection of Crash Prone Conditions and Safety Evaluation

Zhi Chen
University of Wisconsin-Milwaukee

ADVANCED QUANTITATIVE METHODS FOR IMMINENT DETECTION OF CRASH PRONE CONDITIONS AND SAFETY EVALUATION

by

Zhi Chen

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

August 2018

ABSTRACT

ADVANCED QUANTITATIVE METHODS FOR IMMINENT DETECTION OF CRASH PRONE CONDITIONS AND SAFETY EVALUATION

by

Zhi Chen

The University of Wisconsin-Milwaukee, 2018
Under the Supervision of Professor Xiao Qin

Crashes can be accurately predicted through reliable data sources and rigorous statistical models; and prevented through data-driven, evidence-based traffic control strategies. Both predictive analysis and analysis to estimate the causal effect of traffic variables of real-time crashes are instrumental to crash prediction and a better understanding of the mechanism of crash occurrence. However, the research on the second analysis type is very limited for real-time crash prediction; and the conventional predictive analysis using inductive loop detector data has accuracy issues related to inconsistently and distantly spaced loop detectors. The effectiveness of traffic control strategies for improving safety performance cannot be measured and compared without an appropriate traffic simulation application. This dissertation is an attempt to address these research gaps.

First, it conducts the propensity score based analysis to assess the causal effect of speed variation on crash occurrence using the crash data and ILD data. As a casual analysis method, the propensity score based model is applied to generate samples with similar covariate distributions in both high- and low-speed variation groups of all cases. Under this setting, the confounding effects are removed and the causal effect of speed variation can be obtained.

Second, it conducts a predictive analysis on lane-change related crashes using lane-specific traffic data collected from three ILD stations near a crash location. The real-time traffic data for the two lanes – the vehicle's lane (subject lane) and the lane to which that a vehicle intends to change (target lane) – are more closely related with lane-change related crashes, as opposed to congregated traffic data for all lanes. It is found that lane-specific variables are appropriate to study the lane-change frequency and the resulting lane-change related crashes.

Third, it conducts a predictive analysis on real-time crashes using simulated traffic data. The purpose of using simulated traffic data rather than real data is to mitigate the temporal and spatial issues of detector data. The cell transmission model (CTM), a macroscopic simulation model, is employed to instrument the corridor with a uniform and close layout of virtual detector stations that measure traffic data when physical stations are not available. Traffic flow characteristics at the crash site are simulated by CTM 0-5 minutes prior to a crash. It shows that the simulated traffic data can improve the prediction performance by accounting for the spatial-tempo issue of ILD data.

Fourth, it presents a novel approach to modeling freeway crashes using lane-specific simulated traffic data. The new model can not only account for the spatial-tempo issues of detector data but also account for heterogeneous traffic conditions across lanes using a lane-specific cell transmission model (LSCTM). The LSCTM illustrates both discretionary lane-changing (DLC) and mandatory lane-changing (MLC) activities. This new approach presents a viable alternative for utilizing traffic simulation models for safety analysis and evaluation.

Last, it develops a crash prediction and prevention application (CPPA) based on simulated traffic data to detect crash-prone conditions and to help select the desirable traffic control strategies for crash prevention. The proposed application is tested in a case study with

VSL strategies, and results show that the proposed crash prediction and prevention method could effectively detect crash-prone conditions and evaluate the safety and mobility impacts of various VSL alternatives before their deployment. In the future, the application will be more user-friendly and can provide both online traffic operations support as well as offline evaluation of various traffic control operations and methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Writing dissertation has been a period of intense learning for me, not only in the academic research, but also on a personal level. Without this valuable experience, I will never know how far I can push myself.

I would like to express the deepest appreciation to my advisor, Dr. Xiao Qin, who has provided me with unconditional and countless help and support during the entire graduate study. His inner passion in research and teaching infected me a lot. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank my committee members, Dr. Yue Liu, Dr. Yin Wang, Dr. Vytaras Brazauskas, Dr. Phoenix Do, and Dr. Jun Zhang. Thank you so much for being my committee members, reading my proposal and dissertation draft, and giving me the insightful suggestions and comments.

I am so luck to have the continuing and strong supports from my family, my girlfriend (Yanyan Wang), and my friends. When I was upset and stressed, they always encouraged me and cheered me up.

CHAPTER 1 INTRODUCTION

1.1 Background

Tremendous efforts have been devoted to improving traffic safety by designing safer roads, better managing traffic, and advancing vehicular technologies. These efforts have led to significant decrease in crashes in the last three decades. However, from 2014 to 2015, the number of crashes increased by 3.8% (NHTSA, 2016). The increase in crash occurrence is partially due to the lack of variations in safety treatment strategies. Traditional safety improvements such as roadway design improvements through 3R (resurfacing, restoration, and rehabilitation) projects are effective, but are reactive and restrictive. They are implemented after crashes have occurred, and only at selected locations where crashes are abnormally high. Furthermore, a physical safety improvement is difficult to be alter after it is completed, which does not respond timely to the varying vehicle performance, traffic conditions, and driver behaviors.

Predicting crashes is a common practice to support safety improvement decision-making such as hot spot identification, safety performance prediction, and the cost-benefit analysis. Crash prediction models (CPMs) are the statistical regression model for the crash frequency within a period of time (e.g., one year) developed from predictors including roadway geometries and traffic data averaged or aggregated over the same duration. The culmination of the development of CPMs is the release of the Highway Safety Manual (HSM) in 2010 by American Association of State Highway and Transportation Officials (AASHTO) where safety performance functions and crash modification factors/functions have been developed for all types of highway facilities and crash types. However, the safety studies based on aggregated crashes, or crash count, are incapable of identifying the crash-prone situations for individual

crashes, which are considered by researchers as more natural and direct measures of safety conditions of the traveling population.

In light of the new and emerging technologies such as advanced transportation information systems (ATIS), connected and autonomous vehicles (CAV), rich and large amount of information has been produced to shed the light on the development of more dynamic and proactive crash prevention methodologies and techniques. The increasingly available advanced data collection devices began to make a gradual shift in modern safety research with a different focus on understanding the unique circumstances of individual crashes. For instance, sensor instrumented vehicles and the onboard devices have been used to collect naturalistic driving data such as driver actions and vehicle kinematic data. These highly specified, granular data sources have allowed safety research to reveal crash-prone driving conditions and driver behavior pertaining to individual crashes. Risk factors contributing to crashes can now be identified, and their relationships to crash occurrence can be unraveled in detail.

At the crash event level, the prevailing traffic circumstances prior to and under which a crash takes place are believed to be one of the major contributors. A driver must constantly respond to changes in traffic due to the environment, adjacent vehicles, speed limit, highway curve, and pavement conditions. Increasing driver anxiety and traffic congestion leave little room for mistakes. It is best to use prevailing traffic circumstances to identify the causal factors pertaining to crashes and better understand the crash mechanism; as a result, targeted countermeasures or strategies can be proposed and implemented to effectively prevent crashes. Countermeasures can be both adaptive and proactive. If the traffic is continuously monitored, any traffic anomalies pertaining to crash hazards can be detected at an early stage. Drivers can be informed and/or appropriate traffic control strategies can be applied to mitigate crash risk.

Therefore, it is imperative to identify patterns and trends of traffic conditions associated with a crash before the crash happens.

The wide deployment of ATIS has made the collection, storage, and processing of real-time traffic data readily available, meaning researchers can now gather real-time information pertaining to crash occurrence. Among all types of traffic sensors, inductive loop detectors (ILD) have been a popular data source for real-time crash prediction. Figure 1-1 illustrates a layout of two ILD stations. Each station consists of two ILDs, one in each lane. A loop detector vehicle detection device is embedded under pavement. The wired loop is activated by the metal part of a passing vehicle, and the activation time is time-stamped and recorded. After a short period of time (e.g., 30 s), traffic volume can be obtained as the number of activations, and traffic occupancy can be measured as the proportion of the detector's on-time in the period. There are two types of ILDs, single ILDs and dual ILDs. Single ILDs record only traffic volume and occupancy, while dual ILDs record mean traveling speed and vehicle type in addition to volume and occupancy.

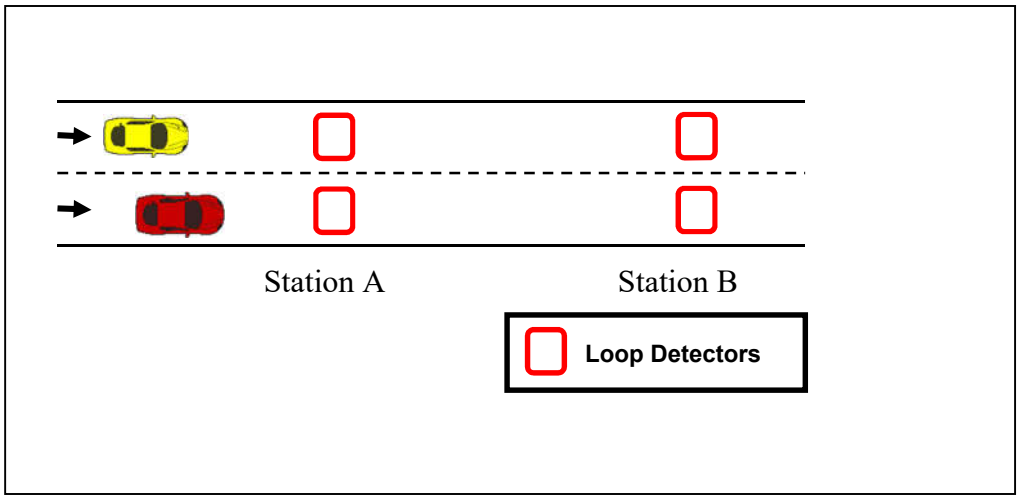


Figure 1-1 Illustration of a loop detector layout.

Figure 1-2 shows a sample of 1-min ILD data in Wisconsin. Detector ID is a unique ID associated with each lane at each ILD station. Date and Time record when the traffic data is recorded. Three traffic measurements are collected, Volume, Speed, and Occupancy. For example, the first record shows that between 9:00 and 9:01 on 4/16/2014, 20 vehicles passed the detector with an average speed of 67.17 MPH and a 5.67% occupancy.

Region	Detector ID	Date	Time	Volume (VPL)	Speed (MPH)	Occupancy (%)
SE	5513	4/16/2014	9:00	20	67.17	5.67
SE	5513	4/16/2014	9:01	20	62.85	7.17
SE	5513	4/16/2014	9:02	14	65	4.33
SE	5513	4/16/2014	9:03	14	68.9	3.83
SE	5513	4/16/2014	9:04	15	67.67	5.83
SE	5513	4/16/2014	9:05	13	64.2	4.17

Figure 1-2 Sample ILD data.

Figure 1-3 shows sample crash records in Wisconsin. Each crash has a unique document number in the DOCTNMBR field. The date and time are shown in ACCDDATE and ACCDTIME fields, respectively. The highway one crash happened on is in ONHWY field, and the direction is in ONHWYDIR field. Each crash also has accurate geo-location including both the latitude and longitude in WISLR_LATDECDG and WISLR_LONDECDG fields, respectively. The weather condition information is in WTHRCOND field. For example, the first record represents a crash that happens at 20:40 on 1/1/2012 on I94 EB. Its geo-location is (43.0289308, -88.1420277) and it was snowing when the crash happened.

DOCTNMBR	ACCDDATE	ACCDTIME	REGION	COUNTY	CNTYCODE	MUNICIPALITY	ONHWY	ONHWYDIR	WTHRCOND	WISLR_LATDECDG	WISLR_LONDECDG
C311WBS	1/1/2012	20:40	SE	WAUKESHA	67	BROOKFIELD	94	E	SNOW	43.0289308	-88.1420277
DHLT79F	1/7/2012	13:12	SE	WAUKESHA	67	BROOKFIELD	94	E	CLR	43.0289302	-88.1416941
DHLT799	1/10/2012	17:31	SE	WAUKESHA	67	BROOKFIELD	94	E	CLDY	43.0244371	-88.1063855
C2ZKG06	1/17/2012	9:55	SE	WAUKESHA	67	BROOKFIELD	94	E	SNOW	43.0244062	-88.1010924
C2ZKG07	1/18/2012	8:20	SE	WAUKESHA	67	BROOKFIELD	94	E	CLR	43.0247588	-88.1132117
C2ZL670	1/20/2012	12:20	SE	WAUKESHA	67	BROOKFIELD	94	E	SNOW	43.0244062	-88.1010924
C3030L0	1/24/2012	17:55	SE	WAUKESHA	67	BROOKFIELD	94	E	CLR	43.0289276	-88.14036
C2ZVBQR	1/29/2012	7:23	SE	WAUKESHA	67	BROOKFIELD	94	E	SNOW	43.0259169	-88.087746
DHSWFX9	2/7/2012	10:55	SE	WAUKESHA	67	BROOKFIELD	94	E	CLDY	43.0289219	-88.1373583

Figure 1-3 Sample crash records.

Real-time CPMs (RTCPM) can be developed using real-time traffic data collected from ILDs near the prospective crashes to identify crash-prone conditions. The rapid growth of this initiative prompts the genesis of a new research direction in real-time crash prediction and road safety surrogate methods. Previous research has investigated a variety of crash scenarios, including rear-end crashes (Pande & Abdel-Aty, 2006b), lane-change crashes (Chen, Qin, & Shaon, 2017; Lee, Abdel-Aty, & Hsia, 2006; Pande & Abdel-Aty, 2006a), crashes in different speed regimes (Mohamed Abdel-Aty, Uddin, & Pande, 2005; Pande & Abdel-Aty, 2006b), and visibility-related crashes (M. A. Abdel-Aty, Hassan, Ahmed, & Al-Ghamdi, 2012). Efforts have been continued to improve the prediction performance of RTCPMs through applying rigorous study designs (M. Abdel-Aty & Pande, 2005; Mohamed Abdel-Aty & Pemmanaboina, 2006; Mohamed Abdel-Aty, Uddin, Pande, Abdalla, & Hsia, 2004; Xu, Liu, Wang, & Li, 2014; Zheng, Ahn, & Monsere, 2010) and data quality (Mohamed Abdel-Aty et al., 2004; Lee, Hellinga, & Saccomanno, 2003; Zheng et al., 2010). Many attempts have been made to utilize sophisticated modeling techniques (Hossain & Muromachi, 2012; Pande & Abdel-Aty, 2006c; Jie Sun & Sun, 2015; Jian Sun, Sun, & Chen, 2014; Xu, Wang, & Liu, 2013).

1.2 Problem Statements

All previous studies on RTCPMs carried out predictive analysis that aims to predict crash probability but lacks the the focus on assessing the causal effect of individual traffic factors. According to a predictive analysis, traffic factors could appear significantly related to crash occurrence, but the relationship can be spurious due to the confounding factors that are related to both crash probability and the traffic variables of interest.

Predictive analysis in previous studies have temporal issues that may undermine the validity of their findings. Temporal proximity says that the traffic conditions a vehicle experiences immediately prior to or at the time of a crash are more relevant than the traffic conditions happening earlier or later. This phenomenon has been supported when comparing prediction accuracy from ILD data based on different crash lead times (Mohamed Abdel-Aty et al., 2004). However, many studies did not consider the traffic conditions that occur right before a crash (e.g., 0-5-min period) because of the assumption that preventative actions may take extra time in a real-time crash identification, notification, and prevention system. Most studies use traffic data from earlier time periods (e.g., the 5-10-min period before a crash) (Mohamed Abdel-Aty et al., 2004; Hossain & Muromachi, 2012; Pande & Abdel-Aty, 2006c; Jie Sun & Sun, 2015). Utilizing data in this manner ignores the fact that a crash can be abrupt and caused by traffic conditions occurring right before or during the crash, which can only be reflected with a closer temporal proximity (e.g., 0-5-min). Even if the crash is not a sudden event, crash-prone situations can intensify as approaching the crash occurrence time. Real-time traffic conditions in a closer temporal proximity may be more effective in distinguishing true crash-prone situations from false crash-prone situations.

Next, the spacing between ILD stations varies substantially from site to site. Large spacing leads to limited quality control for crashes that occur between stations and casts doubts regarding the consistency and transferability of findings in different studies. Given the discrepancies in the spatio-temporal domain, RTCPMs developed with traffic data collected directly from ILD stations may be inadequate in unraveling the intrinsic relationship between crash risk and traffic conditions. Such data issues would undermine the prediction power of developed models and should therefore be addressed.

Furthermore, most studies have focused on the rear-end crashes on the freeway due to the prevalence of ATIS on the freeway compared to other roadway types. Research on sideswipe crashes is rather limited when compared with the amount of studies on rear-end collisions crashes (Li, Ahn, et al., 2014; Oh, Park, & Ritchie, 2006; Pande & Abdel-Aty, 2006b; Pande & Abdel‐Aty, 2008; Qu, Wang, Wang, Liu, & Noyce, 2012).

Even when a reliable crash prediction model is available, the issue of selecting effective preventative countermeasures remains unsolved. Compared to the large body of real-time crash prediction studies, the research on evaluating the safety impacts of traffic control strategies using real-time traffic studies is limited to a few brief reports (Mohamed Abdel-Aty, Cunningham, Gayah, & Hsia, 2008; Mohamed Abdel-Aty, Pande, Lee, Gayah, & Santos, 2007; Li, Li, Liu, Wang, & Xu, 2014; Li, Liu, Wang, & Xu, 2014; Li, Liu, Xu, & Wang, 2016). A performance assessment tool is indispensable to evaluate the effectiveness of intervening strategies and promote the research findings from well-developed RTCPMs. A crash prediction and prevention application (CPPA) that combines both the RTCPM and the performance assessment tool is desirable as it can help detect crash-prone traffic conditions, distribute crash warnings, and evaluate traffic control countermeasures before their deployment.

In summary, the issues of current real-time crash prediction and prevention studies include:

1. Previous studies focused on predictive analysis and no rigorous analysis in estimating the causal effect of traffic variables has been performed, which would render biased estimates of traffic variables due to the existence of confounding factors.
2. Studies on lane-changing related crashes are limited compared to plentiful studies on rear-end crashes or total crashes.
3. Studies that used ILD traffic data overlooked the spatial and temporal issues associated with the ILD data, while those issues would compromise the prediction performance of resultant prediction models and undermine the validity of consequent findings.
4. A systematic safety assessment tool that can effectively measure the impacts of traffic control strategies before their deployment does not exist.

1.3 Research Objectives

The objective of this dissertation is to assess causal effects of traffic factors, develop an advanced methodology to detect crash-prone conditions in real time, and evaluate the effectiveness of traffic control strategies to reduce crash risk. More specifically, this dissertation aims to:

1. Conduct analysis to estimate the causal effect of traffic variables with actual traffic data collected from ILD stations and evaluate causal effects of contributing traffic factors;
2. Conduct predictive analysis with observed lane-specific traffic data collected from ILD stations and identify crash-prone traffic patterns for lane-changing related crashes;

3. Conduct predictive analysis using simulated traffic data from the output of traditional cell transmission models (CTM) to bridge the spatial and temporal gaps introduced by the traffic data from ILD stations;
4. Conduct predictive analysis using simulated traffic data from a lane-specific CTM (LSCTM) specifically developed to simulate lane-specific traffic data.
5. Design a crash prediction and prevention application (CPPA) that combines both the RTCPM and the performance assessment tool to help detect crash-prone traffic conditions, distribute crash warnings, and evaluate traffic control countermeasures before their deployment.

1.4 Dissertation Organization

To achieve all research objectives, the remaining dissertation is organized into seven chapters and the organization chart is presented in Figure 1-4:

Chapter 2 provides a summary of existing real-time crash prediction studies and relevant CTM studies including recent model improvements and their applications in traffic management. The chapter reviews the detector data specification, study design, methodology, crash scenarios and associated risk factors in real-time crash prediction studies; and the CTM framework and simulated safety related traffic control strategies.

Chapter 3 presents the analysis to estimate causal effects of traffic variables using actual traffic data collected from ILD stations. The causal effects of speed variations are evaluated using the propensity score-based method. The propensity score-based method estimates the propensity score of each case and then generates a weighted sample based on it. In the weighted sample, variables have similar distributions across two speed variation groups. Then the causal effect of the treatments can be impartially estimated without the nuisance due to other variables.

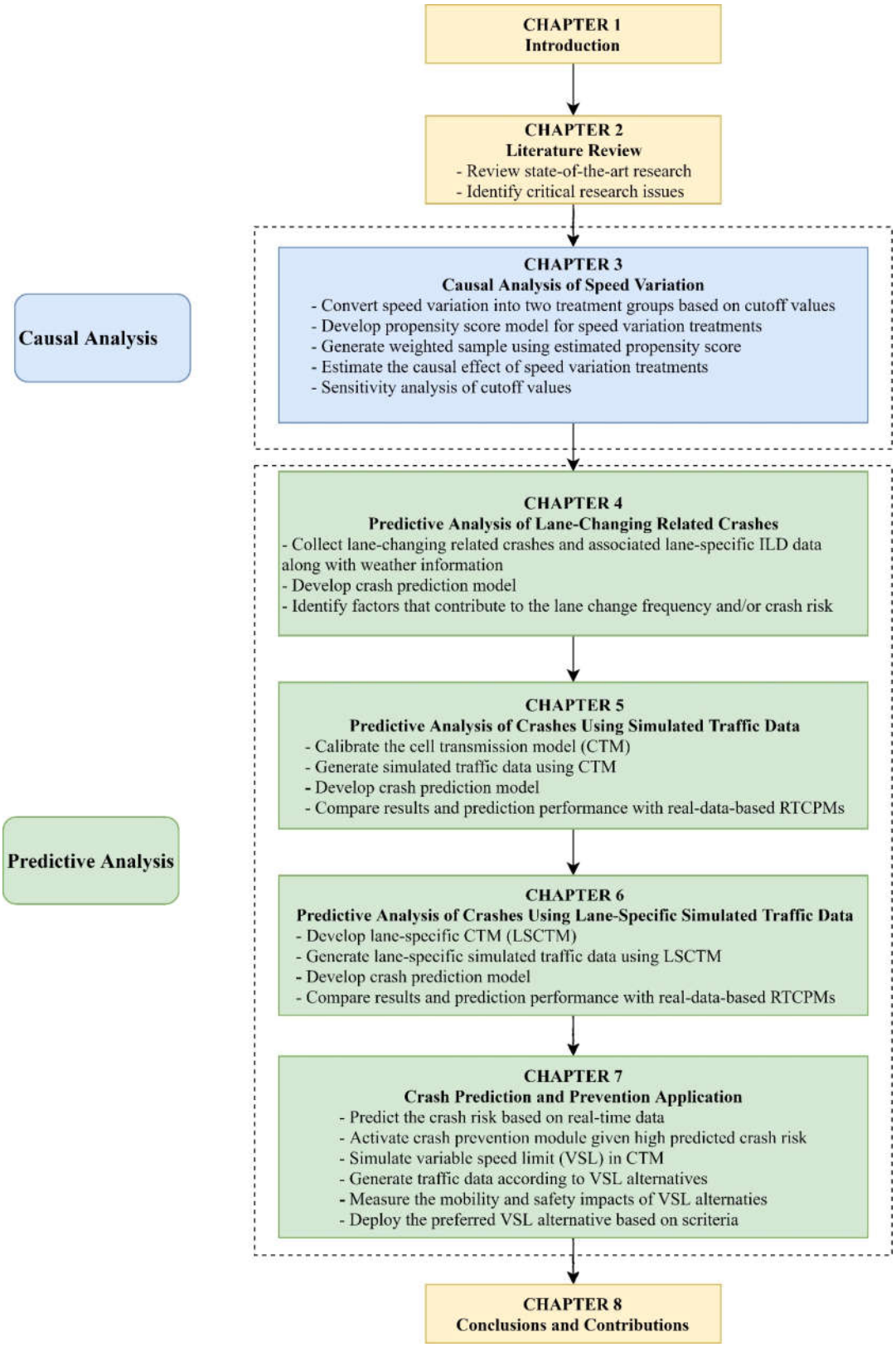


Figure 1-4 Dissertation organization.

Chapter 4 presents the predictive analysis on lane-changing related crashes using lane-specific traffic data collected from ILD stations. It is anticipated that the real-time traffic data for the two lanes – the vehicle's lane (subject lane) and the lane to which that a vehicle intends to change (target lane) – are more closely related with lane-change related crashes, as opposed to congregated traffic data for all lanes. Factors related to the lane-changing frequency and to the crash risk are investigated. The impact of weather conditions on the crash probability is explored.

Chapter 5 presents the predictive analysis on crashes based on simulated traffic from macroscopic traffic simulation CTM to account for the spatial and temporal issues related to traffic data from ILD stations. CTM is employed to instrument the corridor with a uniform and close layout of virtual detector stations that measure traffic data when physical stations were not available. Traffic flow characteristics at the crash site are simulated by CTM 0-5 minutes prior to a crash. Then, crash prediction models are developed using the binary logistic regression with traffic flow characteristics of simulated traffic data. The model developed with simulated traffic data is compared with that developed with observed traffic data collected from physical stations to assess the performance of crash models with simulated traffic.

Chapter 6 proposes a lane-specific cell transmission model (LSCTM) to simulate lane-specific traffic data for crash modeling. A LSCTM is developed to account for heterogeneous traffic conditions across lanes. The LSCTM illustrates both discretionary lane-changing (DLC) and mandatory lane-changing (MLC) activities. A case study is performed to demonstrate the method for modeling freeway crashes.

Chapter 7 develops a crash prediction and prevention application (CPPA) that combines both the RTCPM and the performance assessment tool to help detect crash-prone traffic

conditions, distribute crash warnings, and evaluate traffic control countermeasures before their deployment. The proposed application is tested in a case study with variable speed limit (VSL) strategies for demonstration.

Chapter 8 provides the conclusions and contributions of this dissertation.

1.5 References

Abdel-Aty, M., Cunningham, R., Gayah, V., & Hsia, L. (2008). Dynamic Variable Speed Limit Strategies for Real-Time Crash Risk Reduction on Freeways. *Transportation Research Record: Journal of the Transportation Research Board, 2078*, 108-116. doi:10.3141/2078-15

Abdel-Aty, M., & Pande, A. (2005). Identifying crash propensity using specific traffic speed conditions. *Journal of safety research, 36*(1), 97-108. doi:10.1016/j.jsr.2004.11.002

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., & Santos, C. D. (2007). Crash Risk Assessment Using Intelligent Transportation Systems Data and Real-Time Intervention Strategies to Improve Safety on Freeways. *Journal of Intelligent Transportation Systems, 11*(3), 107-120. doi:10.1080/15472450701410395

Abdel-Aty, M., & Pemmanaboina, R. (2006). Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. *IEEE Transactions on Intelligent Transportation Systems, 7*(2), 167-174. doi:10.1109/TITS.2006.874710

Abdel-Aty, M., Uddin, N., & Pande, A. (2005). Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1908), 51-58.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*(1897), 88-95.

Abdel-Aty, M. A., Hassan, H. M., Ahmed, M., & Al-Ghamdi, A. S. (2012). Real-time prediction of visibility related crashes. *Transportation Research Part C: Emerging Technologies, 24*, 288-298.

Chen, Z., Qin, X., & Shaon, M. R. R. (2017). Modeling Lane-change Related Crashes with Lane-specific Real-time Traffic and Weather Data. *Journal of Intelligent Transportation Systems*(just-accepted).

Hossain, M., & Muromachi, Y. (2012). A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention, 45*, 373-381. doi:10.1016/j.aap.2011.08.004

Lee, C., Abdel-Aty, M., & Hsia, L. (2006). Potential real-time indicators of sideswipe crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 41-49.

Lee, C., Hellinga, B., & Saccomanno, F. (2003). Real-time crash prediction model for application to crash prevention in freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board*(1840), 67-77.

Li, Z., Ahn, S., Chung, K., Ragland, D. R., Wang, W., & Yu, J. W. (2014). Surrogate safety measure for evaluating rear-end collision risk related to kinematic waves near freeway recurrent bottlenecks. *Accident Analysis & Prevention, 64*, 52-61.

Li, Z., Li, Y., Liu, P., Wang, W., & Xu, C. (2014). Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers. *Accident Analysis & Prevention, 72*, 134-145. doi:10.1016/j.aap.2014.06.018

Li, Z., Liu, P., Wang, W., & Xu, C. (2014). Development of a Control Strategy of Variable Speed Limits to Reduce Rear-End Collision Risks Near Freeway Recurrent Bottlenecks. *IEEE Transactions on Intelligent Transportation Systems, 15*(2), 866-877. doi:10.1109/TITS.2013.2293199

Li, Z., Liu, P., Xu, C., & Wang, W. (2016). Optimal Mainline Variable Speed Limit Control to Improve Safety on Large-Scale Freeway Segments: Optimal mainline variable speed limit. *Computer-Aided Civil and Infrastructure Engineering, 31*(5), 366-380. doi:10.1111/mice.12164

NHTSA. (2016). 2015 Motor Vehicle Crashes: Overview. *Traffic safety facts research note, 2016*, 1-9.

Oh, C., Park, S., & Ritchie, S. G. (2006). A method for identifying rear-end collision risks using inductive loop detectors. *Accident Analysis & Prevention, 38*(2), 295-301.

Pande, A., & Abdel-Aty, M. (2006a). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention, 38*(5), 936-948. doi:10.1016/j.aap.2006.03.004

Pande, A., & Abdel-Aty, M. (2006b). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 31-40.

Pande, A., & Abdel-Aty, M. (2006c). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board, 1953*(1), 31-40.

Pande, A., & Abdel‐Aty, M. (2008). A computing approach using probabilistic neural networks for instantaneous appraisal of rear‐end crash risk. *Computer‐Aided Civil and Infrastructure Engineering, 23*(7), 549-559.

Qu, X., Wang, W., Wang, W., Liu, P., & Noyce, D. A. (2012). *Real-time prediction of freeway rear-end crash potential by support vector machine.* Paper presented at the Transportation Research Board 91st Annual Meeting.

Sun, J., & Sun, J. (2015). A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies, 54*, 176-186. doi:10.1016/j.trc.2015.03.006

Sun, J., Sun, J., & Chen, P. (2014). Use of Support Vector Machine Models for Real-Time Prediction of Crash Risk on Urban Expressways. *Transportation Research Record: Journal of the Transportation Research Board, 2432*, 91-98. doi:10.3141/2432-11

Xu, C., Liu, P., Wang, W., & Li, Z. (2014). Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. *Transportation Research Part A: Policy and Practice, 69*, 58-70. doi:10.1016/j.tra.2014.08.011

Xu, C., Wang, W., & Liu, P. (2013). A genetic programming model for real-time crash prediction on freeways. *IEEE Transactions on Intelligent Transportation Systems, 14*(2), 574-586.

Zheng, Z., Ahn, S., & Monsere, C. M. (2010). Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention, 42*(2), 626-636. doi:10.1016/j.aap.2009.10.009

CHAPTER 2 LITERATURE REVIEW

This chapter presents a comprehensive review of real-time crash prediction studies and relevant CTM studies that provide information regarding recent CTM improvements and their applications in traffic management. The first section summarizes the characteristics, strengths, and deficiencies of the state-of-the-art research efforts that investigate the relationships between crash risk and real-time traffic along with operational factors. The second section explores the improvements of CTMs in terms of how they are applied and their ability to generate more reliable traffic simulation.

2.1 Overview of Real-time Crash Prediction

Oh et al. (Oh, Oh, Ritchie, & Chang, 2001) and Golob and Recker (Golob & Recker, 2001) were the first to analyze crash patterns based on real-time traffic flow. Since the early 2000s, substantial research has been devoted to incident detection and traffic management due to the emergence of ATIS; however, little research involved incident prevention. In response, Oh et al. (Oh et al., 2001) decided to measure accident likelihood using real-time traffic data from ILDs. Their study is based on the assumption that the disruptive traffic, represented by high temporal and spatial variation in traffic parameters, contributes to accidents. Two traffic conditions were defined in their study: normal condition (a 5-minute period 30 minutes before a crash) and disruptive condition (a 5-minute period right before a crash). The authors aggregated 10-second traffic flow, speed, and occupancy data into 5-minite intervals and derived the mean and standard deviation of these three factors as indicators. Speed variation was found to be the best indicator of a disruptive condition that contributes to crash occurrence. A real-time application that dynamically monitors the crash likelihood was proposed, and its performance showed the potential for identifying crash-prone conditions using real-time traffic data.

In traditional safety studies, the relationship between crash rates and highly aggregated traffic data (e.g., daily or hourly traffic counts) is a common subject. Golob and Recker (Golob & Recker, 2001) proposed research to solve two outstanding problems in such studies – argument averaging and function averaging. Argument averaging is the use of average traffic flow over a long period rather than the measure of traffic data just prior to an accident. Function averaging relates to the use of the same functions for all types of collisions under all conditions (e.g., weather and lighting conditions). Nonlinear canonical correlation analysis (NLCCA) was applied to investigate the relationship between three sets of variables. The first set included one variable defining the weather and lighting condition at the time of crash; the second set consisted of three accident characteristics: collision type, crash location, and crash severity; the third set comprised real-time traffic flow variables. Real-time traffic flow variables were obtained by first aggregating 30-s lane-specific volume and speed data collected from the ILD station nearest to each crash before the crash occurrence and then applying principal component analysis to the aggregated variables. When controlling for weather and lighting conditions, the authors found that collision type is the best-explained accident characteristic related to median speed and left- and interior-lane speed variation, and that crash severity is influenced more by volume than by speed.

Inspired by both aforementioned groundbreaking articles, a large number of studies have been conducted to develop RTCPMs that identify real-time crash-prone traffic patterns and quantify their effects in crash forecasting. A preponderance of these studies used traffic data collected from ILDs since this data has been proven a useful data source and the stations are widely available and accessible. Other popular data sources utilized in real-time safety studies include video surveillance and Automatic Vehicle Identification (AVI) sensors. Video

surveillance traces all vehicles passing the coverage area via video footage, and AVI sensors collect the passage time of vehicles with AVI tags passing consecutive AVI tag readers.

Video surveillance allows individual vehicle trajectories to be extracted and derived to generate disaggregated traffic characteristics such as speed and time headway. Such data can provide a more explicit view of how vehicles interact before the crash occurrence. ILD data-based studies relate crash likelihood to aggregated traffic characteristics (e.g., variation in speed), whereas video surveillance data research can connect crash propensity to the probability of a vehicle failing to make evasive movements in order to avoid a collision based on the vehicle kinematics and its surrounding vehicles. Although several studies based on surveillance videos have discovered meaningful findings (Chatterjee & Davis, 2016; Davis & Swenson, 2006; Hourdos, Garg, & Michalopoulos, 2008; Hourdos, Garg, Michalopoulos, & Davis, 2006), video surveillance has limitations that hinder its wide use. Due to the high cost of setting up video cameras, video surveillance covers only limited segments, making it difficult to collect a sufficient size of crashes for analysis. Moreover, video requires intensive labor to retrieve and process vehicle trajectories.

AVI sensors collect the travel time of a vehicle equipped with a tag for each AVI segment and then derive the average traveling speed. Traffic data collected by AVI sensors have been utilized in several real-time safety studies (M. Abdel-Aty, Pande, Lee, Gayah, & Santos, 2007; M. A. Abdel-Aty, Hassan, & Ahmed, 2012; Ahmed & Abdel-Aty, 2013; Al-Deek, Venkata, & Ravi Chandra, 2004; Hosmer Jr & Lemeshow, 2004; Shi & Abdel-Aty, 2015; Shi, Abdel-Aty, & Yu, 2016; Yu & Abdel-Aty, 2013). However, one critical issue of this data source is that the sensors can record only vehicles with AVI tags. Therefore, the flow rate and speed data are derived based on a sample of vehicles, and the sample may not be representative. In

addition, AVI sensors cannot collect the occupancy information which has been proven to be a critical variable in predicting crash occurrence in real-time safety studies (M. Abdel-Aty & Pande, 2006; M. Abdel-Aty, N. Uddin, A. Pande, F. Abdalla, & L. Hsia, 2004a; Xu, Liu, & Wang, 2016). Hence, RTCPMs developed with AVI data may be biased given unreliable flow and speed data and unavailable occupancy data.

Due to the limitations of data collected by video surveillance and AVI sensors, ILD data are selected as the data source in this dissertation, and only real-time safety studies using ILD data are reviewed in the following section. Various aspects of those studies are discussed, including data specification, study design, methodologies, crash type, and risk factors. The concept of crash/non-crash cases was adopted in almost all real-time crash studies. A crash case represents the traffic conditions prior to a crash, while a non-crash case represents normal traffic conditions.

2.1.1 Traffic Detector Data Specification

ILDs record three indexes: volume, occupancy, and speed within a short period of time (e.g., 10 s, 20 s, 30 s and 1 min). Most previous studies have aggregated raw data into a longer period (e.g., 2 min and 5 min) and calculated the mean and variation of the three indexes following the procedure in (Oh et al., 2001) and (Golob & Recker, 2001). Time duration and the lead time before a crash occurrence define the time interval within which the data is aggregated. For example, the time interval would be the 5-10-min interval before crash occurrence if the time duration is 5 min and the lead time is 5 min. Researchers aim to identify crash-prone patterns based on the data within time intervals prior to crashes.

As pointed out in a review paper on real-time crash studies by Roshandel et al. (Roshandel, Zheng, & Washington, 2015), different time durations for data aggregation would

significantly impact the study quality and modeling results; however, a guide for selecting appropriate time durations is lacking in most studies. Similarly, lead time would be of the same importance as time duration.

The time interval defined by the lead time and duration varies across studies based on whether they apply to crash detection or crash prevention. Crash detection aims to identify the most relevant and significant factors in maximizing the success of crash prediction. Crash prevention aims to identify crash-prone conditions and prevent crashes, and is more time-sensitive because it takes time to deploy the measures necessary to avoid negative consequences. Some crash detection studies used the time interval right before the crash occurred, or a lead time of 0 min (Lee, Hellinga, & Saccomanno, 2003b; Lee, Saccomanno, & Hellinga, 2002; Oh et al., 2001; Zheng, Ahn, & Monsere, 2010), as the traffic conditions in the period right before the crash occurrence is most likely to be associated with the crash occurrence. This reasoning is supported by a study that tested different time intervals (M. Abdel-Aty et al., 2004a). Lee et al. (Lee et al., 2003b) defined the optimal time duration as that which maximizes the difference in crash precursor values between crash and non-crash cases. The authors also found that 2, 3 and 5 min are the optimal time durations for three crash precursors - longitudinal variation of speed, average upstream and downstream speed difference, and average density. Zheng et al. (Zheng et al., 2010) investigated the impact of the traffic oscillation on the crash occurrence and selected 10 min as the optimal time duration because it was found to be the typical duration of traffic oscillation.

Crash prevention studies are intended to develop practical RTCPMs that can be applied in the real world. The lead times for these studies are not 0 min to allow enough time for taking prevention measures; therefore, both the duration and the lead time need to be determined. Only

a few studies have proposed approaches for choosing time intervals. Abdel-Aty (M. Abdel-Aty et al., 2004a) and Xu et al. (Xu, Liu, Wang, & Li, 2012) compared different 5-min slices (i.e., 0-5-min, 5-10-min, …, 25-30-min) and found that the 5-10-min slice is the most appropriate in terms of performance and practicality. Pande et al. (A. Pande, Abdel-Aty, Hsia, & Trb, 2005) compared a 3-min slice (i.e., 0-3-min, 3-6-min,…, 12-18-min) and a 5-min slice (i.e., 0-5-min, 5-10-min, …, 25-30-min), finding that data aggregated into 5-min slices shows a stronger association with crash occurrence than 3-min slices. Additionally, among 5-min slices, 5-10-min and 10-15-min are preferred as they provide superior modeling results and are more practical. In contrast, most of the other studies only arbitrarily selected time duration, most commonly using the 5-10-min interval prior to the crash occurrence.

Roshandel et al. pointed out that ILD data collected far from crash locations (Roshandel et al., 2015) may be limited because although ILD stations can be equally spaced (e.g., 0.5 mi apart in some studies) (M. Abdel-Aty & Pande, 2005; M. Abdel-Aty & Pemmanaboina, 2006; M. Abdel-Aty et al., 2004a; Anurag Pande & Abdel-Aty, 2006b), spacing can vary significantly within and across studies. For example, the spacing ranges from 0.2 to 1.3 mi with an average of 0.5 mi in (Xu, Liu, & Wang, 2016), from 0.15 to 1.68 mi with an average of 0.5 mi in (Xu, Tarko, Wang, & Liu, 2013) and from 0.34 to 2.37 mi with an average of about 1.06 mi in (Zheng et al., 2010). Studies have shown that the sensor location may affect the estimation of the traffic flow (Danczyk & Liu, 2011; Hong & Fukuda, 2012; Kwon, Petty, & Varaiya, 2007; H. X. Liu & Danczyk, 2009). Hong and Fukuda (Hong & Fukuda, 2012) studied the impacts of the ILD station count, spacing, and layout on the estimation accuracy of travel speed, finding that even with the same station count, one layout provided a balance of under- and over-estimation of speed across stations, while a different one reported over-estimated speed at most stations. The

findings also showed that distant sensors could lead to over-estimation of travel speed. In addition, Kwon et al. (Kwon et al., 2007) observed that the accuracy of measuring traffic congestion drops as the distance between ILD sensors increases. Liu and Danczyk (H. X. Liu & Danczyk, 2009) found that the sensor location had an impact on the accurate detection of bottlenecks, and proposed a method to locate sensors for bottleneck detection optimally. The findings suggest that station spacing and station layout may affect the estimation of traffic flow characteristics which are key input variables of RTCPMs.

Based on the above discussion, one can conclude that data for real-time safety studies involving crash detection and crash prevention are susceptible to both temporal and spatial issues. Temporal issues arise due to a lack of rigor in selecting the appropriate time intervals before the crash occurrence. Spatial issues arise due to different spacing between ILD stations within studies and varying layouts of ILD stations across studies.

2.1.2 Study Design of Real-time Crash Prediction

It takes both crash and non-crash events to develop RTCPMs and identify crash-prone patterns. A crash case involves the traffic conditions prior to a crash occurrence and is restricted by the crash. However, a non-crash case involves crash-free traffic conditions and could include any traffic conditions during crash-free days.

Two primary study designs – matched the case-control design and unmatched design – determine how non-crash cases are collected. The matched case-control design is an efficient means of studying rare diseases, and is widely applied in epidemiological studies (Niven, Berthiaume, Fick, & Laupland, 2012). Abdel-Aty (M. Abdel-Aty et al., 2004a) introduced this study design to real-time crash studies. The matched case-control design compares the level of risk factors in two similar groups, one that includes the outcome and one that does not (Cornfield

et al., 1959). In a matched case-control design, each crash is considered as a "case", and non-crash cases are selected as "controls" by matching confounding factors (i.e., location and time). Confounding factors are correlated with traffic conditions and contribute to the crash occurrence. One example of a matched case-control design involves a crash occurring at 11:01 a.m. on Jan 17st, 2012 that uses traffic measurements from 10:56 to 11:01 a.m. (0-5-min period before the crash time) on the same day from its immediately upstream and downstream ILD stations. Traffic measurements are then collected from the same stations during the same period on crash-free days in 2012 as controls. The matched case-control design is constructed to remove the noise of confounding factors and investigate the risk factors of interest. The matched case-control design is expected to increase the accuracy of variable estimates in RTCPMs by controlling the confounding bias. This study design can greatly reduce the required size of non-crash cases. Due to the efficiency of the matched case-control study design, most real-time crash studies adopt this design for data collection (M. Abdel-Aty & Pande, 2005; M. Abdel-Aty & Pemmanaboina, 2006; M. Abdel-Aty, Uddin, & Pande, 2005; M. Abdel-Aty et al., 2004a; A. Pande et al., 2005; Xu et al., 2012; Xu, Liu, Wang, & Li, 2014; Zheng et al., 2010).

In an unmatched study design, non-crash cases are randomly selected. The safety impacts of all factors, including risk factors (e.g., traffic flow variables) and confounding factors (e.g., geometric design) are estimated based on a large sample. Compared to the matched case-control design, the unmatched design does not require matching confounding factors, and it therefore requires less effort to identify non-crash cases. However, a sufficiently large sample size is required to ensure accurate estimation, especially when the variable number is high (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). This drawback may be the reason why only a few studies have employed the unmatched study design (Anurag Pande & Abdel-Aty,

2006a, 2006b; Xu, Liu, & Wang, 2016; Xu, Wang, & Liu, 2013). Unlike the matched case-control study design, the unmatched design allows for the estimation of both risk factors and confounding factors.

Although the matched case-control design seems superior, there has been no systematic comparison between these two until recently. Xu et al. (Xu, Liu, & Wang, 2016) proposed a measure of the model prediction performance called "predictability" which compared the predictability of the two designs. The authors found that given a predefined specificity (the proportion of crash cases that are correctly classified), the predictability of the RTCPM developed with unmatched data always outperformed that of the matched case-controlled data. This empirical finding is enlightening, but may be data dependent. Additional research is needed to evaluate the two study designs.

2.1.3 Methodology

Specific techniques are required to sort out the relationships between the relatively low number of crashes and the massive volume of real-time traffic data. In general, the approaches to real-time crash prediction can be categorized as either statistical regression models or data mining techniques. Statistical regression models can build clear connections between crash probability and traffic flow variables, which is vital to helping develop proactive safety approaches. Other than two early studies which used the log-linear model (Lee et al., 2003b; Lee et al., 2002), almost all later studies used logistic models. The two main types of logistic models are conditional logistic and regular logistic. The conditional logistic model can be applied only to the data collected using the matched case-control study, while the regular logistic model can be applied even when the data are randomly collected. The matched case-control study design controls the confounding factors of non-crash cases, but the conditional logistic model does not

provide estimates of those confounding factors and cannot predict the crash risk of a given traffic case. For a given traffic case, its matched non-crash cases are first selected and then used to calculate the predicted odds ratio of crash risk based on the conditional logistic modeling results. A threshold value of odds ratio can be established to classify crashes from non-crashes. A regular logistic model gives estimates of all parameters in the model and can be directly applied to predict the crash risk of a given case. Crashes can then be classified based on a pre-established threshold value. Since most real-time crash studies adopted the matched case-control design, the conditional logistic model is more widely applied (M. Abdel-Aty & Pande, 2005; M. Abdel-Aty & Pemmanaboina, 2006; M. Abdel-Aty et al., 2005; M. Abdel-Aty et al., 2004a; A. Pande et al., 2005; Xu et al., 2012, 2014; Zheng et al., 2010) than the regular logistic model (Anurag Pande & Abdel-Aty, 2006a; Xu, Liu, & Wang, 2016; Xu, Wang, et al., 2013).

Contrary to statistical regression models, data mining techniques do not identify explicit relationships between crash probability and traffic flow variables. Various data mining techniques such as the support vector machine (SVM) (Jian Sun, Sun, & Chen, 2014; Yu & Abdel-Aty, 2013), neural networks (NN) (Anurag Pande & Abdel-Aty, 2006a, 2006b), the genetic algorithm (Xu, Wang, et al., 2013) and the Bayesian network (Hossain & Muromachi, 2012; Jie Sun & Sun, 2015) have been applied in real-time safety studies. These data mining methods treat the crash prediction problem as a classification problem, and their aim is to achieve the optimal classification accuracy. The SVM constructs a high-dimensional space based on factors contributing to the crash outcome (e.g., traffic flow variables) and identifies the optimal hyperplane to separate crashes from non-crashes (Jian Sun et al., 2014; Yu & Abdel-Aty, 2013). The NN is comprised of multiple layers as shown in Figure 2-1, which presents a three-layer NN. The first layer represents the vector of contributing factors, and different weight

vectors are applied to these factors to get the subsequent layers until the output layer is computed to get the classification results (i.e., crash or non-crash) (Anurag Pande & Abdel-Aty, 2006a, 2006b). Although data mining methods have high prediction performance (Hossain & Muromachi, 2012; Jie Sun & Sun, 2015; Jian Sun et al., 2014; Xu, Wang, et al., 2013) and can accommodate correlation within variables for speed, flow, and occupancy (Hossain & Muromachi, 2012), they cannot provide explicit connections between crash probability and contributing factors. It is therefore difficult to interpret the crash mechanism and develop effective crash prevention countermeasures.

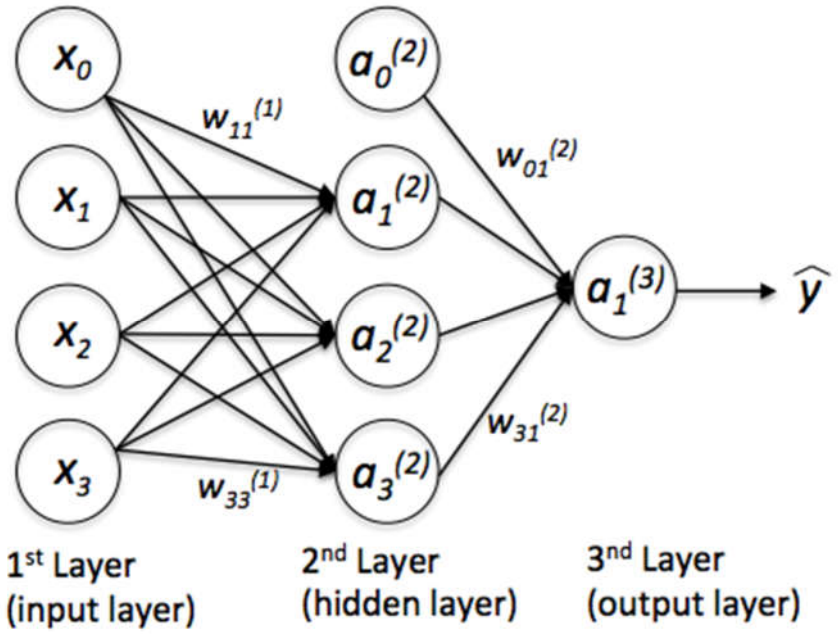


Figure 2-1 Illustration of neural network.

2.1.4 Crash Scenarios and Risk Factors

Various crash scenarios involving difference crash types (e.g., rear-end crashes, lane-change related crashes) and traffic conditions (e.g., different speed regimes, different traffic states) have

been investigated in real-time safety studies. Table 2-1 summarizes the studies that have reviewed different crash scenarios and their associated risk factors.

Most real-time crash studies analyze total crashes, and some focus on specific crash scenarios. Table 2-1 shows that risk factors vary across crash scenarios, suggesting that the crash mechanism may be different depending on the scenario, and it is therefore better to model crash scenarios separately. The table also indicates that different crash prevention strategies should be implemented for different scenarios. Although different crash scenarios are associated with different risk factors, traffic variations such as speed and volume stand out in most scenarios, implying that traffic stability is a significant factor contributing to the crash which needs to be managed through traffic control strategies.

Table 2-1 Summary of Real-Time Safety Studies by Crash Scenario

Crash Scenarios	**Studies**	**Risk Factors**
Total crashes	Oh et al. (Oh et al., 2001), Lee et al. (Lee, Hellinga, & Saccomanno, 2003a), Abdel-Aty et al. (M. Abdel-Aty, N. Uddin, A. Pande, F. M. Abdalla, & L. Hsia, 2004b), Abdel-Aty and Pande (M. Abdel-Aty & Pande, 2006), Abdel-Aty and Pemmanaboina (M. Abdel-	Average speed, Speed variation, Speed difference between upstream and downstream stations, Density variation, Average volume

	Aty & Pemmanaboina, 2006), Xu et al. (Xu et al., 2012), Yu and Abdel-Aty (Yu & Abdel-Aty, 2013), Xu et al. (Xu, Liu, & Wang, 2016)	
Rear-end crashes	Pande and Abdel-Aty (Anurag Pande & Abdel-Aty, 2006b)	In low-speed regime: Average occupancy, Speed variation In high-speed regime: Average speed, Average volume
Lane-change related crashes	Lee et al. (Lee, Abdel-Aty, & Hsia, 2006), Pande and Abdel-Aty (Anurag Pande & Abdel-Aty, 2006a)	Average speed, Speed variation, Volume variation, Occupancy difference between adjacent lane
Visibility related crashes	Abdel-Aty et al. (M. A. Abdel-Aty, Hassan, Ahmed, & Al-Ghamdi, 2012)	Average speed, Coefficient of variation* (CV) in speed
Secondary crashes	Xu et al. (Xu, Liu, Yang, & Wang, 2016)	Average volume, Average speed, Occupancy variation, Volume difference between adjacent lanes
Congestion related crashes	Zheng et al. (Zheng et al., 2010)	Speed variation

Crashes in different speed regimes	Abdel-Aty et al. (M. Abdel-Aty et al., 2005)	In low-speed regime: Average occupancy, Speed variation, CV in speed In high-speed regime: Average occupancy, Average volume, Volume variation
Crashes in different traffic states	Xu et al. (Xu et al., 2012), Li et al. (Z. B. Li, Wang, Chen, Liu, & Xu, 2013), Sun and Sun (Jie Sun & Sun, 2015)	F-F**: Average occupancy F-C**: Average occupancy, Average speed C-F**: Average speed, Speed variation C-C**: Average occupancy, Speed variation
Crashes in different levels of service (LOS)	Xu et al. (Xu et al., 2014)	LOS A&B: First order autocorrelation of speed, Occupancy difference between two periods LOS C: Cross correlation of occupancy between left- and right-most lane LOS D: Cross correlation of occupancy between left- and right-most lane, Occupancy difference between two periods LOS E: Average volume, Volume variation

		LOS F: Volume variation, Cross correlation of occupancy between left- and right-most lane

* Coefficient of variation = Standard deviation/Average.

** Represents upstream traffic state-downstream traffic state with F for free flow and C for congestion.

2.2 Cell Transmission Model

Traffic simulation models have increasingly been used to examine the effects of different intelligent transportation systems (ITS) on traffic flow. Two main types of traffic simulation models exist: microscopic simulation models and macroscopic simulation models. Microscopic models simulate the movements of individual vehicles, while macroscopic models simulate the evolvement of traffic flows. Microscopic simulation software such as VISSIM and PARAMICS has been applied to evaluate the safety impacts of various traffic control strategies, including variable speed limit (VSL) and ramp metering (M. Abdel-Aty, Cunningham, Gayah, & Hsia, 2008; M. Abdel-Aty, Dilmore, & Dhindsa, 2006; M. Abdel-Aty et al., 2007; Allaby, Hellinga, & Bullock, 2007; Lee, Hellinga, & Saccomanno, 2006). These studies developed RTCPMs using ILD data to predict the crash risk before and after the deployment of control strategies to assess their safety effects. However, the ILD data cannot be efficiently used to calibrate microscopic simulation models. Driver behavior parameters like target headway and reaction time are usually empirically adjusted through multiple trials with the intent to reflect the real traffic flow as accurately as possible. It is not guaranteed that these parameters are optimal. Simulated microscopic traffic cannot yet be used by RTCPMs after the model is calibrated, as it needs to first be aggregated into microscopic traffic flow variables such as the average flow, speed, and

occupancy. The whole procedure reveals a gap between RTCPMs and microscopic simulation models.

In contrast, macroscopic simulation models are compatible with RTCPMs. A macroscopic simulation model, CTM, has been applied for assessing VSL regarding its safety effect (Z. Li, Liu, Wang, & Xu, 2014; Z. Li, Liu, Xu, & Wang, 2016). In these studies, ILD data were used to develop RTCPMs and calibrate the simulation model through an analytical approach. The calibrated simulation model generated simulated macroscopic traffic flow data which can be easily used as the input of RTCPMs. Therefore, the macroscopic simulation model, CTM, was selected as the simulation tool in this dissertation.

A CTM can take aggregated data (e.g., flow and density) from detector stations as input variables to simulate traffic conditions at unmeasured collections. A highway segment is first divided into several user-defined cells. As shown in Figure 2-2, a segment is divided into four cells with two ILD stations at the beginning of Cell 1 and the end of Cell 4. The CTM takes data from these two stations as inputs and estimates traffic conditions of locations without ILD stations. Equivalently, one can consider that the CTM instruments virtual loop detector stations function exactly as physical detector stations. CTM makes it possible for traffic conditions from virtual detectors close to the crash location to be used to develop RTCPMs as opposed to those from physical detector stations located farther away. For example, if a crash happens in Cell 2 as shown in Figure 2-2, traditional studies would collect traffic data from physical Station A and E, which are far away from the crash location. If the segment is instrumented with virtual stations, traffic conditions can be retrieved from virtual Station B, C, and D which are much closer to the crash location, and may better reflect the crash-prone traffic conditions.

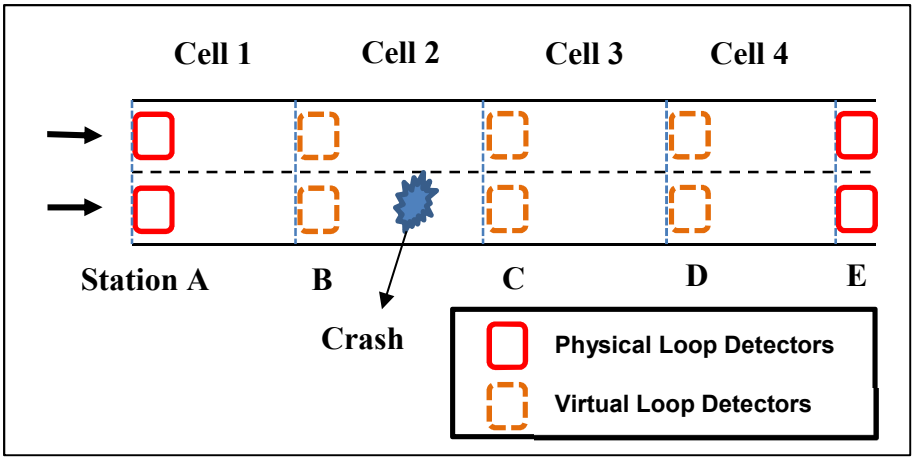


Figure 2-2 Illustration of a CTM setup.

Virtual stations can be placed anywhere without any restriction, so one unique layout of equally spaced virtual stations can be instrumented in different studies, and the developed RTCPMs and findings are comparable across studies. In this way, the spatial gap in traditional real-time safety studies can be resolved. CTM can also help resolve the temporal gap in traditional real-time safety studies. CTM can also be used to simulate future traffic conditions with suitable inputs. Previous crash detection studies using the period right before the crash occurrence cannot be applied to prevent crashes due to the lack of buffer time, but simulated future traffic conditions generated by CTM can be used by crash detection studies to predict crash likelihood for the period after the current moment.

In addition to addressing both the temporal and spatial gaps in traditional real-time safety studies, CTM can leverage all the development in the CTM field and develop robust applications for traffic planning and operations. A description of CTM and its improvements for more accurate traffic estimation are presented, followed by a summary of applications of safety-related traffic control strategies (TCSs) in CTM.

2.2.1 CTM for Traffic Estimation

CTM is a macroscopic traffic flow simulation model first proposed by Daganzo (Daganzo, 1994). CTM is a discretized framework for solving the Lighthill-Whitham-Richards (LWR) Model (Lighthill & Whitham, 1955; Richards, 1956). It partitions a highway into a series of cells and time into discretized time steps. The traffic density in each cell follows the law of conservation, thus evolving based on the relationship defined by the fundamental diagram.

CTM introduces the demand (sending flow) and supply (receiving flow) as functions of density in each cell. The flow entering into one cell is determined as the minimum of the demand of its upstream cell and the supply of its downstream cell.

In CTM, a fundamental diagram governs the flow-density relationship of each cell, and a triangular fundamental diagram (FD) (Drake, Schofer, & May Jr, 1967; Munjal, Hsu, & Lawrence, 1971) is often used. A typical FD is shown in Figure 2-3, where Q_C is the capacity flow, ρ_C is the critical density, ρ_J is the jam density, v is the free-flow speed, and w is the shockwave speed.

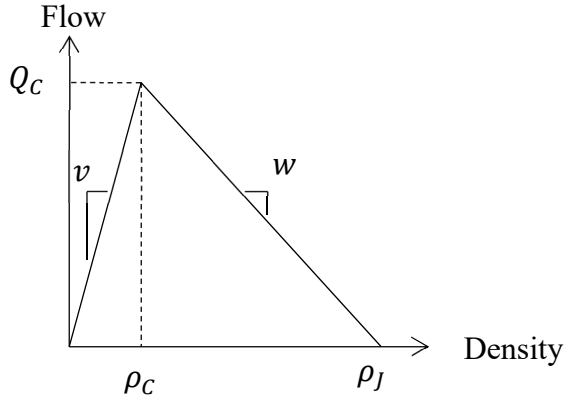


Figure 2-3 Triangular fundamental diagram.

The density for Cell *i* without on- or off-ramps is determined by Equation 2-1:

$$\rho_i(k+1) = \rho_i(k) + \frac{T}{l_i}(q_i(k+1) - q_i(k)) \quad (2\text{-}1)$$

where k is the time step index, $\rho_i(k)$ is the density of Cell *i* during the *k*th time step, T is the length of the time step, l_i is the length of Cell *i*, and $q_i(k)$ is the flow rate into Cell *i* during the *k*th time step. The flow rate is determined by the sending and receiving functions. For Cell *i*, the sending function $S_i(k)$ represents the maximum flow that may be supplied during the *k*th time step, and the receiving function $R_i(k)$ represents the maximum flow that may be received. The two functions are determined in Equations 2-2 and 2-3, respectively:

$$S_i(k) = \min(v_i\rho_i(k), Q_{C,i}) \quad (2\text{-}2)$$

$$R_i(k) = \min(Q_{C,i}, w_i(\rho_{J,i} - \rho_i(k))) \quad (2\text{-}3)$$

The entering flow rate into Cell *i*, $q_i(k)$, is determined by:

$$q_i(k) = \min(S_{i-1}(k), R_i(k)) \quad (2\text{-}4)$$

CTM can capture many important traffic phenomena including queue formation and dissipation, as well as shockwave propagation (Daganzo, 1994). CTM can take aggregated data such as flow and density as inputs meaning it operates sufficiently with aggregated data measured from ILD stations and can be applied to simulate traffic conditions at unmeasured locations. A variety of CTM variants have been developed, based on the original, to improve the accuracy of traffic estimation.

Muñoz et al. (Muñoz, Sun, Horowitz, & Alvarez, 2003) proposed a piecewise-linearized version of CTM, the switching-mode model (SMM). The SMM switches between five sets of linear difference equations, referred to as modes, according to measured mainline boundary data as inputs and the congestion status of cells in a roadway section. Its linear structure simplifies the control analysis, design, and data estimation. A 2-mile section of I-210 West with three

mainline ILD stations was used to observe morning rush-hour periods over several days to test the performance of both SMM and CTM in traffic data estimation. The traffic data from the first and last stations were used as inputs to both models, and the density at the second station was estimated by two models and compared with observed data. It showed that both models made approximately a 13% mean percentage error on average over all test days for estimating density. Sun et al. (X. Sun, Muñoz, & Horowitz, 2003) incorporated mixture Kalman filtering into SMM to simplify its logical mode-selection rules by considering only two modes instead of five., finding that the new model achieved an average of 10% mean percentage error for the density estimation.

The travel demand can yield some extent of variability, which is regarded as recurrent uncertainty or disturbance in traffic flow dynamics (Sumalee, Zhong, Pan, & Szeto, 2011). Therefore, CTM needs to be extended to account for those stochastic features of the traffic flow. Boel and Mihaylova (Boel & Mihaylova, 2006) proposed a stochastic compositional model which extends CTM by defining sending and receiving functions as random variables and specifying the dynamics of the average speed in each cell. The authors considered two extreme cases of traffic states - very light traffic and extremely congested conditions. Vehicles do not interact much during very light traffic conditions, so the sending function is defined by a binomial distribution. In contrast, vehicles interact often during extremely congested conditions, so the sending function is defined by a Gaussian distribution. This stochastic model was validated with both synthetic data and real data and found to provide a satisfactory performance. This model was then improved by incorporating a particle filtering (PF) framework (Mihaylova, Boel, & Hegyi, 2007). Similar to the Monte Carlo simulation, the PF framework can capture the

uncertainty of the traffic state by generating multiple samples. A stochastic component was also added to the sending function by Li et al. (Z. Li et al., 2016) to simulate the stop-and-go traffic.

Although those CTM variants can handle the uncertain travel demand, they depend on determined FDs and therefore fail to capture the uncertain travel supply. However, it was found that the FD can yield large variations due to congestion, driver behavior, and other conditions (Kim & Zhang, 2008; J. Li, Chen, Wang, & Ni, 2012; Wang, Li, Chen, & Ni, 2009). Sumalee et al. (Sumalee et al., 2011) proposed the stochastic CTM (SCTM) based on SMM to capture both the randomness in travel demand and supply. In the SCTM, the stochastic demand is characterized by random in-flow patterns, and the stochasticity of the sending and receiving functions is governed by the random FD parameters such as the free-flow speed, critical density, and so on. In contrast to SMM, the traffic state of each cell in the SCTM is not deterministic but stochastic meaning any of the five modes are possible. The proposed SCTM was validated with real data and found to be reliable by achieving an average error rate of approximately 7% error rate in density estimation.

The CTM variants provide more accurate traffic estimates and reproduce real-world traffic phenomena; therefore, these variants can be applied to accurately simulate traffic conditions where ILD stations are not available. As the foundation of simulation data-based RTCPMs, accurately simulated traffic data generated by CTM variants warrant the crash prediction performance. CTM can be customized to improve the RTCPMs. Although weather and lighting conditions are not a focus in most CTM improvement studies, these conditions need to be taken into account when applying the CTM in real-time safety studies. Moreover, most CTM improvement studies focus on improving the estimation accuracy of the overall traffic such as flow or density rather than traffic variation variables (e.g., speed variation and flow variation),

which are significant real-time safety-related variables. The CTM has the potential to consider these aspects and better serve real-time safety studies.

2.2.2 Safety Related Traffic Control Strategies in CTM

This section summarizes two traffic control strategies (TCSs) – the variable speed limit and ramp metering – that have shown promise in improving the safety and have been implemented in the CTM framework (including CTM and its variants). The CTM model is trustworthy in simulating TCSs, as it is founded on sound traffic theory (Hadiuzzaman & Qiu, 2013). It also has other attractive features: 1) it is parsimonious as it only needs a few parameters which can be estimated both online and off-line; 2) it requires quite low computation effort to predict the traffic variables in real-time (Hadiuzzaman & Qiu, 2013).

Variable speed limit (VSL) is a traffic control technique that is used to increase mobility and reduce crash risks on freeway mainlines. Unlike typical static speed limit signs, the VSL dynamically posts a speed limit based on current traffic, weather, traffic safety level or other conditions. Although the VSL is mainly designed to improve mobility, its effect on safety has also been demonstrated. VSL has been reported to reduce the crash risks by 10-80% (M. Abdel-Aty et al., 2008; M. Abdel-Aty et al., 2006; M. Abdel-Aty et al., 2007; Allaby et al., 2007; Choi & Oh, 2016; Hellinga & Mandelzys, 2011; Lee & Abdel-Aty, 2008; Lee, Hellinga, & Saccomanno, 2006; Z. Li, Li, Liu, Wang, & Xu, 2014; Z. Li, Liu, et al., 2014; Z. Li et al., 2016).

While the effect of VSL on mobility has been extensively evaluated using the CTM (Hadiuzzaman & Qiu, 2013; Han, Hegyi, Yuan, & Hoogendoorn, 2017; Han, Hegyi, Yuan, Hoogendoorn, et al., 2017; Muralidharan & Horowitz, 2012), limited research has been conducted to assess the safety impact of VSL (Z. Li, Liu, et al., 2014; Z. Li et al., 2016). Li et al. (Z. Li, Liu, et al., 2014) developed a VSL control strategy that considers both the travel time and

crash risk near freeway recurrent bottlenecks. The study considered only rear-end collisions, as they are the primary crash type on freeways, especially during congestion. RTCPM was proposed to predict the rear-end crash risk using ILD data. A genetic algorithm (GA) was applied to determine the optimal control factors of VSL strategies. The simulation results showed that the VSL control reduced the rear-end crash risk at freeway recurrent bottlenecks by approximately 70% in the high demand scenario and approximately 82% in the moderate demand scenario. The same authors (Z. Li et al., 2016) then proposed the VSL control strategy in the CTM to reduce both the crash risk and injury severities on large-scale freeway segments. The CTM was modified to handle both the capacity drop and the stop-and-go traffic. Three scenarios with various VSL sign placements were evaluated, and the corresponding optimal control factors were determined using the GA. The results showed that the optimal VSL control strategy could reduce the crash risk and injury severity by approximately 23% and 15%, respectively.

Ramp metering is another effective TCS. It controls the on-ramp vehicle flow allowed to enter the freeway to avoid the traffic breakdown due to oversaturation. Its effectiveness in reducing the crash risk has been demonstrated by several studies (Lee, Hellinga, & Ozbay, 2006; C. Liu & Wang, 2013; Robinson & Doctor, 1989). Ramp metering has been proposed in the CTM framework, but only its mobility effect has been evaluated (G. Gomes & Horowitz, 2006; Gabriel Gomes, Horowitz, Kurzhanskiy, Varaiya, & Kwon, 2008; Muralidharan & Horowitz, 2012). Therefore, its safety impact needs to be assessed in the CTM framework using an RTCPM.

2.3 Summary of Critical Issues

This chapter provides a summary of existing real-time crash prediction studies and relevant CTM studies including recent model improvements and their applications in traffic management. Critical issues that need further investigation are summarized below:

1. Most previous real-time crash studies conducted predictive analysis, and the analysis to estimate causal effects of single traffic variables is lacking.
2. Most previous real-time crash studies did not use the time interval right before the crash occurred, though such a time interval has been proven to provide the best prediction performance. The models' prediction power was further compromised.
3. Different station layouts within and across real-time crash studies pose doubts on the consistency of findings in different studies.
4. Most studies used traffic data collected directly from loop detector stations nearest to the crash location. However, the distance between the crash location and nearest detector stations varies substantially. It is therefore uncertain how the traffic conditions at stations can reflect the actual conditions at the crash location.
5. Although RTCPMs are more compatible with macroscopic simulation models than microscopic simulation models, a very limited number of studies have used macroscopic simulation models to evaluate the safety impacts of TCSs.

2.4 References

Abdel-Aty, M., Cunningham, R., Gayah, V., & Hsia, L. (2008). Dynamic Variable Speed Limit Strategies for Real-Time Crash Risk Reduction on Freeways. *Transportation Research Record: Journal of the Transportation Research Board, 2078*, 108-116. doi:10.3141/2078-15

Abdel-Aty, M., Dilmore, J., & Dhindsa, A. (2006). Evaluation of variable speed limits for real-time freeway safety improvement. *Accident Analysis & Prevention, 38*(2), 335-345.

Abdel-Aty, M., & Pande, A. (2005). Identifying crash propensity using specific traffic speed conditions. *Journal of safety research, 36*(1), 97-108. doi:10.1016/j.jsr.2004.11.002

Abdel-Aty, M., & Pande, A. (2006). ATMS implementation system for identifying traffic conditions leading to potential crashes. *IEEE Transactions on Intelligent Transportation Systems, 7*(1), 78-91. doi:10.1109/tits.2006.869612

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., & Santos, C. D. (2007). Crash Risk Assessment Using Intelligent Transportation Systems Data and Real-Time Intervention Strategies to Improve Safety on Freeways. *Journal of Intelligent Transportation Systems, 11*(3), 107-120. doi:10.1080/15472450701410395

Abdel-Aty, M., & Pemmanaboina, R. (2006). Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. *IEEE Transactions on Intelligent Transportation Systems, 7*(2), 167-174. doi:10.1109/TITS.2006.874710

Abdel-Aty, M., Uddin, N., & Pande, A. (2005). Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1908), 51-58.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., & Hsia, L. (2004a). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*(1897), 88-95.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F. M., & Hsia, L. (2004b). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board, 1897*(1), 88-95.

Abdel-Aty, M. A., Hassan, H. M., & Ahmed, M. (2012). *Real-Time Analysis of Visibility Related Crashes: Can Loop Detector and AVI Data Predict Them Equally?* Paper presented at the Transportation Research Board 91st Annual Meeting.

Abdel-Aty, M. A., Hassan, H. M., Ahmed, M., & Al-Ghamdi, A. S. (2012). Real-time prediction of visibility related crashes. *Transportation Research Part C: Emerging Technologies, 24*, 288-298.

Ahmed, M., & Abdel-Aty, M. (2013). Application of Stochastic Gradient Boosting Technique to Enhance Reliability of Real-Time Risk Assessment: Use of Automatic Vehicle Identification and Remote Traffic Microwave Sensor Data. *Transportation Research Record: Journal of the Transportation Research Board*(2386), 26-34.

Al-Deek, H. M., Venkata, C., & Ravi Chandra, S. (2004). New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *Transportation Research Record: Journal of the Transportation Research Board, 1867*(1), 116-126.

Allaby, P., Hellinga, B., & Bullock, M. (2007). Variable Speed Limits: Safety and Operational Impacts of a Candidate Control Strategy for Freeway Applications. *IEEE Transactions on Intelligent Transportation Systems, 8*(4), 671-680. doi:10.1109/TITS.2007.908562

Boel, R., & Mihaylova, L. (2006). A compositional stochastic model for real time freeway traffic simulation. *Transportation Research Part B: Methodological, 40*(4), 319-334. doi:10.1016/j.trb.2005.05.001

Chatterjee, I., & Davis, G. A. (2016). Analysis of Rear-End Events on Congested Freeways by Using Video-Recorded Shock Waves. *Transportation Research Record: Journal of the Transportation Research Board*(2583), 110-118.

Choi, S., & Oh, C. (2016). Proactive Strategy for Variable Speed Limit Operations on Freeways Under Foggy Weather Conditions. *Transportation Research Record: Journal of the Transportation Research Board, 2551*, 29-36. doi:10.3141/2551-04

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Nat. Cancer Inst, 22*, 173-203.

Daganzo, C. F. (1994). The Cell Transmission Model: Network Traffic. *Transportation Research Part B-Methodological, 29*(2), 79-93.

Danczyk, A., & Liu, H. X. (2011). A mixed-integer linear program for optimizing sensor locations along freeway corridors. *Transportation Research Part B: Methodological, 45*(1), 208-217.

Davis, G. A., & Swenson, T. (2006). Collective responsibility for freeway rear-ending accidents? An application of probabilistic causal models. *Accident Analysis & Prevention, 38*(4), 728-736.

Drake, J. S., Schofer, J. L., & May Jr, A. D. (1967). A statistical analysis of speed-density hypotheses. in vehicular traffic science. *Highway Research Record*(154).

Golob, T. F., & Recker, W. W. (2001). Relationships Among Urban Freeway Accidents, Traffic Flow, Weather and Lighting Conditions. *California Partners for Advanced Transit and Highways (PATH)*.

Gomes, G., & Horowitz, R. (2006). Optimal freeway ramp metering using the asymmetric cell transmission model. *Transportation Research Part C-Emerging Technologies, 14*(4), 244-262. doi:10.1016/j.trc.2006.08.001

Gomes, G., Horowitz, R., Kurzhanskiy, A. A., Varaiya, P., & Kwon, J. (2008). Behavior of the cell transmission model and effectiveness of ramp metering. *Transportation Research Part C: Emerging Technologies, 16*(4), 485-513.

Hadiuzzaman, M., & Qiu, T. Z. (2013). Cell transmission model based variable speed limit control for freeways. *Canadian Journal of Civil Engineering, 40*(1), 46-56. doi:10.1139/cjce-2012-0101

Han, Y., Hegyi, A., Yuan, Y., & Hoogendoorn, S. (2017). Validation of an extended discrete first-order model with variable speed limits. *Transportation Research Part C: Emerging Technologies, 83*, 1-17.

Han, Y., Hegyi, A., Yuan, Y., Hoogendoorn, S., Papageorgiou, M., & Roncoli, C. (2017). Resolving freeway jam waves by discrete first-order model-based predictive control of variable speed limits. *Transportation Research Part C: Emerging Technologies, 77*, 405-420.

Hellinga, B., & Mandelzys, M. (2011). Impact of Driver Compliance on the Safety and Operational Impacts of Freeway Variable Speed Limit Systems. *Journal of transportation engineering, 137*(4), 260-268. doi:10.1061/(ASCE)TE.1943-5436.0000214

Hong, Z., & Fukuda, D. (2012). Effects of traffic sensor location on traffic state estimation. *Procedia-Social and Behavioral Sciences, 54*, 1186-1196.

Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*: John Wiley & Sons.

Hossain, M., & Muromachi, Y. (2012). A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention, 45*, 373-381. doi:10.1016/j.aap.2011.08.004

Hourdos, J., Garg, V., & Michalopoulos, P. (2008). Accident Prevention Based on Automatic Detection of Accident Prone Traffic Conditions: Phase I.

Hourdos, J., Garg, V., Michalopoulos, P., & Davis, G. (2006). Real-time detection of crash-prone conditions at freeway high-crash locations. *Transportation Research Record: Journal of the Transportation Research Board*(1968), 83-91.

Kim, T., & Zhang, H. (2008). A stochastic wave propagation model. *Transportation Research Part B: Methodological, 42*(7), 619-634.

Kwon, J., Petty, K., & Varaiya, P. (2007). Probe vehicle runs or loop detectors?: Effect of detector spacing and sample size on accuracy of freeway congestion monitoring. *Transportation Research Record: Journal of the Transportation Research Board*(2012), 57-63.

Lee, C., & Abdel-Aty, M. (2008). Testing effects of warning messages and variable speed limits on driver behavior using driving simulator. *Transportation Research Record: Journal of the Transportation Research Board*(2069), 55-64.

Lee, C., Abdel-Aty, M., & Hsia, L. (2006). Potential real-time indicators of sideswipe crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 41-49.

Lee, C., Hellinga, B., & Ozbay, K. (2006). Quantifying effects of ramp metering on freeway safety. *Accident Analysis & Prevention, 38*(2), 279-288.

Lee, C., Hellinga, B., & Saccomanno, F. (2003a). Proactive freeway crash prevention using real-time traffic control. *Canadian Journal of Civil Engineering, 30*(6), 1034-1041.

Lee, C., Hellinga, B., & Saccomanno, F. (2003b). Real-time crash prediction model for application to crash prevention in freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board*(1840), 67-77.

Lee, C., Hellinga, B., & Saccomanno, F. (2006). Evaluation of variable speed limits to improve traffic safety. *Transportation Research Part C: Emerging Technologies, 14*(3), 213-228.

Lee, C., Saccomanno, F., & Hellinga, B. (2002). Analysis of crash precursors on instrumented freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1784), 1-8.

Li, J., Chen, Q.-Y., Wang, H., & Ni, D. (2012). Analysis of LWR model with fundamental diagram subject to uncertainties. *Transportmetrica, 8*(6), 387-405.

Li, Z., Li, Y., Liu, P., Wang, W., & Xu, C. (2014). Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers. *Accident Analysis & Prevention, 72*, 134-145. doi:10.1016/j.aap.2014.06.018

Li, Z., Liu, P., Wang, W., & Xu, C. (2014). Development of a Control Strategy of Variable Speed Limits to Reduce Rear-End Collision Risks Near Freeway Recurrent Bottlenecks. *IEEE Transactions on Intelligent Transportation Systems, 15*(2), 866-877. doi:10.1109/TITS.2013.2293199

Li, Z., Liu, P., Xu, C., & Wang, W. (2016). Optimal Mainline Variable Speed Limit Control to Improve Safety on Large-Scale Freeway Segments: Optimal mainline variable speed limit. *Computer-Aided Civil and Infrastructure Engineering, 31*(5), 366-380. doi:10.1111/mice.12164

Li, Z. B., Wang, W., Chen, R. Y., Liu, P., & Xu, C. C. (2013). Evaluation of the Impacts of Speed Variation on Freeway Traffic Collisions in Various Traffic States. *Traffic Injury Prevention, 14*(8), 861-866. doi:10.1080/15389588.2013.775433

Lighthill, M. J., & Whitham, G. B. (1955). *On kinematic waves. II. A theory of traffic flow on long crowded roads.* Paper presented at the Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences.

Liu, C., & Wang, Z. (2013). Ramp Metering Influence on Freeway Operational Safety near On-ramp Exits. *International Journal of Transportation Science and Technology, 2*(2), 87-94.

Liu, H. X., & Danczyk, A. (2009). Optimal sensor locations for freeway bottleneck identification. *Computer‐Aided Civil and Infrastructure Engineering, 24*(8), 535-550.

Mihaylova, L., Boel, R., & Hegyi, A. (2007). Freeway traffic estimation within particle filtering framework. *Automatica, 43*(2), 290-300. doi:10.1016/j.automatica.2006.08.023

Munjal, P., Hsu, Y.-S., & Lawrence, R. (1971). Analysis and validation of lane-drop effects on multi-lane freeways. *Transportation Research, 5*(4), 257-266.

Muñoz, L., Sun, X., Horowitz, R., & Alvarez, L. (2003, 2003). *Traffic density estimation with the cell transmission model*.

Muralidharan, A., & Horowitz, R. (2012). *Optimal control of freeway networks based on the link node cell transmission model.* Paper presented at the American Control Conference (ACC), 2012.

Niven, D. J., Berthiaume, L. R., Fick, G. H., & Laupland, K. B. (2012). Matched case-control studies: a review of reported statistical methodology. *Clinical Epidemiology, 4*, 99-110. doi:10.2147/CLEP.S30816

Oh, C., Oh, J.-S., Ritchie, S., & Chang, M. (2001). *Real-time estimation of freeway accident likelihood.* Paper presented at the 80th Annual Meeting of the Transportation Research Board, Washington, DC.

Pande, A., & Abdel-Aty, M. (2006a). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention, 38*(5), 936-948. doi:10.1016/j.aap.2006.03.004

Pande, A., & Abdel-Aty, M. (2006b). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 31-40.

Pande, A., Abdel-Aty, M., Hsia, L., & Trb. (2005). Spatiotemporal variation of risk preceding crashes on freeways *Statistical Methods; Highway Safety Data, Analysis, and Evaluation; Occupant Protection; Systematic Reviews and Meta-Analysis* (pp. 26-36).

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology, 49*(12), 1373-1379.

Richards, P. I. (1956). Shock waves on the highway. *Operations Research, 4*(1), 42-51.

Robinson, J., & Doctor, M. (1989). *RAMP METERING STATUS IN NORTH AMERICA*. Retrieved from

Roshandel, S., Zheng, Z., & Washington, S. (2015). Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis & Prevention, 79*, 198-211. doi:10.1016/j.aap.2015.03.013

Shi, Q., & Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies, 58*, 380-394.

Shi, Q., Abdel-Aty, M., & Yu, R. (2016). Multi-level Bayesian safety analysis with unprocessed Automatic Vehicle Identification data for an urban expressway. *Accident Analysis & Prevention, 88*, 68-76.

Sumalee, A., Zhong, R. X., Pan, T. L., & Szeto, W. Y. (2011). Stochastic cell transmission model (SCTM): A stochastic dynamic traffic model for traffic state surveillance and assignment. *Transportation Research Part B: Methodological, 45*(3), 507-533. doi:10.1016/j.trb.2010.09.006

Sun, J., & Sun, J. (2015). A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies, 54*, 176-186. doi:10.1016/j.trc.2015.03.006

Sun, J., Sun, J., & Chen, P. (2014). Use of Support Vector Machine Models for Real-Time Prediction of Crash Risk on Urban Expressways. *Transportation Research Record: Journal of the Transportation Research Board, 2432*, 91-98. doi:10.3141/2432-11

Sun, X., Muñoz, L., & Horowitz, R. (2003). *Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control.* Paper presented at the Decision and Control, 2003. Proceedings. 42nd IEEE Conference on.

Wang, H., Li, J., Chen, Q.-Y., & Ni, D. (2009). *Speed-density relationship: From deterministic to stochastic.* Paper presented at the The 88th Transportation Research Board (TRB) Annual Meeting. Washington, DC.

Xu, C., Liu, P., & Wang, W. (2016). Evaluation of the predictability of real-time crash risk models. *Accident Analysis & Prevention, 94*, 207-215. doi:10.1016/j.aap.2016.06.004

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

Xu, C., Liu, P., Wang, W., & Li, Z. (2014). Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. *Transportation Research Part A: Policy and Practice, 69*, 58-70. doi:10.1016/j.tra.2014.08.011

Xu, C., Liu, P., Yang, B., & Wang, W. (2016). Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data. *Transportation Research Part C: Emerging Technologies, 71*, 406-418. doi:10.1016/j.trc.2016.08.015

Xu, C., Tarko, A. P., Wang, W., & Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention, 57*, 30-39. doi:10.1016/j.aap.2013.03.035

Xu, C., Wang, W., & Liu, P. (2013). A genetic programming model for real-time crash prediction on freeways. *IEEE Transactions on Intelligent Transportation Systems, 14*(2), 574-586.

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention, 51*, 252-259.

Zheng, Z., Ahn, S., & Monsere, C. M. (2010). Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention, 42*(2), 626-636. doi:10.1016/j.aap.2009.10.009

CHAPTER 3 ESTIMATING CAUSAL EFFECTS OF CONTRIBUTING FACTORS ON CRASHES

3.1 Introduction

The wide deployment of advanced transportation information systems (ATIS) has made the collection, storage, and processing of real-time traffic data readily available. Now researchers can gather real-time information pertaining to crash occurrence. Based on real-time traffic data, various real-time crash prediction models (RTCPM) have been proposed to identify the contributing factors of crashes. Crash is usually considered as a binary variable (yes/no) in almost all real-time crash prediction studies. A crash case represents the traffic conditions prior to a crash, while a non-crash case represents crash-free traffic conditions. The traffic condition in a short time interval before a crash is determined by reviewing reported crash time and location. Any traffic condition that is unrelated to crashes could be a non-crash case.

Crashes are rare events and therefore, non-crash cases are large in volume. Most previous studies randomly sampled non-crash cases, either by matched case-control design or unmatched design. The matched case-control design compares the level of risk factors in two similar groups, one that includes the outcome and one that does not. In a matched case-control design, each crash is considered as a "case", and a non-crash case is treated as "control" by matching confounding factors that are related to both the crash probability and the traffic variables of interest. The matched case-control design is used to remove the noise of confounding factors and investigate the underpinning risk factors. Although the matched case-control design is expected to increase the accuracy of variable estimates in a crash prediction model, most studies only treated non-traffic variables such as weather, location, and time as confounding factors while overlooking the potential confounding effect within traffic variables.

Therefore, the true causal effects of one traffic factor would be compromised given the presence of other confounding traffic factors. As a result, the model findings can be biased or inaccurate.

The objective of this chapter is to measure the causal effects of speed variation on the probability of a crash using the propensity score based method. The propensity score based method can eliminate the nuisance of confounding traffic factors related to the traffic factor that is being assessed. This is done by generating a sample of non-crash cases which have similar distributions of confounding traffic factors related to the traffic factor of interest. Then a binary logit model will be applied to assess the causal effect of that factor.

3.2 Literature Review

3.2.1 Study Design in Traditional Real-Time Crash Studies

Both crash and non-crash cases are needed to develop RTCPMs and identify crash-prone patterns. A crash case involves the traffic conditions prior to a crash occurrence and is restricted by the crash. A non-crash case involves crash-free traffic conditions and could include any traffic conditions during crash-free days.

Two primary study designs – matched the case-control design and unmatched design – determine how non-crash cases are collected. The matched case-control design is an efficient means of studying rare diseases, and is widely applied in epidemiological studies (Niven, Berthiaume, Fick, & Laupland, 2012). Abdel-Aty (Mohamed Abdel-Aty, Uddin, Pande, Abdalla, & Hsia, 2004) introduced this study design to real-time crash studies. The matched case-control design compares the level of risk factors in two similar groups, one that includes the outcome and one that does not (Cornfield et al., 1959). In a matched case-control design, each crash is considered as a "case", and non-crash cases are selected as "controls" by matching confounding factors (i.e., location and time). Confounding factors are correlated with traffic

conditions and contribute to the crash occurrence. One example of a matched case-control design involves a crash occurring at 11:01 a.m. on Jan 17st, 2012 that uses traffic measurements from 10:56 to 11:01 a.m. (0-5-min period before the crash time) on the same day from its immediately upstream and downstream ILD stations. Traffic measurements are then collected from the same stations during the same period on crash-free days in 2012 as controls. The matched case-control design is constructed to remove the noise of confounding factors and investigate the risk factors of interest. The matched case-control design is expected to increase the accuracy of variable estimates in RTCPMs by controlling the confounding bias. This study design can greatly reduce the required size of non-crash cases. Due to the efficiency of the matched case-control study design, most real-time crash studies adopt this design for data collection (M. Abdel-Aty & Pande, 2005; Mohamed Abdel-Aty & Pemmanaboina, 2006; Mohamed Abdel-Aty, Uddin, & Pande, 2005; Mohamed Abdel-Aty et al., 2004; A. Pande, Abdel-Aty, Hsia, & Trb, 2005; Xu, Liu, Wang, & Li, 2012, 2014; Zheng, Ahn, & Monsere, 2010).

In an unmatched study design, non-crash cases are randomly selected. The safety impacts of all factors, including risk factors (e.g., traffic flow variables) and confounding factors (e.g., geometric design) are estimated based on a large sample. Compared to the matched case-control design, the unmatched design does not require matching confounding factors, and it therefore requires less effort to identify non-crash cases. However, a sufficiently large sample size is required to ensure accurate estimation, especially when the variable number is high (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). This drawback may be the reason why only a few studies have employed the unmatched study design (Anurag Pande & Abdel-Aty, 2006a, 2006b; Xu, Liu, & Wang, 2016; Xu, Wang, & Liu, 2013). Unlike the matched case-

control study design, the unmatched design allows for the estimation of both risk factors and confounding factors.

Although the matched case-control design is superior to unmatched study design as it requires smaller sample size and would provide more accurate estimates, almost all studies only control non-traffic variables such as weather, location, and time while overlooking the confounding effect within traffic variables. However, traffic volume, speed, and density are correlated with each other. Then among derived traffic variables such as the average and variance of these traffic measures, there may exist confounding variables for one traffic variable. These confounding variables could make that traffic variable appear significantly contributing to the crash occurrence, but the relationship could be in fact spurious.

3.2.2 Causal Effect

Consider a population of subjects, one subject receives a treatment if it is assigned to the treated group; or it is untreated if it is assigned to the control group. The subject would have a response, r_1, if it had been assigned to the treated group; and a response, r_0, if it had been assigned to the control group. The causal effect of this treatment is based on the comparison of r_1 and r_0, i.e., $r_1 - r_0$ or r_1/r_0 (Rosenbaum & Rubin, 1983). In fact, one subject can only be assigned to one treatment group, and only one response can be observed. Therefore, the causal effect of the treatment cannot be directly measured.

An alternative way is through a randomized experiment that randomly assigns subjects to different groups. In a randomized experiment, the responses of two groups can be compared because the subjects are likely to be similar in characteristics across groups. That means a randomized experiment removes the nuisance of confounding factors and yields unbiased estimates of average treatment effects (Rosenbaum, 2002; Rosenbuam, 2010). However, a

randomized experiment cannot be achieved in many situations such as in observational studies. Observational studies do not have control on the assignments of subjects to a treated group or a control group and simply collect after-fact data. Therefore, subjects in a treat group may be very different from those in a control group in observational studies. The causal effect of the treatment could be biased in observational studies if other confounding variables are not properly controlled for. As one type of observational study, real-time crash studies also suffers from this issue.

One method to reduce confounding is through multivariate regression that regresses the outcome on the treatment and other confounding variables. One critical issue of this method is that when groups have different variable distributions, the results are dependent on the specific form of the model and are determined by unreliable extrapolations (Rubin, 1997). The extent of unreliability could be exacerbated when many confounding variables lack adequate overlap (Rubin, 1997). Another popular method is the propensity score based method that can mimic randomized experiments. The propensity score, $P(T = 1|\boldsymbol{X})$, denotes the probability that a subject is assigned to the treated group ($T = 1$) given its characteristics (Rosenbaum & Rubin, 1983). The covariate distribution of $\boldsymbol{X}$ may be different across the treated group and the control group. However, conditional on the propensity score, the covariate distribution should be similar between the two groups (Rosenbaum & Rubin, 1983). Therefore, the propensity score based method is able to reconstruct the treatment and control group so that they are similar in variable distributions. Compared to multivariate regression, the propensity score based method is less sensitive to model misspecification (Drake, 1993). Moreover, the propensity score based method is more robust when outcome (e.g., crash occurrences) are rare and treatment is common (Braitman & Rosenbaum, 2002).

The application of propensity score based methods in traffic safety is scarce. The first application in traffic safety by Davis (Davis, 2000) applied the propensity score based method to account for the site selection bias in estimating the accident reduction factor of safety treatments. Other researchers applied the propensity score based method to estimate the Crash Modification Factor (CMF) of signal installation (Aul & Davis, 2006), evaluate the effectiveness of child safety restraint (Durbin, Elliott, & Winston, 2009), examine the effectiveness of lighting at intersections (Sasidharan & Donnell, 2013) and assess the effects of continuous green T intersections (Wood & Donnell, 2016). The propensity score based method applied in most of these studies is the propensity score matching which only matches treated subjects and control subjects with similar propensity score to generate a sub-population.

Another propensity score based method, inverse probability of treatment weighting (IPTW) using the propensity score, has gained popularity in observational studies (Austin & Stuart, 2015). IPTW using the propensity score assigns weights based on the propensity score to subjects. This method can create a synthetic sample in which the covariate distribution is independent of treatment assignment (Joffe, Ten Have, Feldman, & Kimmel, 2004). In a recent study, this method has been used to assess the safety effectiveness of 20 MPH zones in London (Li & Graham, 2016).

3.3 Methodology

In this study, a logit model is used to estimate the propensity score:

$$P(T = 1|\boldsymbol{X}) = \frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})} \quad (3\text{-}1)$$

where $\boldsymbol{X}$ is the vector of explanatory covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients. Based on the estimated propensity score, $\hat{P}(T = 1|\boldsymbol{X})$, the IPTW method assigns weight to each subject. The weight for each subject depends on its propensity score as defined by:

$$\widehat{w}_i(T,X) = \frac{T}{\hat{P}(T=1|\boldsymbol{X})} + \frac{1-T}{1-\hat{P}(T=1|\boldsymbol{X})} \quad (3\text{-}2)$$

Therefore, the weight is $\frac{1}{\hat{P}(T=1|\boldsymbol{X})}$ for a subject assigned to the treated group and $\frac{1}{1-\hat{P}(T=1|\boldsymbol{X})}$ for a subject assigned to the control group. The weight for a subject is equal to the inverse of the probability of being assigned to the group that the subject is actually assigned to.

3.4 Data Description and Processing

Three data sources were used to develop a comprehensive approach:

a) 1-min time interval traffic information from the WisTransPortal V-SPOC (Volume, Speed, and Occupancy) application suite (Parker & Tao, 2006);

b) crash data from the web-based query and retrieval facility for Wisconsin Department of Transportation crash data and from reports archived in the WisTransPortal data management system; and

c) weather information (e.g. snow, rain) from the Road Weather Information System (RWIS) in WisTransPortal.

A 4.15-mile corridor on I-94 East in Waukesha, WI was selected as the study site. The site was selected based on the following criteria: spacing of loop detector stations, traffic data quality, and crash sample size. The selected roadway corridor, as shown in Figure 3-1, has three lanes with one on-ramp and one off-ramp. The corridor consists of three segments, $\boldsymbol{S_1}$, $\boldsymbol{S_2}$, and $\boldsymbol{S_3}$, which are 1.77-mile, 0.79-mile and 1.59-mile long, respectively. Segment $\boldsymbol{S_2}$ starts at the end of the off-ramp and ends at the beginning of the on-ramp. The posted speed limit was 65 MPH in $\boldsymbol{S_1}$, and 55 MPH in $\boldsymbol{S_2}$ and $\boldsymbol{S_3}$. Other roadway characteristics such as lane width and shoulder width did not change along the corridor.

The corridor was instrumented with seven mainline loop detector stations: $\mathbf{N_1}$, $\mathbf{N_2}$, …, $\mathbf{N_7}$. The stations are referred to as physical stations so as to differentiate them from the virtual detectors introduced in the later chapters. The seven stations space between 0.50 and 1.00 mile, with an average of 0.69 mile and a standard deviation of 0.20 mile. One loop detector station was located on the off-ramp, but no stations exist on the on-ramp. The traffic flow of the on-ramp can be imputed based on the conservation of vehicles using the flows from the nearest upstream and downstream detector stations.

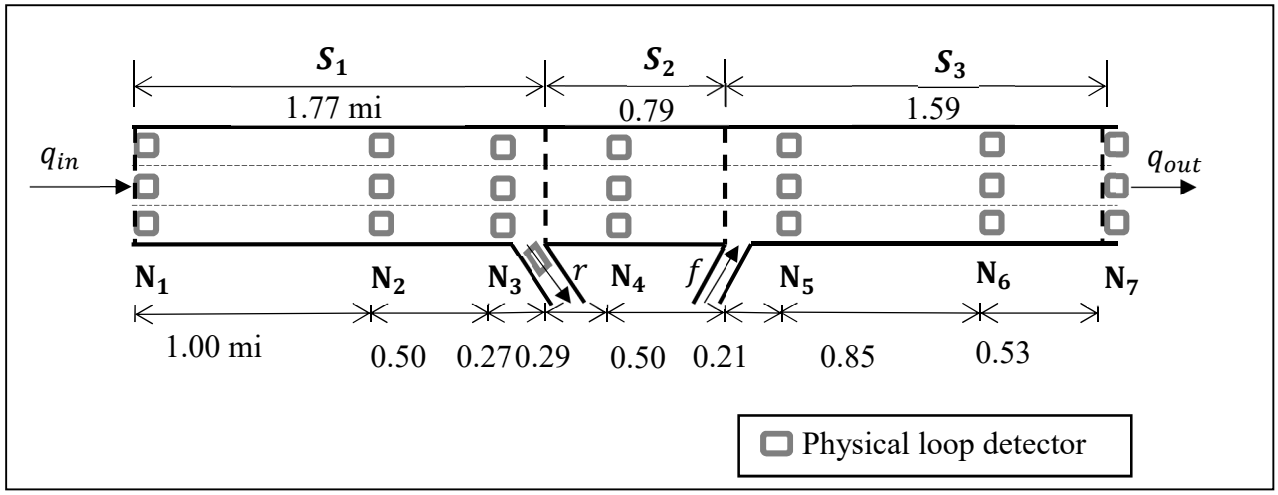


Figure 3-1 Layout of physical loop detector stations.

Crashes occurred from 2012 to 2014 were used. Any crash that happened within one hour after a crash occurrence was considered a secondary crash and was subsequently removed as indicated in (Hirunyanitiwattana & Mattingly, 2006). Crashes with missing times were excluded, as crash time is required to retrieve the traffic data.

A critical component of developing a crash prediction model is the knowledge of the traffic conditions experienced by the vehicle right before a crash; therefore, it is important to know the exact time in which a crash occurs. Crash times are sometimes rounded to the nearest

5-minute time stamp, and are therefore not reliable (Golob & Recker, 2003; Kockelman & Ma, 2010). Crash times in this study were carefully reviewed, and no rounding issue was found. Crashes were randomly sampled and validated with the time when abrupt changes in traffic conditions were observed (Mohamed Abdel-Aty et al., 2005; Zheng et al., 2010). The validation result was positive, and the crash times from the database were used as the actual crash occurrence times.

After crash cases with missing ILD data were removed, a total of 113 crashes remained for crash analysis. For each crash case, its exact location on the freeway was determined based on its longitude and latitude information. Based on the location, its nearest upstream and downstream ILD stations were located. 1-minute interval ILD data from one upstream and one downstream ILD station from each crash location were collected 0 to 5 minutes prior to the crash. 2,260 non-crash cases with a 20:1 non-crash to crash case ratio were randomly selected from 1,578,240-min intervals in 2012-2014 (60 min×24 h×1096 days in 2012-2014). Only the non-crash cases that are not within 2 hours from any crash were selected. The 5-min traffic data consisting of data from five 1-min intervals were retrieved from physical stations for non-crash cases in the same way that data were retrieved for crash cases.

Given the 1-minute interval ILD data from one upstream and one downstream ILD station from each crash location, means and standard deviations of flow, speed, and occupancy were calculated for all crash and non-crash events at 0 to 5 minutes prior to the crash. Additional non-traffic variables such as curve presence, ramp presence, and weather condition were included. Table 3-1 presents the candidate variables for analysis.

Table 3-1 Candidate Variables

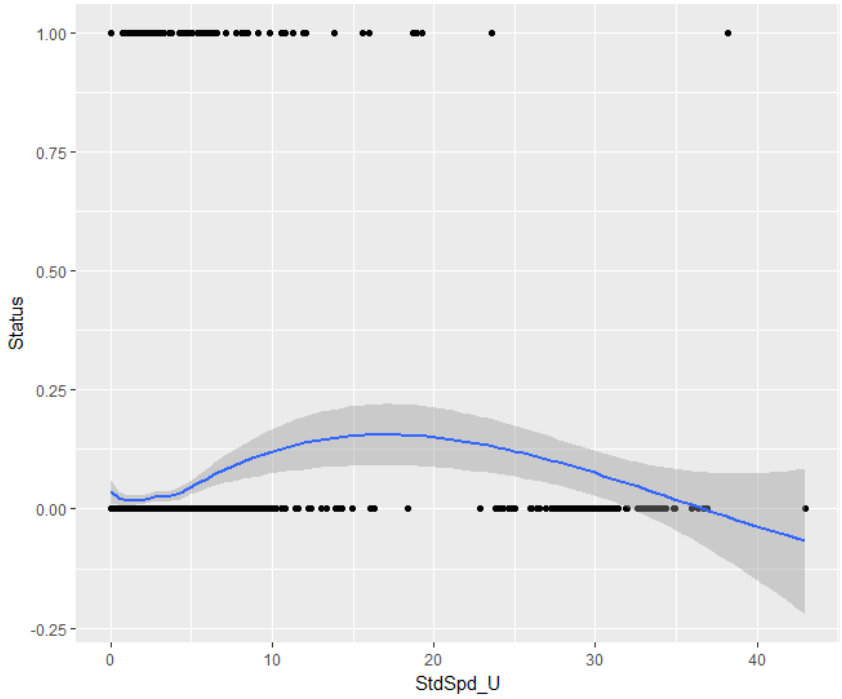
Variable	Description

AvgVol_U	Average 1-min volume at the upstream station (veh/h)
AvgDen_U	Average 1-min density at the upstream station (veh/mi)
AvgSpd_U	Average 1-min speed at the upstream station (mi/h)
StdVol_U	Standard deviation of 1-min volume at the upstream station (veh/h)
StdDen_U	Standard deviation of 1-min density at the upstream station (veh/mi)
StdSpd_U	Standard deviation of 1-min speed at the upstream station (mi/h)
AvgVol_D	Average 1-min volume at the downstream station (veh/h)
AvgDen_D	Average 1-min density at the downstream station (veh/mi)
AvgSpd_D	Average 1-min speed at the downstream station (mi/h)
StdDen_D	Standard deviation of 1-min density at the downstream station (veh/mi)
StdSpd_D	Standard deviation of 1-min speed at the downstream station (mi/h)
StdSpd_D	Standard deviation of 1-min speed at the downstream station (mi/h)
Curve	1 = Horizontal curve section; 0 = otherwise
OnRamp	1 = an on-ramp between upstream and downstream stations; 0 = otherwise
OffRamp	1 = an off-ramp between upstream and downstream stations; 0 = otherwise
Rain	1 = if the weather is rainy; 0 = otherwise
Snow	1 = if the weather is snowy; 0 = otherwise

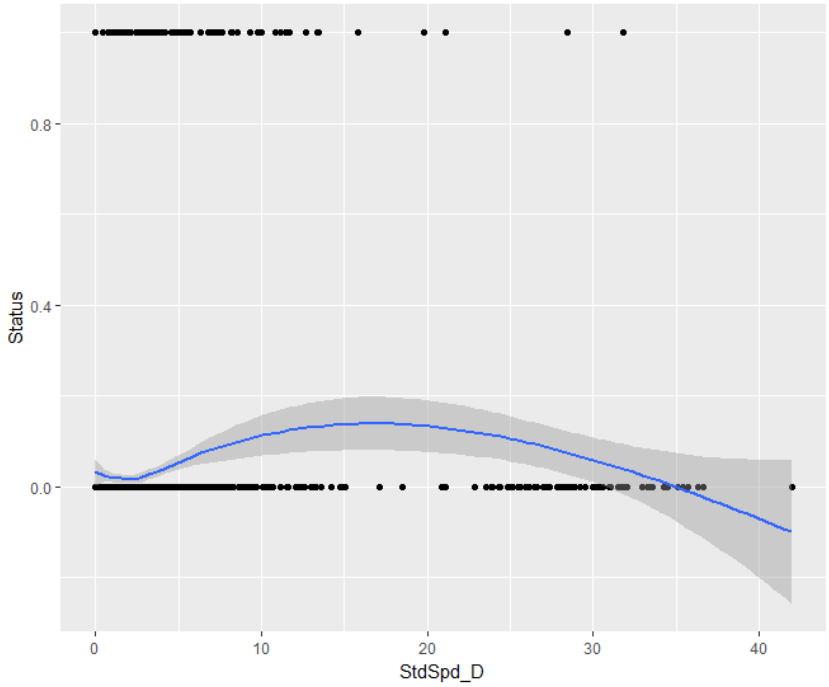
3.5 Analysis

Roshandel (Roshandel, Zheng, & Washington, 2015) conducted a systematic review of real-time contributing factors to the crashes across studies. They found that the speed variation, the standard deviation of speed, carries the highest odds ratio among all contributing factors. In this chapter, both speed variation variables, StdSpd_U and StdSpd_D, were tested for their causal effects on crash risk.

First, the two speed variation variables need to be converted to two treatment groups, high speed variation group (treated group) and low speed variation group (control group) based on a cutoff value. The crash status is plotted against the two speed variation variables as shown by black dots in Figure 3-2. The blue lines are the smooth plots. The crash probability is higher when the blue line is close to the status of 1. Both blue lines first decrease gently, then increase from the value around 4, and decrease from the value around 18. However, the trend is valid when the value is smaller than 15, but it is questionable when the value is above 15. As the value exceeds 15, there are a few isolated non-crash points and some dense non-crash points. Those dense points could be outliers that cause the plausible downtrend. There are 97 out of 2,260 and 79 out of 2,260 non-crash cases with StdSpd_U and StdSpd_D above 15, respectively. It is reasonable to consider these cases as outliers when either StdSpd_U and StdSpd_D is above 15. Subsequently, 11 crash cases and 90 non-crash cases were excluded from the sample. The final sample consists of 102 crash cases and 2,170 non-crash cases.



(a)



(b)

Figure 3-2 (a) Crash status against StdSpd_U; (b) Crash status against StdSpd_D.

The turning points from which the blue lines start to increase monotonically are set as the cutoff values for both speed variation variables. They are 3.33 for StdSpd_U and 2.19 for StdSpd_D, respectively. Two treatments were considered here, high upstream speed variation (HUSV) and high downstream speed variation (HDSV). HUSV is treated when StdSpd_U is above 3.33 and is control otherwise; HDSV is treated when StdSpd_D is above 2.19 and is control otherwise. The distribution of crash/non-crash cases by treatment group is presented in Table 3-2.

Table 3-2 Distribution of Crash Outcomes by Treatment Group

Crash Outcome	HUSV	HDSV

	Treated	Control	Treated	Control
1 (Crash)	44	58	68	34
0 (Non-Crash)	551	1,619	1,081	1,089

Then a propensity score model is developed for the two treatments. According to (Brookhart et al., 2006), all variables associated with the outcome should be included into the propensity score model regardless of their association with the treatment assignment. The correlation of all variables with the crash outcome was checked, and it was found that only Off_Ramp is not significantly related to the crash outcome. Therefore, all variables except Off_Ramp were included into the propensity model.

Two propensity models were developed for HUSV and HDSV using the logit. The results are presented in Table 3-3. Note StdSpd_D was included in the propensity model for HUSV, and StdSpd_U was included in the model for HDSV.

Table 3-3 Propensity Score Model

	HUSV			**HDSV**		
Variable	**Estimate**	**Standard Error**	**P-value**	**Estimate**	**Standard Error**	**P-value**
Intercept	-3.194	0.484	<0.001	-3.594	0.492	<0.001
StdSpd_U				0.161	0.035	<0.001
StdSpd_D	0.177	0.033	<0.001			
AvgVol_U	-0.00034	0.00012	0.005	0.000115	0.000114	0.314

StdVol_U	-0.00227	0.00049	<0.001	0.000376	0.000359	0.295
AvgDen_U	-0.00653	0.00540	0.226	0.013	0.005	0.012
StdDen_U	0.177	0.026	<0.001	-0.043	0.016	0.008
AvgSpd_U	0.04688	0.00676	<0.001	-0.020	0.004	<0.001
AvgVol_D	0.00012	0.00012	0.349	-0.001	0.000	<0.001
StdVol_D	-0.00011	0.00038	0.769	-0.002	0.000	<0.001
AvgDen_D	0.00761	0.00519	0.143	0.013	0.006	0.030
StdDen_D	-0.00197	0.01450	0.892	0.162	0.026	<0.001
AvgSpd_D	-0.022	0.005	<0.001	0.071	0.007	<0.001
Curve1	-0.476	0.133	<0.001	0.464	0.107	<0.001
On_Ramp1	0.636	0.134	<0.001	0.252	0.128	0.049
Snow1	-0.126	0.319	0.692	0.443	0.274	0.106

Based on Equation 3-1 and 3-2, the estimated propensity score and treatment weight for each crash/non-crash case is obtained. Then a weighted sample is generated. The balance of variables is checked for the unadjusted sample and the weighted sample using the standardized mean difference (SMD) between the treated group and the control group. One variable is balanced when it has similar distributions in the treated group and control group. For a continuous variable, the SMD is defined as:

$$\text{SMD} = \frac{\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}}{\sqrt{\frac{s^2_{\text{treatmetn}} + s^2_{\text{control}}}{2}}} \tag{3-3}$$

where $\bar{x}_{\text{treatment}}$ and $\bar{x}_{\text{control}}$ denote the sample mean of the variable in the treated and control subjects, respectively, while $s^2_{\text{treatmetn}}$ and s^2_{control} are their variance, respectively.

For dichotomous variables, the SMD is defined as:

$$\text{SMD} = \frac{\hat{p}_{\text{treatment}} - \hat{p}_{\text{control}}}{\sqrt{\frac{\hat{p}_{\text{treatment}}(1-\hat{p}_{\text{treatment}}) + \hat{p}_{\text{control}}(1-\hat{p}_{\text{control}})}{2}}} \quad (3\text{-}4)$$

where $\hat{p}_{\text{treatment}}$ and $\hat{p}_{\text{control}}$ denote the prevalence or mean of the dichotomous variable in treated and control subjects, respectively. A SMD less than 0.1 indicates that the corresponding variable is well balanced (Austin, 2009).

Table 3-4 presents the balance check results of the unadjusted and weighted samples for the two treatments. It shows that almost all variables in the adjusted sample have SMD values greater than 0.1, indicating they are imbalanced before weighting. However, all variables have SMD values smaller than 0.1 in the weighted sample, indicating they are all balanced after weighting. It implies that all variables have similar distributions in two treatment groups after weighting.

Table 3-4 Balance Check Results of Unadjusted and Weighted Samples

HUSV						
Variable	**Unadjusted Sample**			**Weighted Sample**		
	Control	**Treated**	**SMD**	**Control**	**Treated**	**SMD**
N*	1677	595		494.49	507.22	
StdSpd_D	2.43 (1.61)**	3.04 (2.15)	0.321	2.67 (1.99)	2.68 (1.73)	**0.005**
AvgVol_U	2222.4 (1469.1)	2338.5 (1580.8)	**0.076**	2102.3 (1356.1)	2104.5 (1520.7)	**0.002**
StdVol_U	394.54 (254.66)	411.98 (263.95)	**0.067**	397.13 (247.59)	395.69 (259.51)	**0.006**
AvgDen_U	35.79 (30.16)	53.64 (58.02)	0.386	37.40 (38.93)	38.29 (39.38)	**0.023**
StdDen_U	6.55 (5.40)	11.93 (14.33)	0.497	7.37 (7.44)	7.63 (6.66)	**0.037**
AvgSpd_U	62.50 (17.21)	60.24 (17.30)	0.131	65.24 (13.65)	64.77 (12.58)	**0.036**
AvgVol_D	2302.9 (1448.1)	2503.2 (1721.7)	0.126	2240.0(1527.9)	2242.8 (1654.2)	**0.002**

StdVol_D	413.93 (253.50)	412.67 (266.59)	**0.005**	399.9 (265.0)	400.70 (262.59)	**0.003**
AvgDen_D	38.56 (31.25)	58.02 (55.99)	0.429	43.05 (41.71)	43.76 (41.86)	**0.017**
StdDen_D	7.30 (6.25)	11.11 (12.35)	0.389	8.29 (8.60)	8.41 (8.72)	**0.014**
AvgSpd_D	62.41 (14.83)	54.39 (17.01)	0.503	57.72 (17.89)	58.02 (14.67)	**0.018**
Curve=1 (%)	505 (30.1)	126 (21.2)	0.206	103.8 (21.0)	108.3 (21.4)	**0.009**
On_Ramp=1(%)	233 (13.9)	151 (25.4)	0.292	129.5 (26.2)	132.8 (26.2)	**<0.001**
Snow=1 (%)	46 (2.7)	20 (3.4)	**0.036**	14.7 (3.0)	17.0 (3.4)	**0.022**
HDSV						
Variable	**Unadjusted Sample**			**Weighted Sample**		
	Control	**Treated**	**SMD**	**Control**	**Treated**	**SMD**
N	1123	1149		796.04	805.8	
StdSpd_U	2.52 (1.50)	2.88 (1.98)	0.205	2.60 (1.61)	2.62 (1.51)	**0.012**
AvgVol_U	2274.5 (1412.2)	2231.4 (1580.8)	**0.029**	2153.1(1415.4)	2162.4 (1509.1)	**0.006**
StdVol_U	407.86 (244.26)	390.56 (269.05)	**0.067**	388.81(240.07)	390.04 (261.05)	**0.005**
AvgDen_U	37.02 (29.12)	43.82 (48.38)	0.17	35.71 (30.69)	36.03 (33.92)	**0.01**
StdDen_U	7.23 (6.81)	8.68 (10.66)	0.163	7.02 (7.01)	7.12 (7.66)	**0.013**
AvgSpd_U	63.85 (13.74)	60.01 (19.93)	0.225	63.59 (14.45)	63.59 (16.51)	**<0.001**
AvgVol_D	2390.1 (1519.3)	2321.5 (1533.7)	**0.045**	2299.7(1413.9)	2294.0 (1508.4)	**0.004**
StdVol_D	420.48 (259.39)	406.88 (254.44)	**0.053**	418.46(239.85)	417.83 (261.82)	**0.003**
AvgDen_D	40.33 (31.23)	46.91 (47.10)	0.165	39.11 (31.91)	39.36 (32.44)	**0.008**
StdDen_D	7.18 (5.32)	9.39 (10.56)	0.264	7.20 (5.41)	7.31 (5.13)	**0.02**
AvgSpd_D	59.15 (16.87)	61.44 (14.66)	0.145	63.25 (10.21)	63.11 (10.28)	**0.013**
Curve=1 (%)	236 (21.0)	395 (34.4)	0.302	213.7 (26.8)	211.7 (26.3)	**0.013**
On_Ramp=1(%)	186 (16.6)	198 (17.2)	**0.018**	147.8 (18.6)	148.1 (18.4)	**0.005**
Snow = 1 (%)	26 (2.3)	40 (3.5)	**0.07**	23.7 (3.0)	26.0 (3.2)	**0.015**

Note: all SMDs below 0.1 are in bold.

* Denotes the subject count. It should be integers for unadjusted sample, while it could be decimals due to decimal weights.

** Mean with standard deviation in the parentheses.

A logit model for the crash outcome is developed on each treatment using the weighed sample. The odds ratio (OR) is obtained from the model as the causal effect and is compared with the predictive effect obtained from the unadjusted model. The OR for the two treatments using unadjusted and weighted samples are shown in Table 3-5. If the 95% confidence interval does not include 1, the effect is considered significant. For both HUSV and HDSV treatment, the causal effect is not significant, while it shows spurious significance based on the unadjusted sample.

Since different cutoff values may result in different assignments of subjects to two treatment groups and may yield different causal effects, the sensitivity analysis has been conducted to obtain the causal effect and predictive effect based upon different cutoff values. Cutoff values from 2 to 6 with 0.1 as the increment for both speed variation variables are tested.

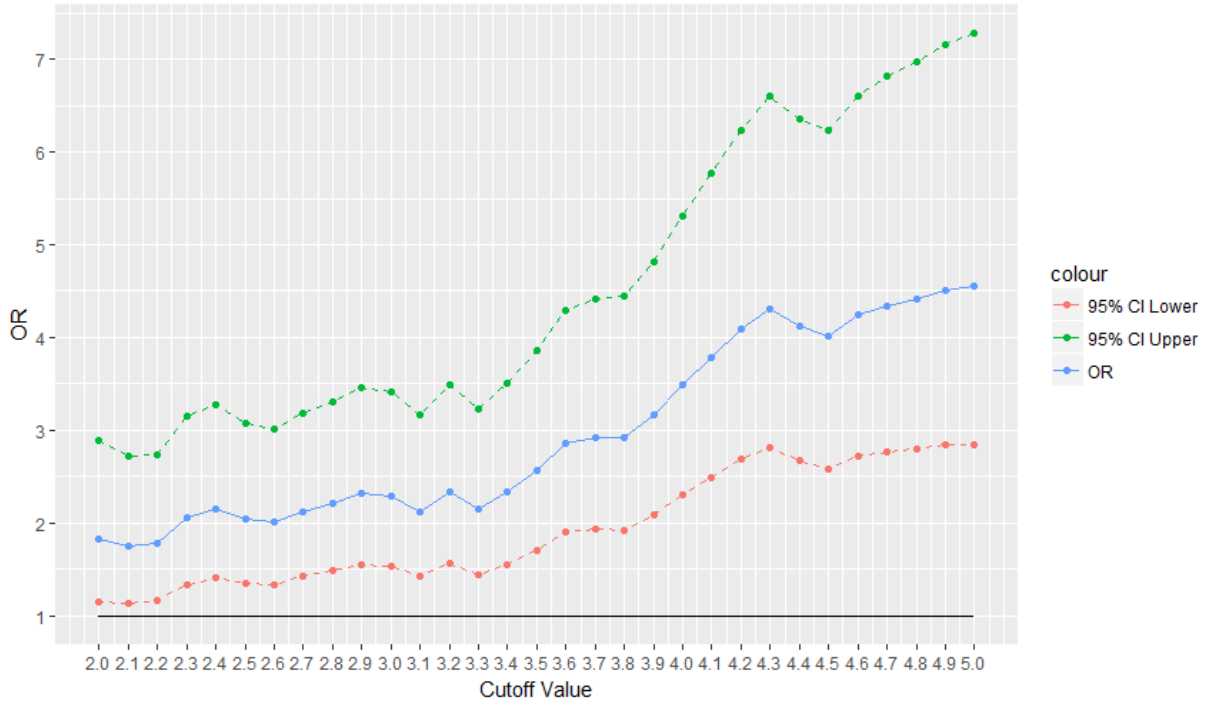
Table 3-5 Odds Ratios for Two Treatments

Treatment	Predictive Effect	Causal Effect
HUSV	2.23 (1.48, 3.33)*	0.99 (0.61, 1.61)
HDSV	2.01 (1.33, 3.10)	1.40 (0.84, 2.33)

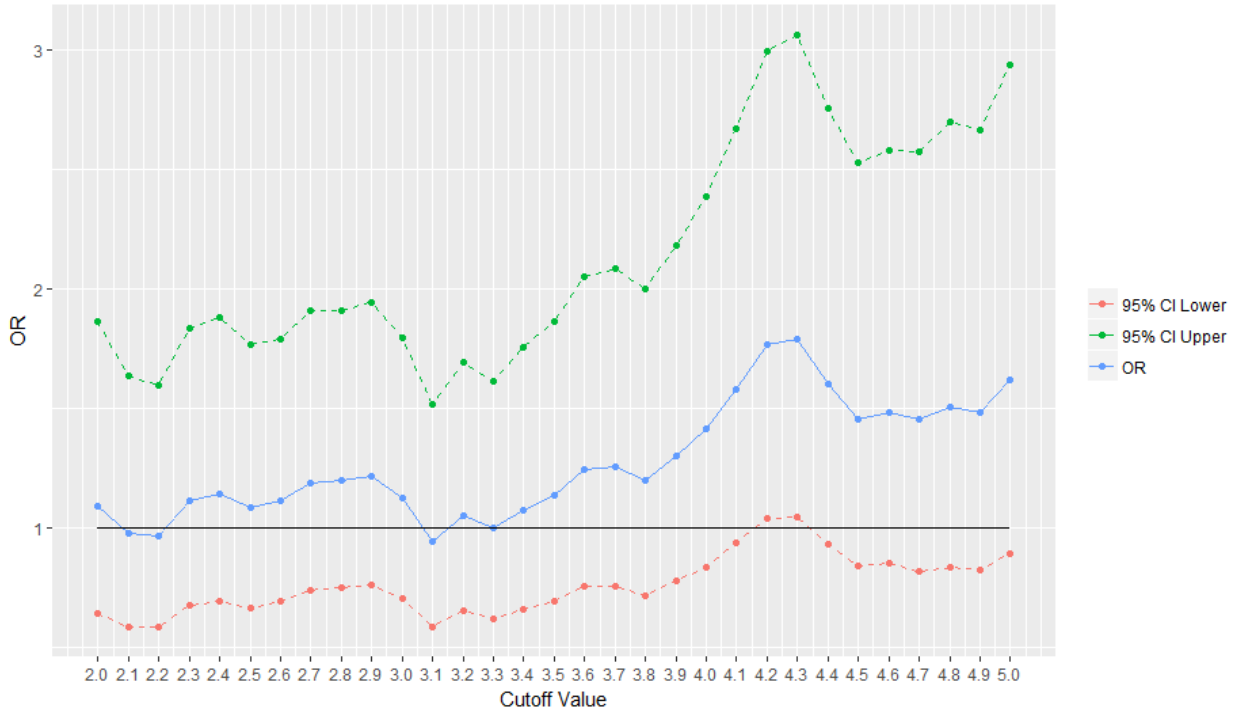
* 95% confidence interval of odds ratio is in parentheses.

The sensitivity analysis results are shown in Figure 3-3 and Figure 3-4. For both treatments, the causal effects are almost consistently insignificant while the predictive effects are

consistently and spuriously significant. Moreover, their causal effects are very consistent regarding the width of the 95% confidence interval. In conclusion, neither the upstream nor downstream speed variations have significant causal effect on the crash risk.

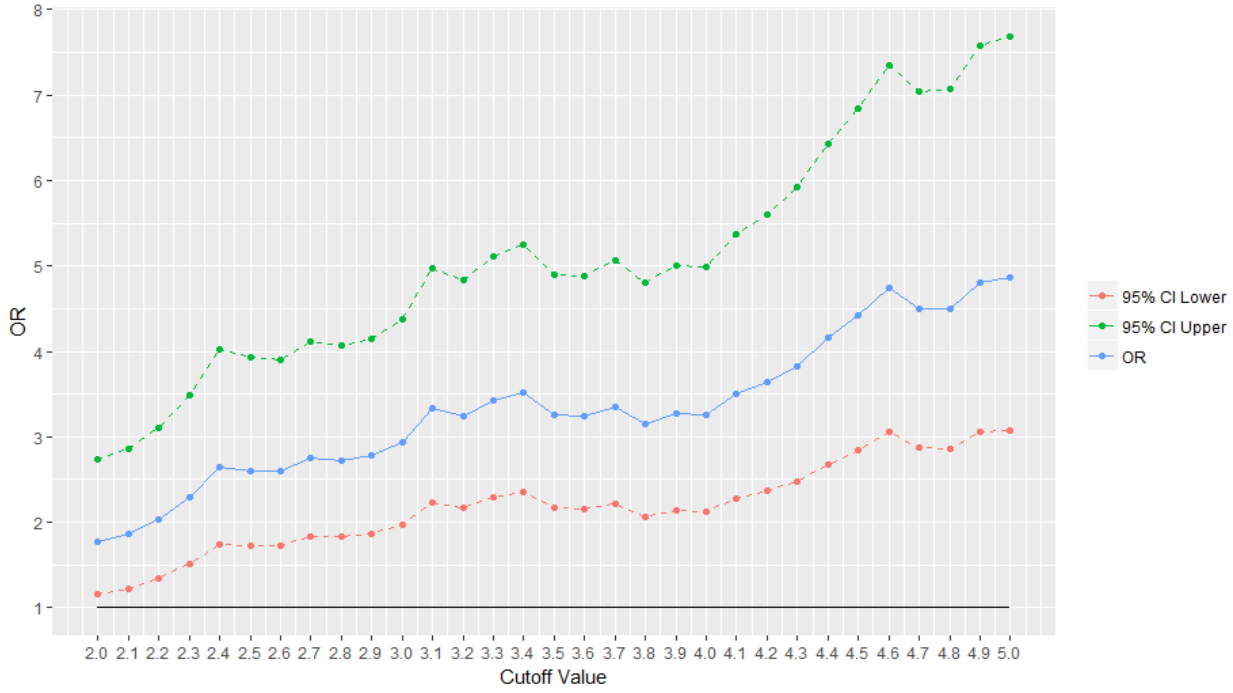


(a) unadjusted sample

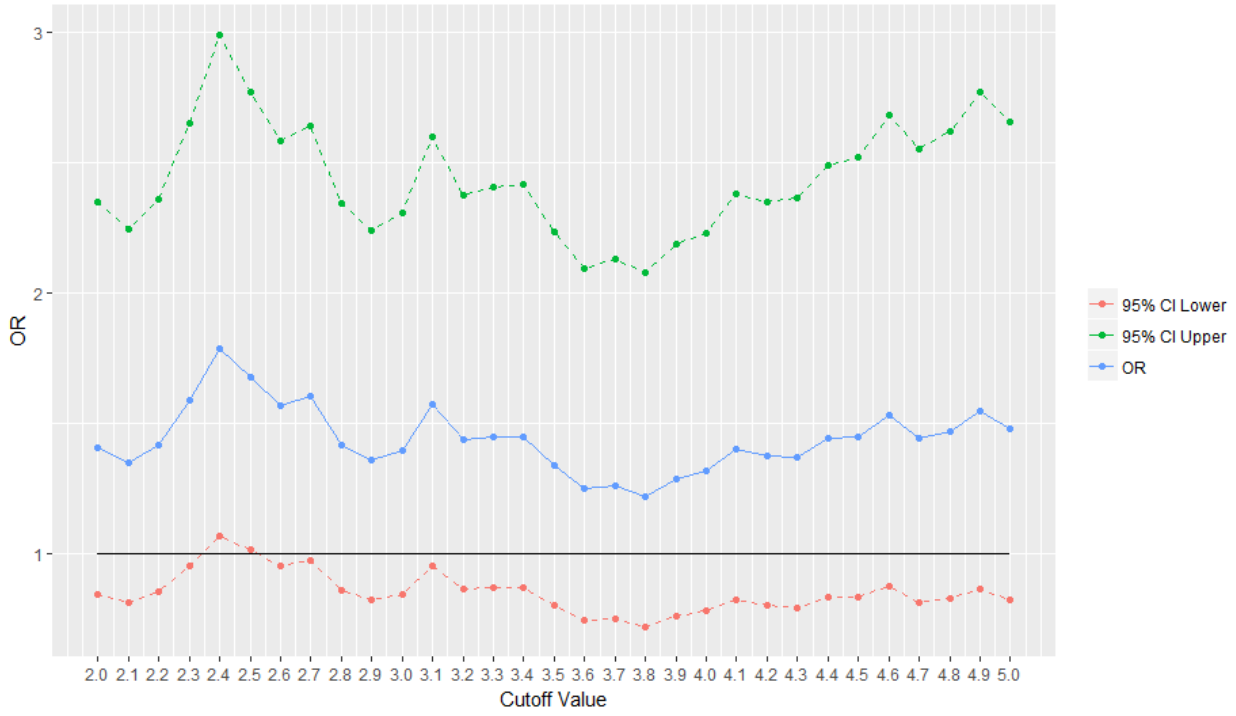


(b) weighted sample

Figure 3-3 Sensitivity analysis of cutoff values for HUSV



(a) unadjusted sample



(b) weighted sample

Figure 3-4 Sensitivity analysis of cutoff values for HUSV.

3.6 Conclusions

In this chapter, the propensity score based method was used to assess the effect of the upstream and downstream speed variation on crash occurrence. The analysis was carried out using the crash data and ILD data on a 4.15-mile corridor on I-94 East in WI where both crash and non-crash associated traffic data were collected. Speed variation was converted into a binary variable (i.e., high/low speed variation) based on cutoff values. The propensity score method found that neither HDSV nor HDSV variable has statistically causal effect on crash occurrence while the predictive method found that both HUSV and HDSV variables are statistically significant. Although it is difficult to prove which conclusion is correct, the propensity score based method is considered superior because of a more rigorous study design.

In a propensity score method, the propensity score for each case was estimated based on the propensity model, and inverse probability of treatment weighting (IPTW) method was applied to generate a weighted sample. In the weighted sample, covariates other than HUSV and HDSV have similar distributions across treated and control groups. Therefore, the causal effect of HUSV and HDSV between the treated and control group can be impartially estimated.

Sensitivity analysis on the cutoff value of speed variation has been performed to test the consistency of the findings. The finding holds with varying cutoff values. Hence, it is concluded with high confidence that speed variation is not one of the causes for a crash. The finding in this chapter demonstrates the necessity of a propensity score based model because it is able to obtain the causal effect of a contributing factor with more accuracy, while the predictive analysis may yield biased or inconsistent effects.

3.7 References

Abdel-Aty, M., & Pande, A. (2005). Identifying crash propensity using specific traffic speed conditions. *Journal of safety research, 36*(1), 97-108. doi:10.1016/j.jsr.2004.11.002

Abdel-Aty, M., & Pemmanaboina, R. (2006). Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. *IEEE Transactions on Intelligent Transportation Systems, 7*(2), 167-174. doi:10.1109/TITS.2006.874710

Abdel-Aty, M., Uddin, N., & Pande, A. (2005). Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1908), 51-58.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*(1897), 88-95.

Aul, N., & Davis, G. (2006). Use of propensity score matching method and hybrid Bayesian method to estimate crash modification factors of signal installation. *Transportation Research Record: Journal of the Transportation Research Board*(1950), 17-23.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity‐score matched samples. *Statistics in medicine, 28*(25), 3083-3107.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine, 34*(28), 3661-3679.

Braitman, L. E., & Rosenbaum, P. R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of internal medicine, 137*(8), 693-695.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Nat. Cancer Inst, 22*, 173-203.

Davis, G. A. (2000). Accident reduction factors and causal inference in traffic safety studies: a review. *Accident Analysis & Prevention, 32*(1), 95-109.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 1231-1236.

Durbin, D. R., Elliott, M. R., & Winston, F. K. (2009). A propensity score approach to estimating child restraint effectiveness in preventing mortality. *Statistics and Its Interface, 2*(4), 437-447.

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

Hirunyanitiwattana, W., & Mattingly, S. P. (2006). *Identifying secondary crash characteristics for California highway system.* Paper presented at the Transportation Research Board 85th Annual Meeting.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician, 58*(4), 272-279.

Kockelman, K. K., & Ma, J. (2010). *Freeway speeds and speed variations preceding crashes, within and across lanes.* Paper presented at the Journal of the Transportation Research Forum.

Li, H., & Graham, D. J. (2016). Quantifying the causal effects of 20 mph zones on road casualties in London via doubly robust estimation. *Accident Analysis & Prevention, 93*, 65-74.

Niven, D. J., Berthiaume, L. R., Fick, G. H., & Laupland, K. B. (2012). Matched case-control studies: a review of reported statistical methodology. *Clinical Epidemiology, 4*, 99-110. doi:10.2147/CLEP.S30816

Pande, A., & Abdel-Aty, M. (2006a). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention, 38*(5), 936-948. doi:10.1016/j.aap.2006.03.004

Pande, A., & Abdel-Aty, M. (2006b). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 31-40.

Pande, A., Abdel-Aty, M., Hsia, L., & Trb. (2005). Spatiotemporal variation of risk preceding crashes on freeways *Statistical Methods; Highway Safety Data, Analysis, and Evaluation; Occupant Protection; Systematic Reviews and Meta-Analysis* (pp. 26-36).

Parker, S. T., & Tao, Y. (2006). *WisTransPortal: A Wisconsin Traffic Operations Data Hub.* Paper presented at the 9th International Conference on Applications of Advanced Technology in Transportation, Chicago, Ill.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology, 49*(12), 1373-1379.

Rosenbaum, P. R. (2002). Observational studies *Observational studies* (pp. 1-17): Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbuam, P. (2010). Design of observational studies: Springer Series in Statistics.

Roshandel, S., Zheng, Z., & Washington, S. (2015). Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis & Prevention, 79*, 198-211. doi:10.1016/j.aap.2015.03.013

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine, 127*(8_Part_2), 757-763.

Sasidharan, L., & Donnell, E. T. (2013). Application of propensity scores and potential outcomes to estimate effectiveness of traffic safety countermeasures: Exploratory analysis using intersection lighting data. *Accident Analysis & Prevention, 50*, 539-553.

Wood, J., & Donnell, E. T. (2016). Safety evaluation of continuous green T intersections: A propensity scores-genetic matching-potential outcomes approach. *Accident Analysis & Prevention, 93*, 1-13.

Xu, C., Liu, P., & Wang, W. (2016). Evaluation of the predictability of real-time crash risk models. *Accident Analysis & Prevention, 94*, 207-215. doi:10.1016/j.aap.2016.06.004

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

Xu, C., Liu, P., Wang, W., & Li, Z. (2014). Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. *Transportation Research Part A: Policy and Practice, 69*, 58-70. doi:10.1016/j.tra.2014.08.011

Xu, C., Wang, W., & Liu, P. (2013). A genetic programming model for real-time crash prediction on freeways. *IEEE Transactions on Intelligent Transportation Systems, 14*(2), 574-586.

Zheng, Z., Ahn, S., & Monsere, C. M. (2010). Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention, 42*(2), 626-636. doi:10.1016/j.aap.2009.10.009

CHAPTER 4 PREDICTIVE ANALYSIS OF CRASH-PRONE CONDITIONS OF LANE-CHANGE RELATED CRASHES

4.1 Background

The advances and wide deployment of detection technologies have greatly increased the transportation agencies' ability to collect, store, and process large-scale traffic data in a real-tie fashion. The enhanced surveillance has empowered traffic engineers to constantly monitor travel conditions and instantly detect accidents, aiding in the comprehensive evaluation of highway traffic and safety performance. ILDs, which are embedded in the pavement, are commonly used in the area of detection technology and are a key component of the freeway management and operations systems. Loop detector data have the potential to be used for traffic crash prediction and prevention. For instance, engineers can identify potentially hazardous situations that are prone to crashes by investigating the relationship between crashes and traffic flow characteristics retrieved from detector data. Subsequently, preventive countermeasures can be developed and implemented to proactively address imminent safety concerns.

Rear-end and sideswipe crashes are the most common types of crashes on the freeway. Research on sideswipe crashes is rather limited when compared with the amount of studies on rear-end collisions crashes (Li et al., 2014; Oh, Park, & Ritchie, 2006; Pande & Abdel-Aty, 2006b; Pande & Abdel‐Aty, 2008; Qu, Wang, Wang, Liu, & Noyce, 2012); however, sideswipe crashes are more prevalent. Wang and Knipling reported that among all lane change/merge crashes in United Sates, rear-end crashes only make up 4.5% (Wang & Knipling, 1994). Sideswipe crashes are often caused by the driver's poor gap judgment and failed gap acceptance after initiating a lane change maneuver. Sideswipe crashes occur when one vehicle changes or merges into the other vehicle's lane and sideswipes (or is sideswiped by) the other vehicle, or

when both vehicles change lanes and collide. Researchers gain more information on how different variables (i.e. changes in traffic flow, speed, and density) may affect the odds of a crash when the traffic parameters are related to specific crash types (i.e. rear-end, same direction sideswipe, etc.), as opposed to the total number of crashes. Due to the prevalence of and lack of research with regard to sideswipe crashes, the focus of this paper is on sideswipe crashes related to lane changes.

Researchers filtered lane-change related crashes from all other crashes occurring in 2012 and 2013 on a 62-mile stretch of freeway from I-94 to I-43 in Southeast Wisconsin. A matched case-control study was adopted to mitigate the nuisance of non-traffic flow parameters such as time of day and geometric design elements. The real-time traffic data for crashes and non-crash events were extracted for the two lanes associated with lane change crashes. Statistical models were developed using conditional logistic methodology. The effects of traffic flow factors on lane-change related crashes were thoroughly investigated and discussed in a lane-specific manner.

4.2 Literature Review

A number of studies have addressed the topic of how drivers' decisions relate to lane changes. Gibbs proposed a hierarchical structure of the decision process to model lane changes on urban highways (Gipps, 1986), concluding that the driver's gap acceptance and vehicle speed influenced the driver's decision to change lanes. Equipped with an instrumental vehicle, Brackstone et al. investigated driver behavior at a more detailed level (Brackstone, McDonald, & Wu, 1998) and added several microscopic parameters to the list of factors affecting drivers' lane change decisions, such as the size of the available gap perceived by the driver, the distance to the preceding vehicle on the same lane, the relative speed, etc. Hill and Elefteriadou collected data

on drivers' desired speed, lane change duration (time lapse between the initiation and completion of lane change), and gap acceptance (Hill & Elefteriadou, 2013), finding that during congested conditions, lane change duration is longer and drivers tend to accept smaller lag gaps (the gap between the vehicle intending to change lanes and its potential following vehicle in the target lane). In a study performed with Next Generation Simulation (NGSIM) traffic data, relative velocity between lanes and relative lead gap (the gap between the vehicle intending to change lanes and its potential preceding vehicle in the target lane) had a positive effect on lane change probability, indicating that the driver tends to change into lanes that are moving faster or have larger lead gaps (J. Lee, Park, & Yeo, 2013).

Patterns in driver lane-change behavior tempted researchers to employ real-time traffic data in the exploration of the factors behind a lane-change related crash. Average Flow Ratio (AFR), the ratio of flows from one lane to its adjacent lane(s), was first proposed by Chang and Gao (Chang & Kao, 1991) who found it to be significantly related to lane change intensity. Lee et al. adopted the idea of AFR to compare the contributing factors to rear-end and sideswipe crashes (C. Lee, Abdel-Aty, & Hsia, 2006). The authors modified the algorithm to calculate the AFR variable and computed Overall Average Flow Ratio (OAFR) as the geometric mean of the modified AFRs of all lanes. OAFR was intended to serve as a surrogate measure of the lane change frequency on all lanes. The findings showed that OAFR contributed more to sideswipe crashes than to rear-end crashes, and that sideswipe crashes were more likely to occur in uncongested traffic conditions. Pande and Abdel-Aty adopted the modified AFRs and the corresponding OAFR proposed by Lee et al. to measure the lane change frequency, and found that OAFR was not significant when modeling crashes related to lane changes (C. Lee et al., 2006; Pande & Abdel-Aty, 2006a).

Lee et al. also estimated the likelihood of a lane change-related crash occurring in the center lane compared to the left or right lane, respectively (C. Lee, En, Young-Jin, & Abdel-Aty, 2009). The study included real-time traffic flow variables and dichotomy road geometric variables, which indicate whether the segment has a curve and whether the segment is close to ramps. Authors used both lane-specific and lane average variables to build models for variables related to traffic conditions. Results showed that flow-related variables rather than speed or occupancy-related variables were significant, and that lane-specific variables (as opposed to lane average variables) more clearly explained why a crash most likely happened in one lane as opposed to another.

The relationship between historical crash data and the recorded traffic flow data from loop detectors near crash locations can be examined to build prediction models that aid in the development of proactive accident prevention approaches. Several studies have attempted to develop prediction models with real-time data using various statistical methodologies and data mining techniques such as the matched case-control logistic model, log-linear model, neural network, support vector machine (SVM), and stochastic gradient boosting (M. Abdel-Aty, Pande, Lee, Gayah, & Santos, 2007; M. Abdel-Aty, Uddin, Pande, Abdalla, & Hsia, 2004; M. Ahmed & Abdel-Aty, 2013; C. Lee, Saccomanno, & Hellinga, 2002; Yu & Abdel-Aty, 2013). The matched case-control study, an analytical approach commonly used in epidemiological research, was introduced in a previous study to increase the accuracy of the predication model (M. Abdel-Aty et al., 2004). The study used key traffic flow information for crash and non-crash events while controlling for other external factors such as location, time of day, day of week, etc.

The rapid development of vehicle tracking and re-identification technologies as well as prolific use of probe vehicle data has allowed researchers to develop crash prediction models

using new data sources. Automatic Vehicle Identification (AVI) sensors collect travel time from each AVI segment and can measure the space mean speed (SMS) rather than time mean speed (TMS) which is measured from conventional inductive loops detectors. AVI sensors are nonintrusive detection devices which are easier to install and maintain than intrusive loop detectors. AVI sensors also have enhanced reliability compared to loop detectors because of their easy maintenance. Loop detectors can fail due to hard pavement conditions and temperature variations. Additionally, maintenance can be delayed by congested roadways as repairs usually involve cutting into the pavement.

Ahmed and Abdel-Aty were the first to utilize AVI data to predict real-time crashes (M. M. Ahmed & Abdel-Aty, 2012). The high classification accuracy of their model proves that AVI data is promising in predicting crashes on expressways. But the authors found that loop detector data, as opposed to AVI data, is better for predicting visibility-related crashes (M. A. Abdel-Aty, Hassan, & Ahmed, 2012). The authors attributed the loop detector success to the closer spacing between detectors (AVI segments were spaced farther apart). Additionally, AVI sensors record only the vehicles with AVI tags, and only about 80% of vehicles using expressways in the study area had AVI tags. Loop detector systems can record data for all vehicles. Many other traffic safety studies implemented AVI data (M. A. Abdel-Aty et al., 2012; M. Ahmed & Abdel-Aty, 2013; M. Ahmed, Abdel-Aty, & Yu, 2012; M. M. Ahmed & Abdel-Aty, 2012; Yu & Abdel-Aty, 2014; Yu, Abdel-Aty, Ahmed, & Wang, 2014), but these studies used only the speed data recorded by AVI sensors. Real-time traffic factors including the flow rate and occupancy have not been utilized, possibly because of the failure of AVI sensors to capture all traversed vehicles. Therefore, if crash type is related to traffic flow, prediction models that use only AVI speed data may generate biased results.

Visual surveillance has been widely used in many areas. Vision-based traffic monitoring systems can track individual vehicles and capture the crash process, while loop detectors can only record aggregated vehicle behaviors (e.g. flow rate and density). Loop detectors also fail to identify the crash process. Video recordings of crashes have also been used to identify causal factors of crashes. Davis and Swenson used video recordings to study the causal factors of three rear-end crashes (Davis & Swenson, 2006). Trajectory information for a platoon of vehicles was obtained from the recordings, and was used to estimate driver information including initial speed, following distance, reaction time, and braking rate. The reaction time of at least one driver ahead of the colliding vehicle was found to be longer than the following time headway of the driver in each crash. Chatterjee and Davis demonstrated this finding in a later study that used video recordings of 41 freeway shock waves, five of which resulted in rear-end crashes and ten of which resulted in swerving behavior (Chatterjee & Davis, 2016). However, no studies have been found to use video data to develop crash prediction models. One possible explanation is that unlike loop detectors, visual surveillance has not recorded a large enough sample of crashes, as crashes are rare events.

4.3 Methodology

The effects of traffic parameters and weather conditions would be more accurate if external variables were controlled. The amount of traffic, environmental conditions, and geometric characteristics all contribute to the occurrence of lane-change related crashes. A matched case-control study can concentrate on specific variables while also controlling for nuisance factors. Each crash should be considered a separate case in order to construct an eligible design for this methodology. Several non-crash events were chosen as controls along with their corresponding non-traffic-flow variables (location, time, season, etc.) matching with those of crash cases; this

way, each case and its corresponding controls constitute a stratum. The controlled, non-traffic-flow variables are the same within each stratum, but are different across strata.

In a matched case-control logistic regression, the crash probability could be expressed as:

$$\text{logit}(p_{ijk}) = \propto_i + \sum_{k=1}^{K} \beta_k x_{ijk} \quad (4\text{-}1)$$

where x_{ijk} is the k^{th} traffic flow variable for the case ($j = 0$) or the j^{th} control in the i^{th} stratum; $i = 1, 2, \ldots, N$; $j = 0, 1, \ldots, J$; and $k = 1, 2, \ldots, K$. N is the number of strata, J denotes the number of controls, and K represents the number of explanatory variables.

The number of parameters increases as more strata are added, as each stratum has its corresponding intercept parameter. It is against the optimality properties of the maximum likelihood method which becomes minimum variance unbiased estimator as the sample size increases while keeping the number of parameters fixed. However, if the stratum-specific parameters, $\propto_i$, are considered to be nuisance parameters, conditional likelihood could be created to yield maximum likelihood estimators which would be expressed as (Hosmer Jr & Lemeshow, 2004):

$$\boldsymbol{l_i(\beta)} = \frac{\exp(\sum_{k=1}^{K} \boldsymbol{\beta_k x_{i0k}})}{\sum_{j=0}^{J} \exp(\sum_{k=1}^{K} \boldsymbol{\beta_k x_{ijk}})} \quad (4\text{-}2)$$

And the full conditional likelihood is the product of the $l_i(\beta)$ over N strata,

$$l(\beta) = \prod_{i=1}^{N} l_i(\beta) \quad (4\text{-}3)$$

Full conditional likelihood is independent of stratum-specific parameters, $\propto_i$, and thus cannot be used to estimate those stratum-specific parameters. Hence, the crash probability in a specific case cannot be estimated by using Equation 4-1. But the slope coefficients β can be estimated by Equation 4-3, and can be used to evaluate the effect of each variable.

4.4 Data Description

All data were collected from WisTransPortal, a comprehensive data depository system which provides a central source of traffic operations, safety, and intelligent transportation systems (ITS) data regarding Wisconsin highways. The corridor used in this study is a continuous corridor stretching from I-94 to I-43 in Southeast Wisconsin, highlighted by black lines in Figure 4-1.

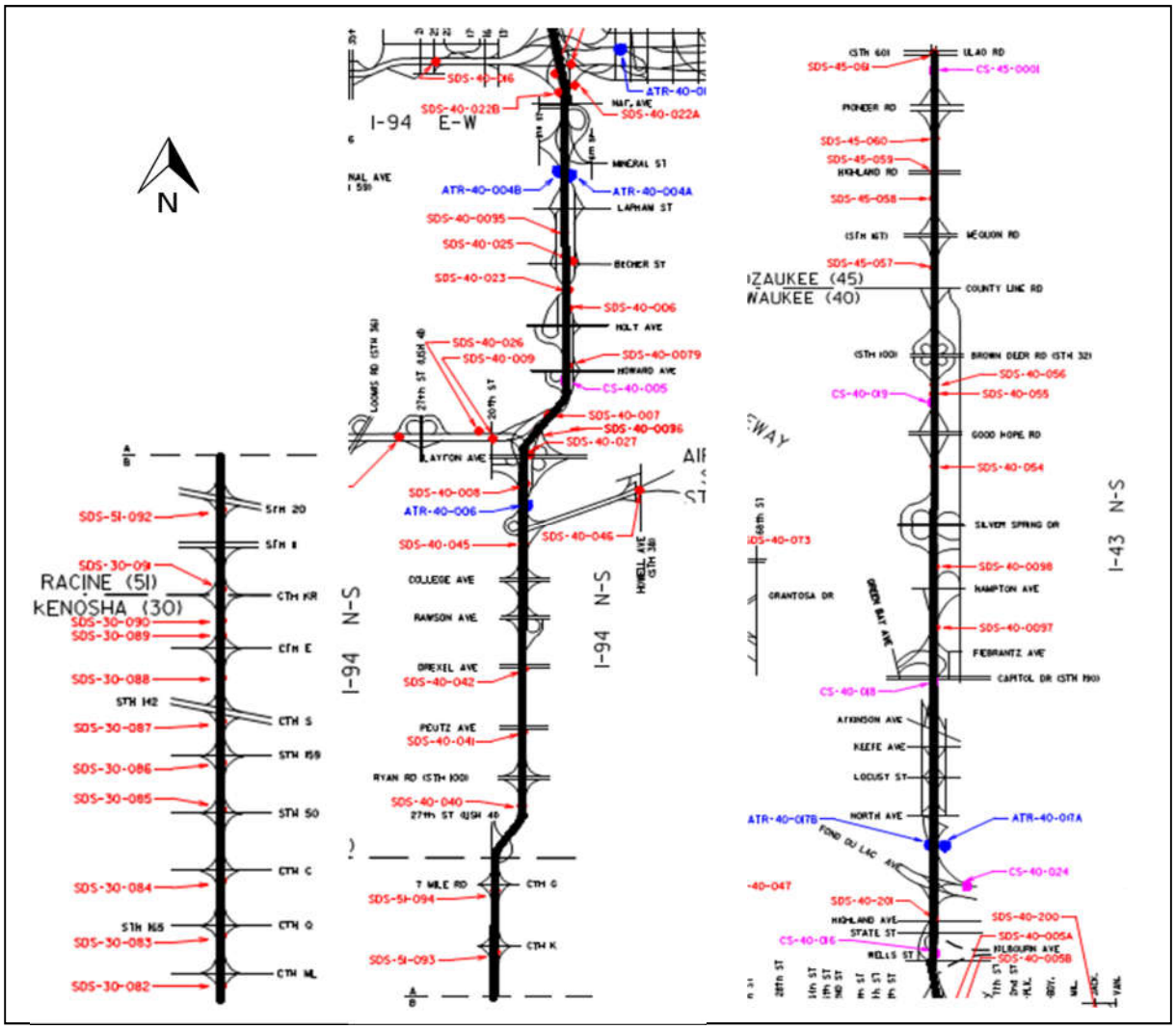


Figure 4-1 Freeway I-94 N-S and I-43 N-S. (One long continuous segment is divided into three short ones for clear layout.)

The 62-mile long corridor has 27 loop detector stations in the northbound lanes and 26 in the southbound lanes. The space between detector stations ranges from 1 to 6.5 miles with a median of 1.5 miles. 996 crashes were reported on this corridor from 2012 to 2013. Real-time traffic data, including average volume (volume per lane), speed (mile per hour), and occupancy (percent of time the loop is occupied) were collected from dual loop detectors in 1-minute, 5-minute, 15-minute, and 60-minute intervals. Finer time intervals can reflect more accurate traffic condition information; thus, 1-minute detector data was used in this chapter.

Sideswipe same direction, sideswipe opposite direction, and angle crashes were removed from the total crash dataset in order to obtain the number of crashes related to lane changes. The total number of these crashes was 369. Sideswipe opposite direction crashes were removed because there is no direct conflict between two opposite directions on a freeway; therefore, these crashes may be wrongly recorded. Angle crashes were also kept for further exploration, as a previous study (Pande & Abdel-Aty, 2006a) indicated they may be caused by lane change activities. Abnormal crashes such as intersection-related crashes and crashes involving one vehicle were discarded. A total of 310 crashes remained after the aforementioned incidents were removed.

Original police accident reports of these crashes were carefully reviewed to remove ramps or construction zone-related crash events based on the crash location description. A crash was considered to be lane-change related if "changing lanes" was recorded as the driver activity. The crash reports offered more accurate crash occurrence times than the crash database, which rounded the time to the nearest five-minute interval. The specific lanes that were involved in lane change maneuvers were obtained from the crash diagram and narrative.

As Figure 4-2 shows, the initial lane just before lane change is referred to as the "subject lane" and the lane the vehicle is moving into is called the "target lane". The subject and target lanes were designated by reviewing the crash diagram for each crash.

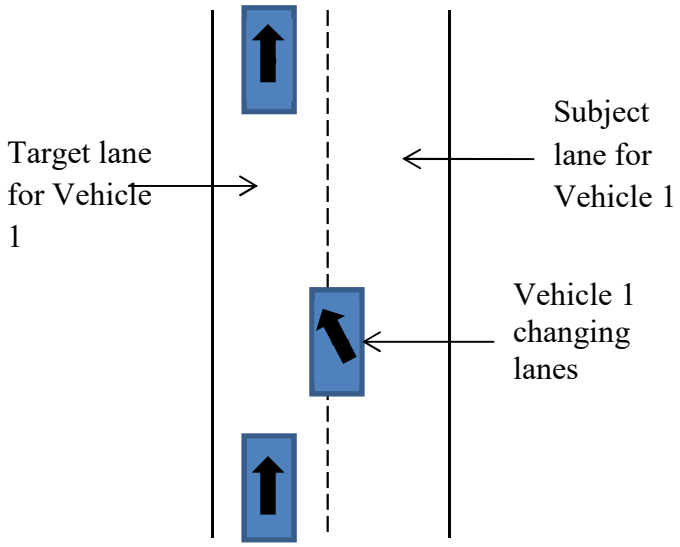


Figure 4-2 Illustration for subject lane and target lane.

After collecting the crash data, the loop detector data for each crash was extracted. The upstream station nearest to the crash location and the two downstream stations nearest to the crash location were chosen as the three detector stations from which to gather real-time traffic data. Figure 4-3 illustrates the layout of three stations. Previous studies have compared traffic data in several time periods prior to a crash, concluding that variables occurring within 5 to 10 minutes of the crash are the most effective variables in modeling lane-change related crashes (C. Lee et al., 2006; C. Lee et al., 2009; Pande & Abdel-Aty, 2006a). Therefore, this study used the traffic data in the time period 5 to 10 minutes before crash occurrence. A crash happening at 1:00 p.m. would use loop detector data from 12:50 p.m. to 12:55 p.m. Each crash was associated with the three detector station locations (nearest upstream and two nearest downstream) and the

crash location, and thus there are three gaps between consecutive locations (see Figure 4-3). Any gap that was found to be longer than 5 miles was excluded during data collection for the sake of consistence, as most gaps were around 2 miles. Blank data was a possibility due to detector dysfunction. If blank data was found in the traffic data collected from one station, all traffic data from that station was excluded. The number of crashes with traffic data available at one station or more was 108.

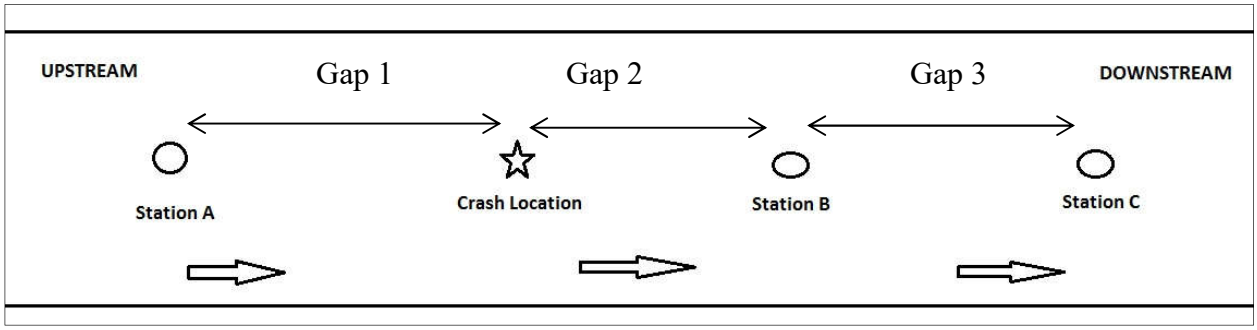


Figure 4-3 Detector stations.

Data from non-crash events also was gathered to conduct a matched case-control analysis. The same time of day and day of the week were adopted for selecting non-crash events in order to ensure similar general traffic patterns, such as commuter types, trip purposes, etc. In order to avoid effects of seasonal changes in weather and daylight, only four non-crash events with the corresponding time and day were chosen. Two of these events were selected from the time period before the crash, while the other two took place after the crash. For example, if a crash happened on Monday, June 4, 2012 at 1:00 p.m., the non-crash event data were taken from 12:50 p.m. to 12:55 p.m. on the following Mondays in 2012: May 21, May 28, June 11, and June 18. It was assumed that the non-crash event location was the same as the crash location; therefore, the same set of detector stations was chosen for data collection.

Loop detectors can encounter random equipment and hardware problems that result in erroneous data. All inaccurate data were eliminated from the raw 1-minute data in this study. The inaccurate traffic data include: 1) occupancy < 0 or > 100; 2) speed < 0 or > 100; 3) volume < 0 or > 50 in 1 min; 4) volume > 0 with speed = 0 or speed > 0 with volume = 0 (Al-Deek, Venkata, & Ravi Chandra, 2004). 63 crash events and 192 non-crash events remained, with detector data available at all three stations.

Weather information was collected from Weather Underground Inc., which archives historical weather information from airports. Four airports existed along the study segments of I-94 to I-43, and each event utilized the nearest airport's historical weather data. Observations from the time stamps before and after the event were collected since weather data are recorded hourly. For instance, if a crash occurred at 1:00 p.m., the weather information at 12:46 p.m. and 1:46 p.m. would be collected. The weather conditions observed from the detector stations were reclassified as clear, cloudy, rainy, and snowy in order to be consistent with the weather types in available crash data. If a conflict existed between the weather conditions recorded at two time stamps, the earlier weather condition was used. In the dataset, only 4% of non-crash events had conflicting weather records. The weather information collected from the nearest airports was compared with the weather information from crash reports for each crash event.

Table 4-1 shows the classification results of weather information from both sources. The two weather sources show consistency for a total of 60 out of 63 cases, indicating a 95% classification rate. Only 3 crashes had conflicting weather records. Therefore, weather information retrieved from the nearest airports was considered reliable and was used as the weather information source for non-crash events. Weather conditions recorded in crash reports were used as the source for crash events. The distribution of different weather conditions is

shown in Table 4-2. It was observed that 21% (13/63) of the crashes occurred on snowy days, while only 2% (4/192) of non-crashes occurred on snowy days; this indicates that snowy weather may influence an increase in lane-change crashes. This hypothesis was investigated in the analysis.

Table 4-1 Classification of Weather Information from Crash Reports and Nearest Airports

		Crash Reports			
		Normal	**Rainy**	**Snowy**	**Total**
Nearest Airports	**Normal**	45	0	0	45
	Rainy	2*	2	0	4
	Snowy	1*	0	13	14
	Total	48	2	13	63

* It represents the number of crashes with conflicting weather records.

Table 4-2 Distribution of Weather Factor

	Crash Status	
Weather Condition	1	0
Normal	48 (76%)	180 (94%)
Rainy	2 (3%)	8 (4%)
Snowy	13 (21%)	4 (2%)
Total	63 (100%)	192 (100%)

4.5 Analysis and Discussion

Crashes involving lane-change maneuvers occurred when a vehicle attempted to change from the subject lane into the target lane when the gap was not large enough for a safe merge. Lane-specific traffic characteristics directly reflect the circumstances when a driver initiates and performs a lane change movement. These traffic characteristics can be represented by measurable traffic-related parameters such as traffic flow, speed, and density or occupancy. Given the 1-minute interval inductive dual-loop detector data from three stations near each crash location, means and standard deviations of flow, speed, and occupancy were calculated for all crash and non-crash events at 5 to 10 minutes prior to the crash. The calculations were performed for three loop detector stations, the station located immediately upstream from the crash, and the two stations immediately downstream.

The coefficient of variation, a unitless measure calculated as 100*standard deviation / mean, was generated for all three elements to show a relative standard deviation. The information for each detector location is formulated by four letters (see Figure 4-4). For example, ASAF stands for the average flow for the subject lane at the immediate upstream detector. Only the crashes with available data at all three stations were analyzed. In total, there were 63 crash events and 192 non-crash events.

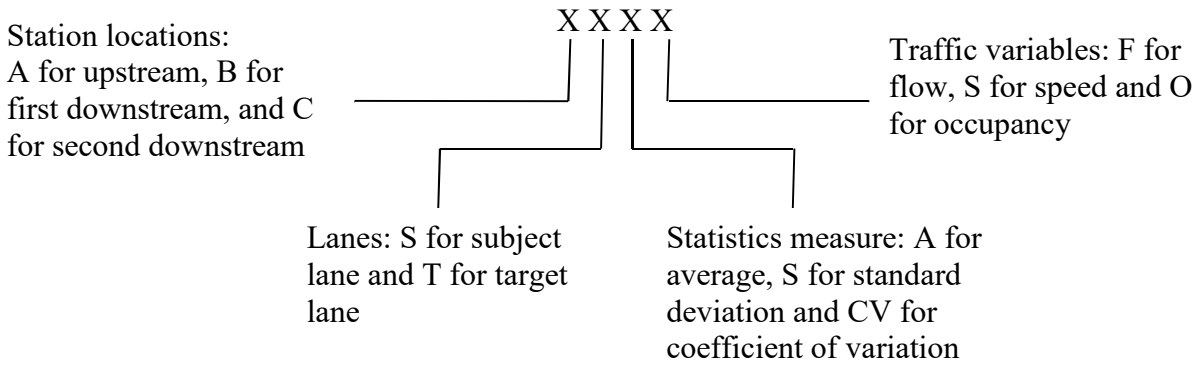


Figure 4-4 Nomenclature method for traffic-related variables.

It is anticipated that between-lane traffic differentials are critical to lane change decisions. In Gipps' theory, drivers adopt discretionary lane changes to gain speed advantage (Gipps, 1986). Higher speed in the adjacent lane and available gaps provide opportunities for drivers to change lanes. The ratios of flow, speed, and occupancy are calculated for the three detector stations to capture the between-lane traffic-related variable differences by dividing the average values of the subject lane by the average values of the target lane at the same detector location. Similar to the lane-specific traffic characteristics, between-lane traffic differences are formulated by three letters with the first letter representing detector location, the second letter representing traffic variable, and the last letter "R" representing the ratio. For example, the ratio of flow at station A, named AFR, is equal to ASAF/ATAF. Weather was used as the non-traffic variable, and four weather categories were identified: clear, cloudy, rainy and snowy. In total, 64 explanatory variables existed (see Table 4-3).

Table 4-3 List of Explanatory Variables

Variables	Description
ASAF, BSAF, CSAF	Average flow in the subject lane at station A, B and C, respectively
ATAF, BTAF, CTAF	Average flow in the target lane at station A, B and C, respectively
ASAS, BSAS, CSAS	Average speed in the subject lane at station A, B and C, respectively
ATAS, BTAS, CTAS	Average speed in the target lane at station A, B and C, respectively
ASAO, BSAO, CSAO	Average occupancy in the subject lane at station A, B and C, respectively
ATAO, BTAO, CTAO	Average occupancy in the target lane at station A, B and C, respectively
ASSV, BSSV, CSSV	Standard deviation of flow in the subject lane at station A, B and C, respectively
ATSV, BTSV, CTSV	Standard deviation of flow in the target lane at station A, B and C, respectively
ASSS, BSSS, CSSS	Standard deviation of speed in the subject lane at station A, B and C, respectively
ATSS, BTSS, CTSS	Standard deviation of speed in the target lane at station A, B and C, respectively
ASSO, BSSO, CSSO	Standard deviation of occupancy in the subject lane at station A, B and C, respectively
ATSO, BTSO, CTSO	Standard deviation of occupancy in the target lane at station A, B and C, respectively

ASCVF, BSCVF, CSCVF	Coefficient of variation of flow in the subject lane at station A, B and C, respectively
ATCVF, BTCVF, CTCVF	Coefficient of variation of flow in the target lane at station A, B and C, respectively
ASCVS, BSCVS, CSCVS	Coefficient of variation of speed in the subject lane at station A, B and C, respectively
ATCVS, BTCVS, CTCVS	Coefficient of variation of speed in the target lane at station A, B and C, respectively
ASCVO, BSCVO, CSCVO	Coefficient of variation of occupancy in the subject lane at station A, B and C, respectively
ATCVO, BTCVO, CTCVO	Coefficient of variation of occupancy in the target lane at station A, B and C respectively
AFR, BFR, CFR	Flow ratio at station A, B and C, respectively
ASR, BSR, CSR	Speed ratio at station A, B and C, respectively
AOR, BOR, COR	Occupancy ratio at station A, B and C, respectively
Weather	The weather condition at the crash location

A matched case-control design was used to collect the data with the purpose of eliminating the effects of non-traffic flow variables such as geometric design, pavement condition, etc. However, collecting data with this design makes it impossible to calculate the estimates of coefficients using traditional logistic regression analysis. Therefore, conditional logistic regression was adopted to investigate the effects of traffic-related parameters.

Proportional hazard regression analysis in SAS was used to calculate the hazard ratios of variables.

The hazard ratio of each variable describes the ratio of change in the probability of a crash with one-unit change in that variable. A hazard ratio greater than 1 means that the chance of a crash increases as the value of that variable increases. The non-crash events happening at the same location and same time on the same day of the week were chosen for data collection. Thus, each crash and its corresponding non-crash events form one stratum, and the whole dataset are stratified based on the location, time and date of each crash. This stratification feature suggests that it may be questionable to apply popular variable selection methods (e.g. decision tree, random forest) to narrow down the number of variables, as they may not be able to handle stratified data.

A conditional logistic regression was run with one independent variable at a time for all 64 variables in order to screen for statistically significant variables. The preliminary results showed that BSAF, BTAF, BTSO, CTAF, CTSS, and CVR were statistically significant at a 5% confidence level. "Snowy" was significant at 5% confidence level while "rainy" was not when "normal" was the reference level. After combing rainy and normal conditions, the weather factor became the snow indicator, a binary variable, which was found to be statistically significant. The six significant traffic variables along with the snow indicator were reviewed further for correlation. Table 4-4 shows three different sets of variables having correlation coefficients smaller than 0.5 between any pair of variables. The stepwise selection procedure was performed to obtain the significant variables from each set, and the results are presented in Table 4-5.

Table 4-4 Candidate Variable Sets

Variable Set	Variables
#1	BSAF, BTSO, CTSS, CFR, Snow
#2	BTAF, CTSS, CFR, Snow
#3	CTAF, CTSS, CFR, Snow

Table 4-5 Stepwise Selection Results

Model	Variable Set	Selected Variables	AIC*
1	#1	BSAF, CFR, Snow	148.2
2	#2	BTAF, CFR, Snow	144.1
3	#3	CFR, Snow	151.5

* Akaike Information Criterion

The matched case-control study intends to eliminate all factors other than traffic flow variables and weather. While fixing the location, time, and same day of the week helps to eliminate most nuisance variables, other variables such as human factors still exist. Therefore, a conditional logistic regression with the random intercept was introduced to help account for the heterogeneity among each stratum. This methodology assumes that the intercept of each event is random, and that therefore the intercepts of cases and controls in the same stratum can be different (Duchesne, Fortin, & Courbin, 2010). The R package "mclogit" was applied for all three models, and the random intercept was found to be not significant in any of the models; this suggests that the heterogeneity between an event and its non-event counterpart has been effectively mitigated using the matched case-control method (Elff, 2014).

The snow factor was significant, and its coefficient is positive in all models; this indicates that snow increases the likelihood of a crash during a lane-change maneuver. Model 2 had the smallest AIC value and was thus chosen as the best model. The results of the best model in Table 4-6 show that downstream traffic conditions are significantly related to the occurrence of lane-change related crashes upstream, as both traffic flow variables were collected from Stations B and C. BSAF has a smaller than 1 hazard ratio, while both CFR and Snow have hazard ratios larger than 1; this suggests that a lower traffic flow rate in the target lane at Station B and/or a higher traffic flow ratio at Station C under snowy conditions may increase the likelihood of a lane-change related crash.

Table 4-6 Model Results for Model 2

Variable	Description	Estimate	Pr > χ^2	Hazard Ratio
CFR	Flow ratio at Station C	1.307	0.0211	3.694
BTAF	Average flow of the target lane at Station B	-0.187	0.0046	0.829
Snow	Snow indicator	2.736	0.0006	15.421

Like any crash, the risk of a lane-change related crash can be formulated as the ratio of crashes to traffic exposure, where exposure is the lane change frequency. Table 4-6 shows that CFR, the flow ratio at the second downstream station, significantly affects the crashes of interest. In Chang and Kao's research, the average flow ratio (AFR) of a specific lane was found to be related to lane change frequency (Chang & Kao, 1991). Later, overall average flow ratio (OAFR), the geometric mean of modified AFR across all lanes, was applied to distinguish lane-change related crashes and rear-end crashes (C. Lee et al., 2006) and to model lane-change related crashes by lane (C. Lee et al., 2009). The lane-specific AFR was also proven to have a

similar effect as OAFR by Lee et al.(C. Lee et al., 2009). The chance of lane changes into the center lane(s) from either the left or right lane was assumed to be 50% for freeways with more than two lanes. However, this assumption is questionable in the absence of any solid evidence. The flow ratio used in this paper is calculated without making such an assumption. Using a lane-specific flow ratio between the subject and target lanes should be more accurate in modeling lane-change related crashes as opposed to using an overall flow ratio across all lanes.

BTAF, the average flow of the target lane at the first downstream station, was found to be significantly associated with lane-change related crashes. This finding suggests that higher volume in the target lane decreases the chances of a crash. Drivers in the subject lane tend to exercise caution or give up discretionary lane changes when few safe gaps exist in the target lane. Caution and discretion lead to a reduced likelihood of a crash occurrence. On the contrary, if the traffic volume in the target lane is low, the speed is usually higher. Drivers in the subject lane may perceive there to be larger gaps due to the low traffic volume, but they may not take into account the increased speed in the target lane. The aforementioned combination could lead to more crashes.

Almost all of the significant traffic variables in all three models are related to flow; this indicates a strong correlation between lane-change related crashes and flow-related variables. Lee et al. studied this phenomenon extensively after finding out that only flow-related variables were helpful in classifying lane-change related crashes and rear-end crashes, and in predicting lane-change related crashes by lane (C. Lee et al., 2006; C. Lee et al., 2009). The authors concluded that lane-change related crashes resulted from drivers' collective behaviors , i.e., traffic flow (C. Lee et al., 2009).

It is not surprising that snowy conditions contribute significantly to the occurrence of lane-change related crashes, as snow deteriorates the driver's visibility making it difficult to judge gap sizes and speeds of other vehicles. Snowy pavement surfaces also have a smaller friction coefficient, meaning it is more difficult to control the vehicle in these conditions. Furthermore, traffic flow parameters during snowy conditions, such as speed, may be rather different from those in non-snowy conditions. Although speed alone is not significantly related to lane-change crash occurrences, it may affect the possibility of crashes on snowy roads. The interaction term between snow and each of the six speed parameters was introduced one at a time in Model 2 in order to test this hypothesis. It was found that the main factor – snow – became insignificant when the interaction term was included, while the interaction term itself showed statistical significance. The interaction term remained significant after the snow factor was removed from the model, and the model presented better goodness-of-fit when CFR and BTAF were also included. Among six interaction terms, the product of snow and BTAS has the smallest AIC value of 140.3, a considerable decrease from 144.1, the AIC value of Model 2. The updated results are shown in Table 4-7.

Table 4-7 Results of the Model with the Interaction Term

Variable	**Description**	**Estimate**	**Pr > χ^2**	**Hazard Ratio**
CFR	Flow ratio at Station C	1.421	0.0154	4.141
BTAF	Average flow of the target lane at Station B	-0.195	0.0038	0.823
Snow×BTAS	Interaction of snow and the average speed of the target lane at Station B	0.059	0.0016	1.061

The signs of CFR and BTAF are the same as those in Model 2, and the coefficients change slightly. The interaction term reveals some interesting findings: on non-snowy days, BTAS (the speed of the target lane at the first downstream station) is not statistically significant; however, the crash likelihood on snowy days increases as speed increases. The finding stresses the aggravated impact of inclement weather on lane-change maneuvers that are combined with a high travel speed.

4.6 Conclusions

A matched case-control analysis was used to investigate the effects of traffic parameters on lane-change related crashes that took place on a corridor on I-94 to I-43 in Southeast Wisconsin during 2012 and 2013. All lane-change related crashes and specific lanes involved were determined by reviewing police crash reports. Non-crash events were identified for each crash in order to eliminate nuisance factors, or variables other than traffic and weather. In this study, traffic data were extracted for each lane related to a crash, and 63 traffic variables were created to better represent the prevailing traffic conditions prior to a crash. In addition, the weather information was collected from a historical weather database.

The matched case-control logistic analysis produced three models, and the model with the smallest AIC value was chosen. The results suggest that lower traffic flow in the target lane and/or a higher traffic flow ratio combined with snowy road conditions may increase the occurrence of upstream lane-change related crashes.

The effect of speed on crash occurrence was investigated, finding that speed itself had little influence on crashes unless snow was present. After improving the goodness-of-fit model to include the interaction of snow and speed (without snow as the main factor), results suggest a higher crash propensity when high travel speeds are combined with snowy conditions.

The study did not find any statistically significant variables related to traffic density. Intuition suggests that a decision to change lanes is based on the speed differential between the two lanes; but the action is contingent upon available gaps, which depend on traffic volume. Nevertheless, traffic density can be determined by flow and speed, and more research should be devoted to exploring the possible gap between the macroscopic features of lane-change related crashes and the microscopic features of driver behavior as they contribute to lane changes.

According to the model, the probability of a lane-change related crash under real-time traffic conditions can aid in flagging potential crash-prone conditions. The identified contributing factors can help traffic operators select traffic control and management countermeasures to proactively mitigate lane-change related crashes. But the development of effective traffic control strategies for crash prevention requires more future research which can identify, investigate, and validate the threshold values of the critical variables related to predicting lane-change related crashes.

4.7 References

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., & Santos, C. D. (2007). Crash Risk Assessment Using Intelligent Transportation Systems Data and Real-Time Intervention Strategies to Improve Safety on Freeways. *Journal of Intelligent Transportation Systems, 11*(3), 107-120. doi:10.1080/15472450701410395

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F. M., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board, 1897*(1), 88-95.

Abdel-Aty, M. A., Hassan, H. M., & Ahmed, M. (2012). *Real-Time Analysis of Visibility Related Crashes: Can Loop Detector and AVI Data Predict Them Equally?* Paper presented at the Transportation Research Board 91st Annual Meeting.

Ahmed, M., & Abdel-Aty, M. (2013). Application of Stochastic Gradient Boosting Technique to Enhance Reliability of Real-Time Risk Assessment: Use of Automatic Vehicle Identification and Remote Traffic Microwave Sensor Data. *Transportation Research Record: Journal of the Transportation Research Board*(2386), 26-34.

Ahmed, M., Abdel-Aty, M., & Yu, R. (2012). Bayesian Updating Approach for Real-Time Safety Evaluation with Automatic Vehicle Identification Data. *Transportation Research Record: Journal of the Transportation Research Board*(2280), 60-67.

Ahmed, M. M., & Abdel-Aty, M. (2012). The viability of using automatic vehicle identification data for real-time crash prediction. *Intelligent Transportation Systems, IEEE Transactions on, 13*(2), 459-468.

Al-Deek, H. M., Venkata, C., & Ravi Chandra, S. (2004). New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *Transportation Research Record: Journal of the Transportation Research Board, 1867*(1), 116-126.

Brackstone, M., McDonald, M., & Wu, J. (1998). *Lane changing on the motorway: Factors affecting its occurrence, and their implications.* Paper presented at the 9th International Conference on Road Transport Information and Control, 1998., London, UK.

Chang, G.-L., & Kao, Y.-M. (1991). An empirical investigation of macroscopic lane-changing characteristics on uncongested multilane freeways. *Transportation Research Part A: General, 25*(6), 375-389.

Chatterjee, I., & Davis, G. A. (2016). Analysis of Rear-End Events on Congested Freeways by Using Video-Recorded Shock Waves. *Transportation Research Record: Journal of the Transportation Research Board*(2583), 110-118.

Davis, G. A., & Swenson, T. (2006). Collective responsibility for freeway rear-ending accidents? An application of probabilistic causal models. *Accident Analysis & Prevention, 38*(4), 728-736.

Duchesne, T., Fortin, D., & Courbin, N. (2010). Mixed conditional logistic regression for habitat selection studies. *Journal of Animal Ecology, 79*(3), 548-555.

Elff, M. (2014). mclogit: Mixed Conditional Logit. http://CRAN.R-project.org/package=mclogit

Gipps, P. G. (1986). A model for the structure of lane-changing decisions. *Transportation Research Part B: Methodological, 20*(5), 403-414.

Hill, C., & Elefteriadou, L. (2013). Exploration of lane changing behavior on freeways. *Operations Traffic Management, Transportation Research Record, Washington, DC, USA*.

Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*: John Wiley & Sons.

Lee, C., Abdel-Aty, M., & Hsia, L. (2006). Potential real-time indicators of sideswipe crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board, 1953*(1), 41-49.

Lee, C., En, P., Young-Jin, P., & Abdel-Aty, M. A. (2009). *Effects of Lane-Change and Car-Following-Related Traffic Flow Parameters on Crash Occurrence by Lane.* Paper presented at the Transportation Research Board 88th Annual Meeting.

Lee, C., Saccomanno, F., & Hellinga, B. (2002). Analysis of crash precursors on instrumented freeways. *Transportation Research Record: Journal of the Transportation Research Board, 1784*(1), 1-8.

Lee, J., Park, M., & Yeo, H. (2013). *Empirical Analysis of Discretionary Lane Changes Using Probabilistic Models.* Paper presented at the Transportation Research Board 92nd Annual Meeting.

Li, Z., Ahn, S., Chung, K., Ragland, D. R., Wang, W., & Yu, J. W. (2014). Surrogate safety measure for evaluating rear-end collision risk related to kinematic waves near freeway recurrent bottlenecks. *Accident Analysis & Prevention, 64*, 52-61.

Oh, C., Park, S., & Ritchie, S. G. (2006). A method for identifying rear-end collision risks using inductive loop detectors. *Accident Analysis & Prevention, 38*(2), 295-301.

Pande, A., & Abdel-Aty, M. (2006a). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention, 38*(5), 936-948. doi:10.1016/j.aap.2006.03.004

Pande, A., & Abdel-Aty, M. (2006b). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 31-40.

Pande, A., & Abdel‐Aty, M. (2008). A computing approach using probabilistic neural networks for instantaneous appraisal of rear‐end crash risk. *Computer‐Aided Civil and Infrastructure Engineering, 23*(7), 549-559.

Qu, X., Wang, W., Wang, W., Liu, P., & Noyce, D. A. (2012). *Real-time prediction of freeway rear-end crash potential by support vector machine.* Paper presented at the Transportation Research Board 91st Annual Meeting.

Wang, J.-S., & Knipling, R. R. (1994). *Lane change/merge crashes: problem size assessment and statistical description*: US Department of Transportation, National Highway Traffic Safety Administration.

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention, 51*, 252-259.

Yu, R., & Abdel-Aty, M. (2014). Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. *Accident Analysis & Prevention, 62*, 161-167. doi:10.1016/j.aap.2013.08.009

Yu, R., Abdel-Aty, M., Ahmed, M. M., & Wang, X. (2014). Utilizing Microscopic Traffic and Weather Data to Analyze Real-Time Crash Patterns in the Context of Active Traffic Management. *Intelligent Transportation Systems, IEEE Transactions on, 15*(1), 205-213.

CHAPTER 5 PREDICTIVE ANALYSIS OF CRASH-PRONE CONDITIONS OF REAL-TIME CRASHES BY ACCOUNTING FOR SPATIAL-TEMPORAL ISSUE

5.1 Introduction

The readily available real-time traffic data from ATIS offer new opportunities for crash prediction and prevention in terms of traffic control and operations. Many studies have used real-time traffic data to investigate the relationship between crash risk and prevailing traffic conditions. Among all types of traffic sensors, inductive loop detectors have been widely used for real-time crash prediction.

The prevailing traffic circumstances prior to and under which a crash takes place are believed to be one of the major contributors to a crash. Additionally, a driver must constantly respond to changes in speed and space when driving in traffic, which can be highly stressful. Early detection of traffic anomalies that may result in crashes can help inform drivers and/or lead to implementation of appropriate traffic control strategies. It is imperative to identify patterns and trends of traffic conditions that lead to crashes so that they can be prevented.

Travel conditions can shift rapidly, and the traffic that a vehicle experienced immediately prior to or at the time of a crash is more relevant than earlier or later traffic conditions. The phenomenon of temporal proximity has been observed and supported in a study that predicted freeway crashes using loop detector data (Mohamed Abdel-Aty, Uddin, Pande, Abdalla, & Hsia, 2004). However, many studies did not consider the traffic conditions occurring right before a crash (e.g. 0-5 minutes period), citing that preventative actions may take extra time in a real-time crash identification, notification, and prevention system. Therefore, traffic data used in these

studies comes from earlier time periods (e.g. 5-10 minutes before a crash) (Mohamed Abdel-Aty et al., 2004; Hossain & Muromachi, 2012; Pande & Abdel-Aty, 2006a; Sun & Sun, 2015).

The time buffer between traffic data and crash occurrence is also related to the consistency between crash modeling and crash prediction, though it has never been explicitly discussed in previous studies. Figure 5-1 illustrates such consistency by considering the 5-min period for both crash modeling and crash prediction. The figure shows that one intends to predict the crash risk in the future moment, or the hypothesized crash time, which is 5 min from now. The traffic conditions in the past 5-min period are known, while those in the future 5-min period are not known. One can use only the known traffic information for crash prediction, thus, the traffic from the past 5-min period is used. However, crash modeling needs to be conducted in a consistent manner so that resultant crash prediction models can be applied. Initially, the historical crash time is consistent with the hypothesized crash time. Then the 0-5-min period before the crash would be consistent with the future 5-min period, and the 5-10-min period before the crash would be consistent with the past 5-min period. Therefore, the data from the 5-10-min period before the crash needs to be used for crash modeling so that the developed crash prediction models can be applied to predict the crash risk in real time based on the known traffic data from the past 5 minutes.

Figure 5-1 Consistent time periods for crash prediction and crash modeling.

The loop detector spacing can also lead to a lack of consistency, as spacings can vary substantially from site to site and across studies. For example, in one study the spacing ranges from 0.2 to 1.3 mi with an average of 0.5 mi (Xu, Liu, & Wang, 2016); in another it ranges from 0.15 to 1.68 mi with an average of 0.5 mi (Xu, Tarko, Wang, & Liu, 2013); and one other example has a range of 0.34 to 2.37 mi with an average of about 1.06 mi (Zheng, Ahn, & Monsere, 2010). Studies have shown that the sensor location may affect the estimation of traffic flow by producing inconsistently biased traffic data (Danczyk & Liu, 2011; Hong & Fukuda, 2012; Kwon, Petty, & Varaiya, 2007; Liu & Danczyk, 2009).

The discrepancies in the spatial-tempo domain mean that crash prediction models developed with traffic data collected directly from loop detector stations may be inadequate in unraveling the intrinsic relationship between crash risk and traffic conditions. Such data issues would undermine the prediction power of developed models. Even when a reliable crash prediction model is available, the issue of deploying effective preventative countermeasures remains. A performance assessment tool is needed to evaluate the effectiveness of intervening traffic control strategies before their deployment.

The objective of this chapter is to develop a method for real-time crash prediction and prevention using traffic simulation. Ideally, the method would be able to identify crash-prone conditions by accounting for the spatio-temporal consistency issue of loop detector data. Inspired by virtual loops extensively applied for vehicle detection, counting, and signal control, the cell transmission model (CTM) was employed to instrument a corridor of highway with virtual detector stations and measure traffic data where physical stations were not available.

5.2 Literature Review

Crashes should be more closely related to the traffic conditions occurring during or around the same time of the crash, as opposed to those occurring hours before. One study examined the impact of traffic variables on crash risk using five time slices: 0-5 minutes before the crash (time slice 1); 5-10 minutes before the crash (time slice 2); and up to 20-25 minutes before the crash (time slice 5) (Mohamed Abdel-Aty et al., 2004). The regression results showed that the traffic variables in time slice 1 are the most statistically significant among all five time slices, which supports the notion that the traffic conditions occurring right before a crash can best model the crash probability. However, most previous studies did not use this time period, citing that extra time was needed to take preventive countermeasures (Mohamed Abdel-Aty et al., 2004; Hossain & Muromachi, 2012; Pande & Abdel-Aty, 2006a; Sun & Sun, 2015). Furthermore, the distance between crash locations and detector locations varies from one case to another, making it impossible to obtain consistent measurements. The aforementioned issues regarding time and distance could undermine the validity and accuracy of real-time crash prediction models.

Ideally, the traffic conditions present at the time of the crash at the crash location should be used in studies that attempt to improve prediction accuracy. Although it is unrealistic to have physical detectors located at every crash location, the development of traffic simulation models has made the virtual detection possible. CTM, a macroscopic traffic flow simulation model that was first proposed by Daganzo (Daganzo, 1994), partitions a highway into continuous cells with user-defined lengths. Under the law of conservation, the traffic density in each cell within the highway evolves and follows the relationships derived from the fundamental diagram.

CTM can well accommodate traffic flow data collected from loop detectors, as they have shown promising results in predicting traffic flows using loop detector data as inputs (Muñoz,

Sun, Horowitz, & Alvarez, 2003, 2006; Sumalee, Zhong, Pan, & Szeto, 2011). Muñoz et al. achieved less than 13% of the mean error when simulating density using both CTM and switching-mode model (SMM) (Muñoz et al., 2003), as opposed to density collected from loop detectors. Muñoz et al. improved parameter calibration methods of CTM and SMM (Muñoz et al., 2006); calibrated CTM and SMM produced a 13% and 14% error, respectively, in estimating density, and a 4% and 5% error in estimating flow. Sumalee et al. proposed a stochastic CTM and achieved a 7.9% error in estimating density (Sumalee et al., 2011). CTM is therefore a reliable simulation tool that can generate trustworthy simulated traffic input for predicting crashes. Moreover, well-established traffic flow theories and emerging simulation algorithms provide timely support to the fast development of real-time crash prediction and prevention methods.

The CTM has the capability of simulating traffic control strategies. The CTM has several attractive features: 1) it is trustworthy in simulating TCS, as it is founded on sound traffic theory; 2) it is parsimonious, as it needs only a few parameters which can be estimated both online and off-line; 3) it requires low computational effort to predict traffic conditions in real-time (Hadiuzzaman & Qiu, 2013). Recently, the CTM has been applied to evaluate the safety effects of variable speed limits (VSL). Li et al. developed VSL in CTM and investigated its control strategy to reduce rear-end crash risks near recurrent bottlenecks on a 6-mile long virtual segment (Z. B. Li, Liu, Wang, & Xu, 2014). Later, Li et al. developed a strategy to optimize VSLs on a 29-mile freeway corridor in California (Z. Li, Liu, Xu, & Wang, 2016). In this study, VSL strategies were optimized to balance the impact on collision risk, injury severity, and travel time.

The relationships between the relatively low number of crashes and the massive volume of real-time traffic data can be sorted out through specific techniques. In general, the approaches for real-time crash prediction can be categorized as either statistical regression models or data mining techniques such as the Kohonen clustering algorithm, neural networks, and the Bayesian network (Hossain & Muromachi, 2012; Pande & Abdel-Aty, 2006a; Sun & Sun, 2015). Although data mining methods can accommodate correlation within independent variables for speed, flow, and occupancy (Hossain & Muromachi, 2012), they cannot identify explicit relationships between crash probability and traffic flow variables. Therefore, it is difficult to interpret the crash mechanism and develop effective crash prevention countermeasures. Statistical models, however, can build clear connections between crash probability and traffic flow variables, which is crucial for the development of proactive safety approaches. Among various statistical models used in real-time crash prediction studies, the binary logistic regression is widely used (Mohamed Abdel-Aty, Uddin, & Pande, 2005; Xu, Wang, & Liu, 2013; Zheng et al., 2010) because it can easily predict the crash probability given the explanatory variables.

5.3 Methodology

Crash probability prediction began with using CTM to simulate spatial and temporal traffic during the time period just prior to a crash. The crash occurrence probability was then estimated with simulated traffic conditions using a binary logistic regression model.

5.3.1 Cell Transmission Model (CTM)

CTM is a macroscopic traffic simulation model proposed by Daganzo (Daganzo, 1994). CTM is a powerful simulation technique which can capture many important traffic phenomena including queue formation and dissipation and shockwave propagation (Daganzo, 1994). CTM is more

computationally efficient and easier to configure and calibrate than microscopic simulation models. CTM also operates sufficiently with aggregated traffic data from detector stations. Figure 5-2 shows the fundamental diagram with and without a capacity drop for developing CTM.

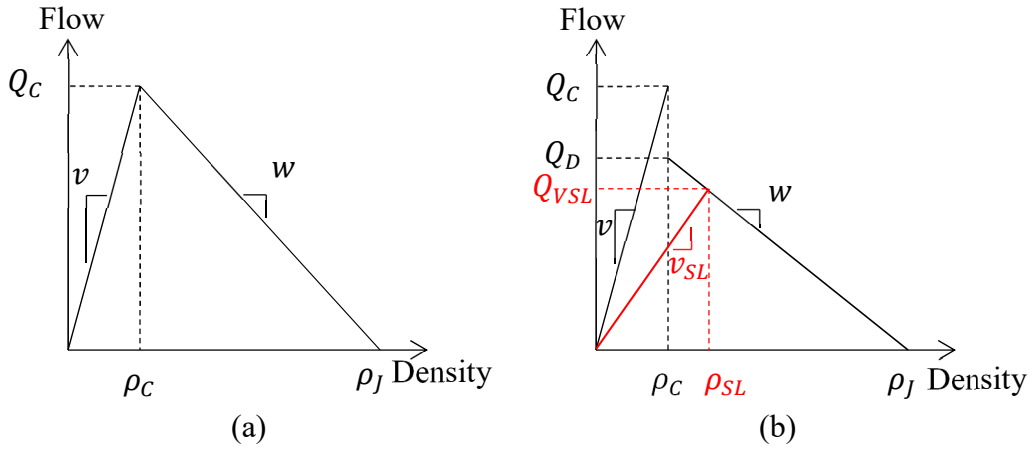


Figure 5-2 (a) Triangular fundamental diagram; (b) Fundamental diagram with capacity drop.

In CTM, a highway segment is divided into a series of cells. The density of each cell evolves following the conservation law of vehicles. Assuming that Cell *i* is characterized by the triangular fundamental diagram in Figure 5-2(a), where Q_C is the capacity flow, ρ_C is the critical density, ρ_J is the jam density, v is the free-flow speed, and w is the shockwave speed. The density for Cell *i* without on- or off-ramps is determined by Equation 5-1:

$$\rho_i(k+1) = \rho_i(k) + \frac{T}{l_i}(q_i(k+1) - q_i(k)) \tag{5-1}$$

where k is the time step index, $\rho_i(k)$ is the density of Cell *i* during the *k*th time step, T is the length of the time step, l_i is the length of Cell *i*, and $q_i(k)$ is the flow rate into Cell *i* during the *k*th time step. The flow rate is determined by the sending and receiving functions. For Cell *i*, the sending function $S_i(k)$ represents the maximum flow that can be supplied during the *k*th time

step, and the receiving function $R_i(k)$ represents the maximum flow that can be received. The two functions are determined in Equations 5-2 and 5-3, respectively:

$$S_i(k) = \min(v_i \rho_i(k), Q_{C,i}) \tag{5-2}$$

$$R_i(k) = \min(Q_{C,i}, w_i(\rho_{J,i} - \rho_i(k))) \tag{5-3}$$

The flow rate, $q_i(k)$, is determined by:

$$q_i(k) = \min(S_{i-1}(k), R_i(k)) \tag{5-4}$$

The fundamental diagram changes when the VSL control is deployed, as shown in Figure 5-2(b). v_{SL} is the deployed speed limit, and Q_{VSL} and ρ_{SL} are the new capacity and critical density after activating the VSL control. A study by Li et al. (Z. B. Li et al., 2014) showed that the sending and receiving functions affected by the VSL control are determined by Equations 5-5 and 5-6, respectively:

$$S_i(k) = \min\left(\min(v_i, v_{SL,i}) * \rho_i(k), Q_{VSL,i}\right) \tag{5-5}$$

$$R_i(k) = \min(Q_{VSL,i}, w_i(\rho_{J,i} - \rho_i(k))) \tag{5-6}$$

A phenomenon called "capacity drop" represents the discharge flow rate dropping below capacity after the congestion forms (Cassidy & Rudjanakanoknad, 2005; Hall & Agyemang-Duah, 1991). Accounting for capacity drop helps to better simulate traffic conditions. Capacity drop is accounted for by adopting the fundamental diagram in Figure 5-2(b) where Q_D is added to the triangular fundamental diagram. The capacity drops from Q_C to Q_D at the onset of congestion. Similar to the study by Li et al. (Z. B. Li et al., 2014), the modified sending and receiving functions are formulated in Equations 5-7 and 5-8, respectively:

$$S_i(k) = \begin{cases} v_i \rho_i(k), & if\ \rho_i(k) \le \rho_{C,i} \\ Q_{D,i}, & if\ \rho_i(k) > \rho_{C,i} \end{cases} \tag{5-7}$$

$$R_i(k) = \begin{cases} Q_{C,i},, & if\ \rho_i(k) \le \rho_{C,i} \\ w_i\left(\rho_{J,i} - \rho_i(k)\right), & if\ \rho_i(k) > \rho_{C,i} \end{cases} \quad (5\text{-}8)$$

5.3.2 Binary Logistic Regression Model

Equation 5-9 shows how the probability of a crash event is formulated in a binary logistic regression model:

$$p(\boldsymbol{X_i}) = \frac{g(X_i)}{1+e^{g(X_i)}} \quad (5\text{-}9)$$

where $p(\boldsymbol{X_i})$ represents the crash probability given $\boldsymbol{X_i} = (x_{i,1}, x_{i,2}, \dots, x_{i,k})$, a set of k explanatory variables for sample i, and $g(X_i)$ is a linear combination of the following variable set:

$$g(X_i) = \beta_0 + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \cdots + \beta_k * x_{i,k} \quad (5\text{-}10)$$

where $(\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ are the corresponding coefficients for $(x_{i,1}, x_{i,2}, \dots, x_{i,k})$.

The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ can be estimated by maximizing the following log-likelihood function:

$$lnL(\boldsymbol{\beta}, \boldsymbol{X_i}) = \sum_{i=1}^{n}[\left(\beta_0 + \beta_1 * x_{i,1} + \cdots + \beta_k * x_{i,k}\right) - \ln(1 + e^{-(\beta_0+\beta_1*x_{i,1}+\cdots+\beta_k*x_{i,k})}] \quad (5\text{-}11)$$

5.4 Data Description and Processing

Three data sources were consulted to develop a comprehensive approach: a) 1-min time interval traffic information from the WisTransPortal V-SPOC (Volume, Speed, and Occupancy) application suite (Parker & Tao, 2006); b) crash data from the web-based query and retrieval facility for Wisconsin Department of Transportation crash data and from reports archived in the WisTransPortal data management system; and c) weather information (e.g. snow, rain) from the Road Weather Information System (RWIS) in WisTransPortal.

5.4.1 Study Site and CTM Setup

A 4.15-mile corridor on I-94 East in Waukesha, WI was selected as the study site. The site was selected based on the following criteria: spacing of loop detector stations, traffic data quality, and crash sample size. The selected roadway corridor, as shown in Figure 5-3, has three lanes with one on-ramp and one off-ramp. The corridor consists of three segments, $\boldsymbol{S_1}$, $\boldsymbol{S_2}$, and $\boldsymbol{S_3}$, which are 1.77-mile, 0.79-mile and 1.59-mile long, respectively. Segment $\boldsymbol{S_2}$ starts at the end of the off-ramp and ends at the beginning of the on-ramp. The posted speed limit was 65 MPH in $\boldsymbol{S_1}$, and 55 MPH in $\boldsymbol{S_2}$, and $\boldsymbol{S_3}$. Other roadway characteristics such as lane width and shoulder width did not change along the corridor.

The corridor was instrumented with seven mainline loop detector stations: $\mathbf{N_1}$, $\mathbf{N_2}$, …, $\mathbf{N_7}$. The stations are referred to as physical stations so as to differentiate them from the virtual detectors introduce later. The seven stations space between 0.50 and 1.00 mile, with an average of 0.69 mile and a standard deviation of 0.20 mile. One loop detector station was located on the off-ramp, but no stations exist on the on-ramp. The traffic flow of the on-ramp can be imputed based on the conservation of vehicles using the flows from the nearest upstream and downstream detector stations.

The corridor was divided into 41 virtual cells for CTM simulation, and the cell length is uniform within each of the three segments. Segment $\boldsymbol{S_1}$ has 17 cells with a length of 0.104 mile; segment $\boldsymbol{S_2}$ has 8 cells with a length of 0.098 mile; segment $\boldsymbol{S_3}$ has 17 cells with a length of 0.099 mile. A virtual detector station was instrumented at the boundaries of cells, so there were 42 virtual detector stations and spacing between consecutive virtual stations averaged 0.1 mile with negligible variation. The off-ramp was located at the end of the 17th cell, while the on-ramp was located at the beginning of the 26th cell.

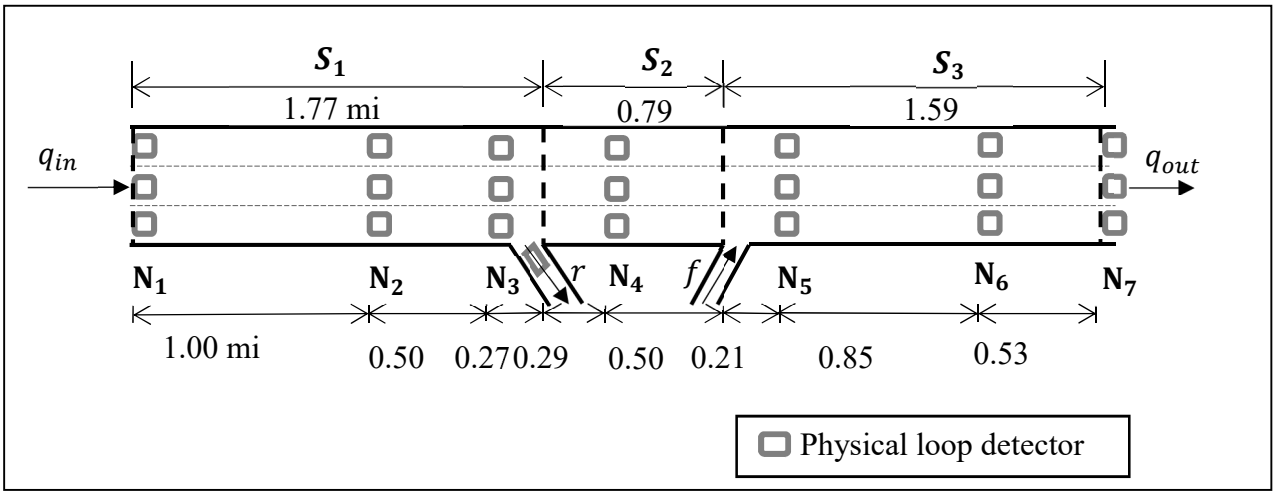


Figure 5-3 Layout of physical loop detector stations.

The virtual stations were set up at cell boundaries, similar to physical detector stations, to measure flow, speed, and density. Virtual stations were expected to capture traffic conditions at locations closer to the crash site.

Crashes that occurred at the study site from 2012 to 2014 were included. Any crash that happened within one hour after a crash occurrence was considered a secondary crash and was subsequently removed as indicated in (Hirunyanitiwattana & Mattingly, 2006). Crashes with missing times were excluded, as crash time is required to retrieve the traffic data.

A critical component of developing a crash prediction model is the knowledge of the traffic conditions experienced by the vehicle right before a crash; therefore, it is important to pinpoint the exact time in which a crash occurs. Crash times are sometimes rounded to the nearest 5-minute time stamp, and are therefore not reliable (Golob & Recker, 2003; Kockelman & Ma, 2010). Crash times in this study were carefully reviewed, and no rounding issue was found. Crashes were then randomly sampled and compared to the abrupt changes in traffic

conditions based on which crash times could be identified (Mohamed Abdel-Aty et al., 2005; Zheng et al., 2010). The validation result was positive, and the crash times from the database were used as the actual crash occurrence times.

5.4.2 CTM Calibration

A fundamental diagram is required to operate the CTM simulation. Differing roadway characteristics (e.g., horizontal curves, distances to on-/off-ramps, posted speed limits) mean different cells could have varying traffic patterns, which lead to different fundamental diagrams. Thus, one fundamental diagram was calibrated using the traffic data collected from each mainline detector station.

The fundamental diagram was based on the flow-density plot. The flows and speeds were collected from the loop detector stations, while the densities were determined by $\text{Density} = \frac{\text{Flow}}{\text{Speed}}$.

The calibration algorithm in Dervisoglu et al. (Dervisoglu, Gomes, Kwon, Horowitz, & Varaiya, 2009) was adopted with modifications to calibrate the fundamental diagram. The full description of the algorithm is summarized as follows:

1. Estimate the free-flow speed, v, using the least-squared method with flow-density pairs in the free-flow conditions. Since the speed limits of the segments are 65 MPH and 55 MPH, data points with speeds exceeding 55 mi/h in segment $\boldsymbol{S_1}$ and 45 mi/h in segments $\boldsymbol{S_2}$, and $\boldsymbol{S_3}$ were deemed to be in free flow conditions.
2. Find the maximum measured flow rate, q_{max}, as the capacity, Q_C. Critical density is determined by $\rho_C = \frac{Q_C}{v}$. Few and unsustainable observations with extremely high flow rates, a phenomenon of capacity overestimation, were observed. The formula to

compute the nominal capacity (in veh/h/lane) of freeways in HCM 2010 was adopted, as opposed to using the high flow rates (Transportation Research Board, 2010):

$$\text{Capacity} = \begin{cases} 2400\,, & \text{if FFS} \geq 70\text{mi/h} \\ 2400 - 10 \times (70 - \text{FFS}), & \text{if FFS} < 70\text{mi/h} \end{cases} \quad (5\text{-}12)$$

The capacity was then determined by taking the minimum of Q_c and the nominal capacity given by Equation 5-12.

3. Estimate the shockwave speed, w, and the jam density, ρ_J, using the least-squared method with flow-density pairs exceeding the critical density. The flow rate after the capacity drop was set as the value on the fitted flow-density line at the critical density.

Following the modified algorithm, fundamental diagram parameters were obtained for each physical detector station as shown in Table 5-1. Note that ρ_C, ρ_J, Q_c and Q_D are for three lanes. The magnitude of the capacity drop is from 2.0% to 6.9% for all physical stations except $\mathbf{N_4}$ which has a 13.9% capacity drop rate. The set of fundamental diagram parameters calibrated for one physical station was assigned to cells near that station.

Table 5-1 Fundamental Diagram Parameters by Physical Station

Station	$\boldsymbol{v}$ (mi/h)	$\boldsymbol{\rho_C}$ (veh/mi)	$\boldsymbol{\rho_J}$ (veh/mi)	$\boldsymbol{Q_C}$ (veh/h)	$\boldsymbol{Q_D}$ (veh/h)	$\boldsymbol{w}$ (mi/h)
$\mathbf{N_1}$	67.0	106.1	486.0	7111	6890	18.1
$\mathbf{N_2}$	68.4	104.6	588.4	7152	6816	14.1
$\mathbf{N_3}$	66.5	106.7	472.2	7095	6603	18.1
$\mathbf{N_4}$	59.8	97.0	799.0	5796	4989	7.1
$\mathbf{N_5}$	60.8	113.9	779.9	6924	6671	10.0
$\mathbf{N_6}$	58.0	118.0	460.4	6839	6703	19.6
$\mathbf{N_7}$	60.1	114.8	375.5	6903	6683	25.6

5.4.3 CTM Simulation

The simulation time step in CTM needs to be chosen so that the Courant–Friedrichs–Lewy (CFL) condition (Courant, Friedrichs, & Lewy, 1967) is fullfilled. A vehicle cannot travel across more than one cell during one simulation step in the CFL condition, i.e., $v_i * \Delta t \leq l_i$ where v_i is the free-flow speed, Δt is the simulation time step, and l_i is the cell length. A 5-second time step was used ($\Delta t = 5s$) based on the lengths of cells.

Entering flow and exiting flow of the highway corridor are required to run the CTM. The four flow inputs were required for the study site, including in-flow, q_{in}, out-flow, q_{out}, off-ramp flow, r, and on-ramp flow f (as shown in Figure 5-3). The 1-min flow data collected from the first physical station, $\mathbf{N_1}$, and the last physical station, $\mathbf{N_7}$, in the 0-5 min period prior to a crash/non-crash were used as the in-flow and out-flow of the corridor. A zeroth-order interpolation was applied to generate the 5-s in-flow, out-flow, on-ramp flow and off-ramp flow data. A CTM was then run to simulate how traffic in cells along the corridor evolves at each time step within the 5-min time interval.

In addition to the flow data, initial densities of cells at the beginning of the simulation interval are also needed for the CTM simulation. The initial density of a cell was obtained from the station's density data as long as the cell had one loop detector station. Densities of cells between two such cells were interpolated using the following approach:

1. Compute the density change rate as the ratio of the difference in densities of two cells with two consecutive loop detector stations and the distance between them: $\nabla\rho = \frac{\rho_{d,0}-\rho_{u,0}}{x_d-x_u}$, where $\nabla\rho$ is the density change rate; $\rho_{d,0}$ and $\rho_{u,0}$ are densities of cells having the downstream and upstream detector stations, respectively; x_d and x_u are the

locations of the beginnings of the two cells, that is, the locations of the two detector stations.

2. Determine the initial density of one cell between those two cells by the following:
3. $\rho_{i,0} = \rho_{u,0} + \nabla\rho * (x_i - x_u)$, where x_i is the location of the beginning of one cell between the two cells.

5.5 Crash Modeling

The simulated traffic data were collected from the virtual upstream and downstream stations to the cell location of each crash/non-crash in the prior 0-5-min period. The time period of 0-5 minutes prior to a crash was used in order to account for the temporal issue of physical station data, as the simulated traffic data in the future 5-min period would be employed for crash prediction. More details will be illustrated in section 6. It is worthwhile to test how the location of virtual upstream and downstream stations would impact the performance of crash prediction models. Therefore, both 0.2 mi and 0.5 mi distances, were selected as the distance from the crash cell location to its upstream and downstream virtual stations. One virtual upstream station and one virtual downstream station that are both 0.2 mi (i.e., two cells) away from the crash cell location were identified as stations from which to collect the simulated traffic data. The spacing between virtual upstream and downstream stations is 0.5 mi (2×0.2-mi+0.1-mi including the crash cell) in the 0.2-mi distance setting. Similarly, two virtual stations that are 0.5 mi (i.e., five cells) away from the crash location were used in the 0.5-mi distance setting. The spacing between the virtual upstream and downstream stations is 1.1 mi (2×0.5-mi+0.1-mi including the crash cell) in the 0.5-mi distance setting.

Having two distance settings also makes it possible to test the feasibility of simulated traffic data in crash modeling by providing a uniform and close layout of virtual stations to

account for spatial issues with physical station data. In the 0.5-mi distance setting, the spacing between virtual upstream and downstream stations is 1.1 mi, which is uniform for all crash and non-crash cases. However, the largest spacing between physical upstream and downstream stations is 1.0 mi, so the virtual upstream or downstream station is farther away from most crash (non-crash) case cells than the corresponding physical station. Therefore, the 0.5-mi distance setting provides traffic data from virtual stations which are only consistently located from but not closer to the crash/non-crash location compared to that from physical stations. The spacing between virtual upstream and downstream stations for the 0.2-mi distance setting, however, is 0.5 mi, which is not larger than the smallest spacing between physical stations. Therefore, the 0.2-mi distance setting provides traffic data from stations with both uniform and short distances from the crash (non-crash) location. The feasibility of uniform and close distances can be tested by comparing the performance of three different models: Model V1, which is developed with virtual station data in the 0.2-mi distance setting; Model V2, which is developed with virtual station data in the 0.5-mi distance setting; Model P, which is developed with physical station data. First, the feasibility of uniform distances can be assessed by comparing the performance of Model V2 and that of Model P. Then, the feasibility of close distances when distances are already uniform can be assessed by comparing the performance of Model V1 and Model V2.

The 5-s traffic data from the two selected virtual stations were aggregated into the 5-min interval for each crash and non-crash case and converted into traffic flow variables in Table 5-2. Due to the intercorrelation between the three traffic parameters of flow, density, and speed, traffic variables related to density and speed were kept to avoid serious correlations between candidate variables.

Table 5-2 Candidate Variables

Variable	Description
$AvgDen_u$	Average 5-s density at the upstream station (veh/mi)
$AvgSpd_u$	Average 5-s speed at the upstream station (mi/h)
$StdDen_u$	Standard deviation of 5-s density at the upstream station (veh/mi)
$StdSpd_u$	Standard deviation of 5-s speed at the upstream station (mi/h)
$AvgTsdDen_u$	Average time-series absolute difference in 5-s density at the upstream station (veh/mi)
$AvgTsdSpd_u$	Average time-series absolute difference in 5-s speed at the upstream station (mi/h)
$StdTsdDen_u$	Standard deviation of time-series difference in 5-s density at the upstream station (veh/mi)
$StdTsdSpd_u$	Standard deviation of time-series difference in 5-s speed at the upstream station (mi/h)
$AvgDen_d$	Average 5-s density at the downstream station (veh/mi)
$AvgSpd_d$	Average 5-s speed at the downstream station (mi/h)
$StdDen_d$	Standard deviation of 5-s density at the downstream station (veh/mi)
$StdSpd_d$	Standard deviation of 5-s speed at the downstream station (mi/h)
$AvgTsdDen_d$	Average absolute time-series difference in 5-s density at the downstream station (veh/mi)
$AvgTsdSpd_d$	Average absolute time-series difference in 5-s speed at the downstream station (mi/h)

$StdTsdDen_d$	Standard deviation of time-series difference in 5-s density at the downstream station (veh/mi)
$StdTsdSpd_d$	Standard deviation of time-series difference in 5-s speed at the downstream station (mi/h)
$AvgDiffDen_{d-u}$	Average difference between 5-s downstream and upstream density (veh/mi)
$AvgDiffSpd_{d-u}$	Average difference between 5-s downstream and upstream speed (mi/h)
$StdDiffDen_{d-u}$	Standard deviation of difference between 5-s downstream and upstream density (veh/mi)
$StdDiffSpd_{d-u}$	Standard deviation of difference between 5-s downstream and upstream speed (mi/h)
FF	1 = if the location is in the free-flow state; 0 = otherwise
BN	1 = if the location is in the bottleneck front state; 0 = otherwise
BQ	1 = if the location is in the back-of-queue state; 0 = otherwise
CT	1 = if the location is in the congestion state; 0 = otherwise
Curve	1 = Horizontal curve section; 0 = otherwise
OnRamp	1 = if there is an on-ramp between upstream and downstream stations; 0 = otherwise
OffRamp	1 = if there is an off-ramp between upstream and downstream stations; 0 = otherwise
Rain	1 = if the weather is rainy; 0 = otherwise
Snow	1 = if the weather is snowy; 0 = otherwise

Three additional groups of traffic variables were considered aside from mean and standard deviation of density and speed, roadway characteristics, and weather factors that have been frequently used in previous studies (M. Abdel-Aty & Pande, 2006; Mohamed Abdel-Aty et al., 2004; M. A. Abdel-Aty, Hassan, Ahmed, & Al-Ghamdi, 2012; Pande & Abdel-Aty, 2006b; Xu et al., 2016; Xu, Liu, Wang, & Li, 2012). The first group is related to the time-series difference in density and speed; the second group is related to the difference between downstream and upstream density and speed; the third group is related to the traffic state of the location.

The first traffic variable group is pertinent to the time-series difference in density and speed. The time-series difference is the difference between the density or speed in the next 5-s and that in this 5-s. Variables such as AvgTsdDen_u and StdTsdDen_u were calculated by Equation 5-13 and 5-14, and AvgTsdSpd_u and StdTsdSpd_u were calculated in the same way,

$$\text{AvgTsdDen}_u = \frac{\sum_{t=1}^{59} |Den_{u,t+1} - Den_{u,t}|}{59} \tag{5-13}$$

$$\text{StdTsdDen}_u = \sqrt{\frac{\sum_{t=1}^{59} [(Den_{u,t+1} - Den_{u,t}) - \frac{\sum_{t=1}^{59} (Den_{u,t+1} - Den_{u,t})}{59}]^2}{59-1}} \tag{5-14}$$

where $Den_{u,t}$ is the 5-s upstream density at time step t=1, 2, …, 60 (60 5-s in one 5-min interval). This variable group measures the traffic trend over time. The average absolute time-series difference in density or speed measures the traffic stability over time, and a large value indicates that the traffic is very unstable. The standard deviation of time-series difference in density or speed measures the consistency of traffic changes, and a large value indicates that the traffic changes are very fluctuant over time.

The second group is related to the difference between downstream and upstream density and speed. Variables such as $\text{AvgDiffDen}_{d\text{-}u}$ and $\text{StdDiffDen}_{d\text{-}u}$ were computed by Equation 5-15

and 5-16, and $\text{AvgDiffSpd}_{\text{d-u}}$ and $\text{StdDiffSpd}_{\text{d-u}}$ were calculated in the same way,

$$\text{AvgDiffDen}_{d-u} = \frac{\sum_{t=1}^{60}(Den_{d,t} - Den_{u,t})}{60} \quad (5\text{-}15)$$

$$\text{StdDiffDen}_{u} = \sqrt{\frac{\sum_{t=1}^{60}[(Den_{d,t} - Den_{u,t}) - \text{AvgDiffDen}_{d-u}]^2}{60-1}} \quad (5\text{-}16)$$

where $Den_{d,t}$ is the 5-s downstream density at time step t=1, 2, …, 60. This variable group indicates the difference between traffic conditions upstream and those downstream from the crash location. A large value of the average differences in density or speed implies that the upstream traffic conditions are very different from the downstream traffic conditions. A large value of the standard deviation of the differences implies that the traffic difference is not very consistent. Although the average absolute difference in upstream and downstream traffic parameters appears to have a significant relationship with the crash occurrence in (Xu et al., 2016; Xu, Liu, Wang, & Li, 2014), the average of the regular difference rather than of the absolute difference was considered because the sign may carry crucial information. For example, a positive AvgDiffSpd_{d-u} means that the downstream speed is higher than the upstream speed, while a negative AvgDiffSpd_{d-u} means the opposite condition. The former traffic condition may be more crash-prone, as high-speed vehicles from upstream may rear-end the slow-moving vehicles downstream. Therefore, this variable group is based on the regular difference which can reflect very different traffic conditions.

The third group is associated with the traffic state at the crash/non-crash location. The average density was used to measure the level of traffic congestion at the virtual upstream and downstream station (Yeo, Jang, Skabardonis, & Kang, 2013). Traffic is congested if the average density is greater than the critical density; otherwise, traffic is in free flow. The traffic state was determined based on the combination of the upstream and downstream traffic conditions:

1. Free Flow (FF): when both upstream state and downstream state are free flow;
2. Bottleneck front (BN): when upstream is congested and downstream is free flow;
3. Back of queue (BQ): when upstream is free flow and downstream is congested; and
4. Congested traffic (CT): when both upstream and downstream are congested.

The CTM cannot run for crash cases that have missing physical detector data, so after such crashes were removed, a total of 113 crashes remained crash modeling. 2,260 non-crash cases with a 20:1 non-crash to crash case ratio were randomly selected from 1,578,240-min intervals in 2012-2014 (60 min×24 h×1096 days in 2012-2014) at one out of 41 cells. Only the non-crash cases that are not within 2 hours from any crash were selected. The 5-min traffic data consisting of data from five 1-min intervals were retrieved from physical stations for non-crash cases in the same way that data were retrieved for crashes. The data were employed to generate simulated traffic data using the CTM. Candidate variables for all non-crash cases were obtained as well. The final dataset consists of 113 crash cases and 2,260 non-crash cases.

Table 5-3 shows the distribution of crash and non-crash cases by traffic state in two different distance settings. The distribution is different across two distance settings because the level of congestion could vary locally. However, two distributions present consistent patterns. Most crashes happened in the FF state for both distance settings, while the fewest happened in the BN state. The ratio of crash cases to non-crash cases indicates the crash probability in each state, and a larger ratio suggests a more crash-prone state. As expected, the ratios in the BN, BQ and CT states were considerably higher than those of the FF state.

Table 5-3 Case Frequency by Traffic State

	0.2-mi			0.5-mi		
Traffic State	**Crash**	**Non-Crash**	**Ratio**	**Crash**	**Non-Crash**	**Ratio**
FF	62	1,978	1:31.9	58	2,037	1:35.1
BN	5	90	1:18	13	90	1:6.9
BQ	15	95	1:6.3	19	82	1:4.3
CT	31	97	1:3.1	23	51	1:2.2

Traffic patterns may vary in different traffic states, so the traffic flow variables could have distinct distributions across traffic states. For example, Xu et al. (Xu et al., 2012) observed varying speed differences between upstream and downstream stations for different traffic states. The hypothesis was tested by dividing the whole dataset into subsets by traffic state. The distributions of traffic flow variables across traffic states were compared using a t-test. The comparison results show that none of the traffic variables have significantly similar distributions across traffic states, indicating that it would not be appropriate to develop a single model for all states without considering the interaction between the traffic variables and traffic states.

Crash-prone variables could vary in different traffic states. Data subsets for different states were used to identify statistically significant variables in each state. In each traffic state, the significance of each candidate variable was identified by developing a binary logit model for that variable only. A 10-fold modeling procedure was conducted to avoid spurious significance; the dataset for one traffic state was randomly split into ten subsets, and all variables' significance was checked for any nine out of the ten data subsets. Table 5-4 reports the number of significant runs for all candidate variables based on the 10% significance level. Variables were identified as

truly significant and were kept for further modeling as long as one variable was significant in at least eight out of ten runs. Correlations between truly significant variables in each traffic state were examined. Candidate models were developed with a maximum number of uncorrelated significant variables for each, and the model with the smallest AIC was selected as the optimal model.

Table 5-4 Number of Significant Runs for Candidate Variables

	0.2-mi				0.5-mi			
Variable	**FF**	**BN**	**BQ**	**CT**	**FF**	**BN**	**BQ**	**CT**
$AvgDen_u$	10	0	0	10	9	0	0	0
$AvgSpd_u$	10	0	0	0	10	0	6	5
$StdDen_u$	10	0	0	10	0	0	10	0
$StdSpd_u$	10	0	0	0	5	0	0	5
$AvgTsdDen_u$	10	1	0	1	10	1	10	0
$AvgTsdSpd_u$	10	0	1	5	10	0	0	3
$StdTsdDen_u$	10	0	0	1	10	0	10	2
$StdTsdSpd_u$	10	0	0	5	9	0	1	7*
$AvgDen_d$	10	0	8	0	10	0	10	1
$AvgSpd_d$	10	0	9	5	10	0	10	6
$StdDen_d$	10	0	0	0	8	0	1	2
$StdSpd_d$	10	0	0	0	8	3	5	7*
$AvgTsdDen_d$	10	1	10	2	10	4	0	2
$AvgTsdSpd_d$	10	1	7	0	10	3	2	6

$StdTsdDen_d$	10	0	6	5	9	0	1	5
$StdTsdSpd_d$	10	0	1	1	9	0	8	6
$AvgDiffDen_{d-u}$	0	0	0	10	4	0	9	0
$AvgDiffSpd_{d-u}$	0	0	0	10	0	0	5	0
$StdDiffDen_{d-u}$	10	0	2	1	10	1	0	0
$StdDiffSpd_{d-u}$	10	1	0	2	9	0	0	0
Curve	10	0	10	0	10	0	6	4
OnRamp	10	0	0	0	10	0	2	0
OffRamp	4	0	0	0	0	0	0	0
Rain	0	0	0	0	0	0	0	0
Snow	9	0	0	0	9	0	0	0

* This variable was considered for further modeling because of it has the highest number of significant runs for the corresponding traffic state though this number was less than eight.

Table 5-5 presents the modeling results by traffic states for both distance settings. The table shows that different traffic states have varying contributing variables for each distance setting, and that contributing variables are not identical in the same traffic state across two different distances. Significant variables in the 0.2-mi setting were checked first. The coefficients of $StdTsdDen_d$ and $StdTsdSpd_d$ for the FF state are positive, indicating that the crash risk increases as density and speed at downstream stations are more fluctuant. This is a logical finding because large variations in time-series changes in density and speed reflect turbulent traffic conditions that could increase crash potential. The negative sign of OnRamp suggests that the crash is less likely to happen near the on-ramp. One possible explanation for this is that drivers tend to be more alert when approaching an on-ramp and may therefore be less likely to

get involved in a crash. The Snow indicator has a positive sign, implying that snow contributes to crash occurrence in free flow traffic. However, this is not significant in the other traffic states, possibly because drivers tend to drive faster in free flow traffic than in the other states. No variables show significance for the BN state, possibly due to the small sample size.

The positive signs of $StdTsdDen_d$ and Curve for the BQ state indicate that the fluctuant time-series density at the downstream station near the curve would contribute to crash occurrence. The curve indicator shows significance only in this state; this could be because vehicles from the upstream free-flow traffic need to slow down to accommodate slow-moving traffic during congestion at downstream stations, and presence of a curve may worsen the deceleration. $AvgDen_u$ is significant and has a positive coefficient in the CT state. The finding indicates that crash risk increases with the increase in density at the upstream station. Upstream traffic is already congested at the upstream station in the CT state, which would increase upstream density and make the small distance headway even smaller, leading to higher crash likelihood.

Table 5-5 Modeling Results of Crash Prediction Models for Two Distances

0.2-mi				**0.5-mi**			
Variable	**Estimate**	**Standard Error**	**P-value**	**Variable**	**Estimate**	**Standard Error**	**P-value**
Traffic State							
FF							
Intercept	-4.444	0.244	<0.001	Intercept	-4.404	0.270	<0.001
$StdTsdDen_d$	0.461	0.082	<0.001	$StdTsdDen_d$	0.438	0.091	<0.001
$StdTsdSpd_d$	0.901	0.257	<0.001	OnRamp	-1.509	0.598	0.012

OnRamp	-1.036	0.473	0.028	Snow*	0.934	0.529	0.078
Snow	1.159	0.496	0.019				
BN							
Intercept	-2.415	0.466	<0.001	Intercept	-1.935	0.296	<0.001
BQ							
Intercept	-3.762	0.912	<0.001	Intercept	-4.351	0.935	<0.001
$StdTsdDen_d$	0.425	0.168	0.011	$StdTsdDen_u$	0.406	0.124	0.001
Curve	2.710	0.842	0.001	$AvgDiffDen_{d-u}$	1.012	0.005	0.016
CT							
Intercept	-2.642	0.865	0.002	Intercept	-1.445	0.432	<0.001
$AvgDen_u$	0.00824	0.00391	0.035	$StdTsdSpd_u$	0.613	0.308	0.046

* Significant at the 90% significance level.

Significant variables in the 0.5-mi setting were also checked. Three out of the four significant variables in the 0.2-mi setting for the FF state showed significance and consistent signs. Similar to the 0.2-mi setting, no variables were significant for the BN state possibly due to the small sample size. The positive sign of $StdTsdDen_u$ for the BQ state indicates that the fluctuant density over time at the upstream station would contribute to a crash occurrence. $AvgDiffDen_{d-u}$ also has a positive coefficient. Since the downstream density is higher than the upstream density in the BQ state, $AvgDiffDen_{d-u}$ is positive. A positive coefficient suggests that the crash risk increases when the downstream traffic gets more congested, while the upstream traffic becomes more of a free-flow state. The findings make sense, as drivers are more likely to make mistakes when they travel on a low-density roadway segment immediately followed by a high-density segment. $StdTsdSpd_u$ shows a positive sign for the CT state, indicating that the fluctuant speed over time at the upstream station would contribute to a crash occurrence. The

finding is logical, as continuous speed changes in already congested traffic would lead to a higher rear-end crash risk.

Separate models by traffic states were combined into one model for each distance to assess the impact of two distances on the prediction performance. Model V1 and Model V2 both include the indicator of BN, BQ, and CT state (FF is the reference state), along with interaction terms of traffic states and other variables. Interaction terms were constructed as the interaction of one traffic state and its significant variables, as identified in Table 5-6. For example, $StdTsdDen_d$ is significant in the FF state, and FF× $StdTsdDen_d$ is then the interaction term in the combined model. The modeling results show that main effects of both BN and CT states are statistically significant, while the main effect of BQ state is not significant. All interaction terms remain significant and their signs remain the same.

Table 5-6 Results of the Combined Models for Two Distances

0.2-mi (Model V1)				**0.5-mi (Model V2)**			
Variable	**Estimate**	**Standard Error**	**P-value**	**Variable**	**Estimate**	**Standard Error**	**P-value**
Intercept	-4.406	0.237	<0.001	Intercept	-4.400	0.260	<0.001
BN	1.990	0.523	<0.001	BN	2.465	0.394	<0.001
CT	1.764	0.897	0.049	CT	2.955	0.505	<0.001
FF× $StdTsdDen_d$	0.452	0.081	<0.001	FF× $StdTsdDen_d$	0.438	0.0892	<0.001
FF× $StdTsdSpd_d$	0.903	0.257	<0.001	FF×OnRamp	-1.510	0.598	0.0116
FF×OnRamp	-1.049	0.473	0.026	FF×Snow	0.933	0.529	0.0778
FF×Snow	1.146	0.495	0.021	BQ×$StdTsdDen_d$	0.410	0.0970	<0.001
BQ×$StdTsdDen_d$	0.530	0.083	<0.001	BQ×$AvgDiffDen_{d-u}$	0.0129	0.0033	<0.001
BQ×Curve	3.111	0.655	<0.001	CT×$StdTsdSpd_u$	0.613	0.308	0.047
CT×$AvgDen_u$	0.00824	0.00392	0.035				

A crash prediction model, Model P, was developed with observed traffic data collected from physical stations for comparison with the two models, Model V1 and Model V2, which were developed using the simulated traffic data from virtual stations. The prediction accuracy of the three models was checked by conducting the 10-fold cross-validation with selected significant variables from each model. The 10-fold cross-validation method first randomly partitions the dataset into ten equally sized subsamples. A single subsample is used as the validation dataset, and the other nine are used as training datasets. A model was then fitted with significant variables given the training dataset, and was then used to predict the crash probability of observation in the validation dataset. This procedure was repeated ten times, with each of the ten subsamples used exactly once as the validation dataset.

Based on the validation results, ROC (receiver operating characteristic) curves for all three models are plotted in Figure 5-4, and the AUC (Area Under Curve) values are presented in Table 5-7. The ROC curve is a plot of sensitivity against 1-specificity for different thresholds of predicted crash risk. The sensitivity represents the proportion of correctly predicted crash cases among all crash cases, or the prediction accuracy of crash cases, while specificity represents the proportion of correctly predicted non-crash cases among all non-crash cases. 1-specificity is the proportion of incorrectly predicted non-crash cases among all non-crash cases, which is also called the false alarm rate. A higher sensitivity along with a lower 1-specificity is preferred. The AUC value represents the total prediction accuracy, and a higher value is favored.

Model V2 provides a higher AUC than Model P which is developed based on observed traffic data from physical stations, though the difference is marginal. It suggests that for modeling real-time crashes, simulated traffic data collected at virtual stations with consistent distances is superior to observed traffic data collected from physical stations with inconsistent

distances. Model V1 provides a higher AUC than Model V2, suggesting that close distances are better when distances are already consistent. These two findings prove that simulated traffic data collected from uniformly and closely spaced virtual stations can provide better model performance by taking into account the spatial issue of physical station data.

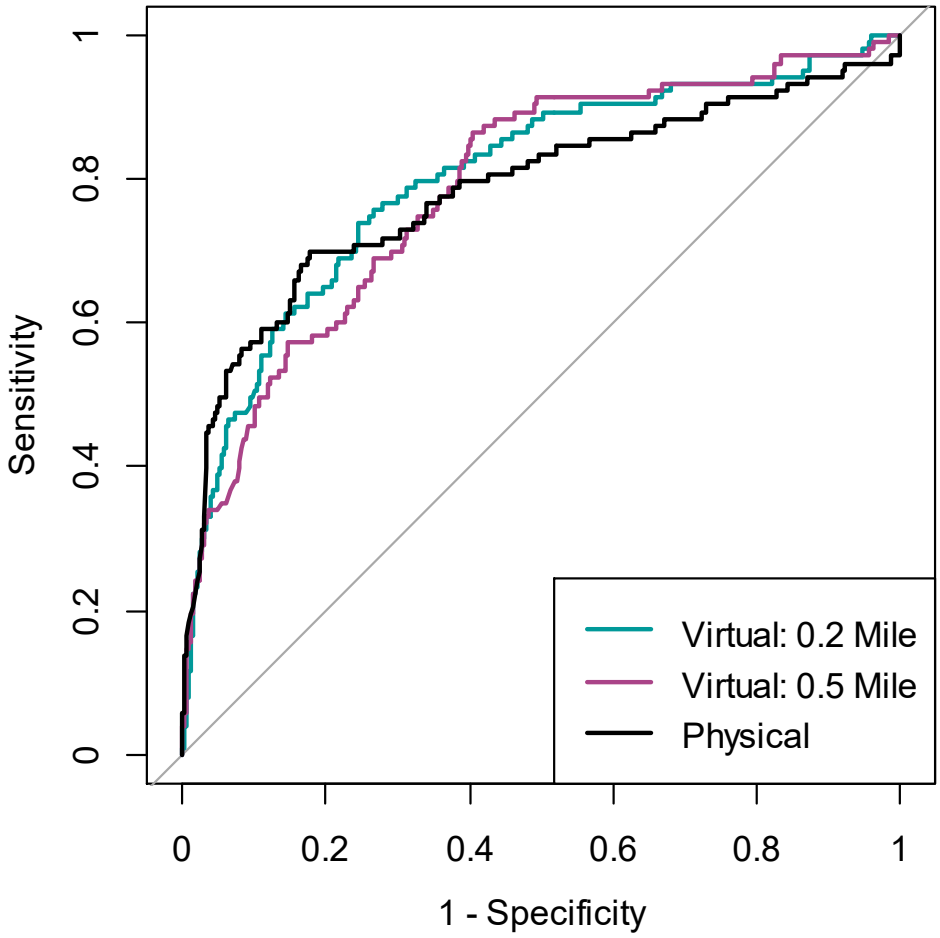


Figure 5-4 ROC curves for three models with different data sources.

Table 5-7 Area Under Curve (AUC) for Three Models

Virtual: 0.2-mi (Model V1)	Virtual: 0.5-mi (Model V2)	Physical (Model P)
0.800	0.787	0.784

5.6 Crash Prediction

In this study, a crash prediction method was proposed to identify crash-prone traffic conditions in real time. The crash prediction method takes the real-time data as the input. It first simulates the traffic in the future 5-min period using CTM and predicts the crash risk for that period based on simulated traffic data.

In the real-time crash prediction method, traffic conditions during the future 5-min period need to first be simulated using CTM. The initial densities of all cells were estimated with densities from the seven physical stations at the current moment. The flow inputs, including in-flow, q_{in}, off-ramp flow, r, and on-ramp flow f (as shown in Figure 5-3) in the future 5-min period were required for CTM simulation and were estimated using the k-nearest neighbor (k-NN) approach. The k-NN approach has been applied in a number of studies to forecast traffic flow rates and has shown promising results (Clark, 2003; Habtemichael & Cetin, 2016; Oswald, Scherer, & Smith, 2001; Smith, Williams, & Oswald, 2002).

The past 30 minutes was considered to be the most recent time period. Flows in the recent time period were considered as the subject flow set. All flow sets during the same time period from last 90 days were considered as candidate flow sets and were matched with the subject flow set. The ten nearest matches with the ten smallest distances were selected. The distance is determined by the following:

$$D(\boldsymbol{X^m}, \boldsymbol{Y}) = \sqrt{\sum_{i=1}^{30}(x_i^m - y_i)^2}, \mathrm{m} = 1, \ldots, 90 \quad (5\text{-}17)$$

where $\boldsymbol{X^m} = (x_1^m, \dots, x_{30}^m)$ is the mth candidate flow set of 30 1-minute flow points; $\boldsymbol{Y} = (y_1, \dots, y_{30})$ represents the subject flow set. The flow in the future 5-min period is calculated as the weighted average of flows in the next 5-min period for those matched flow sets by the following:

$$\boldsymbol{Y^F} = \frac{1}{10}\sum_{k=1}^{10} \frac{(D_k)^2}{\sum_{k=1}^{10}(D_k)^2} \boldsymbol{X^{k,F}} \tag{5-18}$$

where $\boldsymbol{Y^F} = (y_1^F, \dots, y_5^F)$ represents the estimated flow set in the future 5-min period, D_k is the kth smallest distance for kth nearest matched flow sets among those 10 nearest matched sets, and $\boldsymbol{X^{k,F}} = (x_1^{k,F}, \dots, x_5^{k,F})$ is the flow set in the next 5-min period for kth nearest matched flow sets.

After the required flows are estimated, they are used to run the CTM to simulate traffic in the future 5-min period. Simulated traffic is then used to predict the crash risk of each cell. The 0.2-mi distance setting shows better crash prediction performance, and is therefore applied to data collection and crash prediction. Simulated traffic data for each cell is collected from its upstream and downstream virtual stations, both of which are 0.2 mi away, and is then converted into variables as presented in Table 5-6. The predicted crash risk of Cell i is estimated as

$$p_i = \frac{e^{\pi}}{1+e^{\pi}} \tag{5-19}$$

$$\pi = -4.406 + 1.990 * BN + 1.764 * CT + 0.452 * (FF \times StdTsdDen_d)$$
$$+0.903 * (FF \times StdTsdSpd_d) - 1.049 * (FF \times OnRamp) + 1.146 * (FF \times Snow)$$
$$+0.530 * (BQ \times StdTsdDen_d) + 3.111 * (BQ \times Curve) + 0.00824 * (CT \times AvgDen_u)$$

Crash-prone traffic conditions are detected when the predicted crash probability exceeds an established threshold, which is 0.0427. The testing results showed that among the 113 cases, 104 cases exhibit crash-prone conditions, indicating the effectiveness of proposed crash prediction method.

5.7 Conclusions

This study aimed to develop a novel method for crash prediction and prevention which can accurately identify crash-prone conditions by accounting for the spatio-temporal issue of loop detector data.

Conventional real-time freeway crash prediction models identify crash-prone traffic conditions based on live feeds from loop detectors. It is common practice to use traffic data from the 5-10-min period prior to a crash, as this ensures sufficient time for taking the proper precautions. However, the phenomenon of time proximity suggests that traffic conditions occurring within the 0-5-min period of a crash are more relevant when it comes to predicting crashes. Moreover, a crash can happen between two detector stations where traffic information is not available, and the actual traffic conditions at the crash site may deviate from those captured by loop detector stations. Therefore, crash patterns derived from loop detector locations, as opposed to crash locations, are inadequate in accounting for varying distances between crashes and detectors. CTM-simulated traffic data were introduced in this study to fill the spatial and temporal gaps inherent in the observed traffic data collected from physical loop detector stations. Based on the traffic flow theory, CTM can predict traffic conditions anywhere at any time from its virtual detectors.

A real-time crash prediction model was developed with data from a corridor of I-94 in Wisconsin. The corridor was divided into a series of 0.1-mi long cells to create a uniform and close layout of virtual detector stations. Traffic data simulated from virtual upstream and downstream stations with consistent spacing was used for crash modeling to account for the spatial gap in physical station data. The simulated traffic data in the 0-5-min period prior to the crash/non-crash were used for crash modeling, and the traffic in the future 5-min period were

simulated for crash prediction. In this way, the temporal issue of physical station data was also taken into consideration.

Simulated traffic data collected from one virtual upstream station and one virtual downstream station 0.2-mi away were used for crash modeling. The same process was repeated for virtual stations that are 0.5-mi away. The modeling results showed that varying variables are significantly related to the crash occurrence in different traffic states. Observed traffic data collected from physical stations were also employed for crash modeling. The prediction performance of several crash prediction models was compared, showing that the simulated traffic data would improve prediction performance by accounting for the spatial-tempo issue of physical station data. It was also found that the 0.2-mi setting is better than the 0.5-mi setting for collecting simulated traffic data.

A crash prediction and prevention method based on simulated traffic data was proposed to detect crash-prone conditions. Results showed that the proposed crash prediction and prevention method could effectively detect crash-prone conditions.

The crash prediction and prevention method proposed in this study could be applied in ATIS to detect crash-prone traffic conditions and distribute crash warnings. Further improvements of CTM or equivalent simulation models will help to improve the current method.

5.8 References

Abdel-Aty, M., & Pande, A. (2006). ATMS implementation system for identifying traffic conditions leading to potential crashes. *IEEE Transactions on Intelligent Transportation Systems, 7*(1), 78-91. doi:10.1109/tits.2006.869612

Abdel-Aty, M., Uddin, N., & Pande, A. (2005). Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1908), 51-58.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*(1897), 88-95.

Abdel-Aty, M. A., Hassan, H. M., Ahmed, M., & Al-Ghamdi, A. S. (2012). Real-time prediction of visibility related crashes. *Transportation Research Part C: Emerging Technologies, 24*, 288-298.

Cassidy, M. J., & Rudjanakanoknad, J. (2005). Increasing the capacity of an isolated merge by metering its on-ramp. *Transportation Research Part B: Methodological, 39*(10), 896-913.

Clark, S. (2003). Traffic prediction using multivariate nonparametric regression. *Journal of transportation engineering, 129*(2), 161-168.

Courant, R., Friedrichs, K., & Lewy, H. (1967). On the partial difference equations of mathematical physics. *IBM journal of Research and Development, 11*(2), 215-234.

Daganzo, C. F. (1994). The Cell Transmission Model: Network Traffic. *Transportation Research Part B-Methodological, 29*(2), 79-93.

Danczyk, A., & Liu, H. X. (2011). A mixed-integer linear program for optimizing sensor locations along freeway corridors. *Transportation Research Part B: Methodological, 45*(1), 208-217.

Dervisoglu, G., Gomes, G., Kwon, J., Horowitz, R., & Varaiya, P. (2009, 2009). *Automatic calibration of the fundamental diagram and empirical observations on capacity*.

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

Habtemichael, F. G., & Cetin, M. (2016). Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C: Emerging Technologies, 66*, 61-78.

Hadiuzzaman, M., & Qiu, T. Z. (2013). Cell transmission model based variable speed limit control for freeways. *Canadian Journal of Civil Engineering, 40*(1), 46-56. doi:10.1139/cjce-2012-0101

Hall, F. L., & Agyemang-Duah, K. (1991). Freeway capacity drop and the definition of capacity. *Transportation Research Record*(1320).

Hirunyanitiwattana, W., & Mattingly, S. P. (2006). *Identifying secondary crash characteristics for California highway system.* Paper presented at the Transportation Research Board 85th Annual Meeting.

Hong, Z., & Fukuda, D. (2012). Effects of traffic sensor location on traffic state estimation. *Procedia-Social and Behavioral Sciences, 54*, 1186-1196.

Hossain, M., & Muromachi, Y. (2012). A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention, 45*, 373-381. doi:10.1016/j.aap.2011.08.004

Kockelman, K. K., & Ma, J. (2010). *Freeway speeds and speed variations preceding crashes, within and across lanes.* Paper presented at the Journal of the Transportation Research Forum.

Kwon, J., Petty, K., & Varaiya, P. (2007). Probe vehicle runs or loop detectors?: Effect of detector spacing and sample size on accuracy of freeway congestion monitoring. *Transportation Research Record: Journal of the Transportation Research Board*(2012), 57-63.

Li, Z., Liu, P., Xu, C., & Wang, W. (2016). Optimal Mainline Variable Speed Limit Control to Improve Safety on Large-Scale Freeway Segments: Optimal mainline variable speed limit. *Computer-Aided Civil and Infrastructure Engineering, 31*(5), 366-380. doi:10.1111/mice.12164

Li, Z. B., Liu, P., Wang, W., & Xu, C. C. (2014). Development of a Control Strategy of Variable Speed Limits to Reduce Rear-End Collision Risks Near Freeway Recurrent Bottlenecks. *IEEE Transactions on Intelligent Transportation Systems, 15*(2), 866-877. doi:10.1109/Tits.2013.2293199

Liu, H. X., & Danczyk, A. (2009). Optimal sensor locations for freeway bottleneck identification. *Computer‐Aided Civil and Infrastructure Engineering, 24*(8), 535-550.

Muñoz, L., Sun, X., Horowitz, R., & Alvarez, L. (2003, 2003). *Traffic density estimation with the cell transmission model*.

Muñoz, L., Sun, X., Horowitz, R., & Alvarez, L. (2006). Piecewise-linearized cell transmission model and parameter calibration methodology. *Transportation Research Record: Journal of the Transportation Research Board*(1965), 183-191.

Oswald, R. K., Scherer, W. T., & Smith, B. L. (2001). Traffic flow forecasting using approximate nearest neighbor nonparametric regression. *Center for Transportation Studies, University of Virginia*.

Pande, A., & Abdel-Aty, M. (2006a). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board, 1953*(1), 31-40.

Pande, A., & Abdel-Aty, M. (2006b). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 31-40.

Parker, S. T., & Tao, Y. (2006). *WisTransPortal: A Wisconsin Traffic Operations Data Hub.* Paper presented at the 9th International Conference on Applications of Advanced Technology in Transportation, Chicago, Ill.

Smith, B. L., Williams, B. M., & Oswald, R. K. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies, 10*(4), 303-321.

Sumalee, A., Zhong, R. X., Pan, T. L., & Szeto, W. Y. (2011). Stochastic cell transmission model (SCTM): A stochastic dynamic traffic model for traffic state surveillance and assignment. *Transportation Research Part B: Methodological, 45*(3), 507-533. doi:10.1016/j.trb.2010.09.006

Sun, J., & Sun, J. (2015). A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies, 54*, 176-186. doi:10.1016/j.trc.2015.03.006

Transportation Research Board. (2010). *HCM 2010: highway capacity manual.* Washington, D.C.: Transportation Research Board.

Xu, C., Liu, P., & Wang, W. (2016). Evaluation of the predictability of real-time crash risk models. *Accident Analysis & Prevention, 94*, 207-215. doi:10.1016/j.aap.2016.06.004

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

Xu, C., Liu, P., Wang, W., & Li, Z. (2014). Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. *Transportation Research Part A: Policy and Practice, 69*, 58-70. doi:10.1016/j.tra.2014.08.011

Xu, C., Tarko, A. P., Wang, W., & Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention, 57*, 30-39. doi:10.1016/j.aap.2013.03.035

Xu, C., Wang, W., & Liu, P. (2013). A genetic programming model for real-time crash prediction on freeways. *IEEE Transactions on Intelligent Transportation Systems, 14*(2), 574-586.

Yeo, H., Jang, K., Skabardonis, A., & Kang, S. (2013). Impact of traffic states on freeway crash involvement rates. *Accident Analysis and Prevention, 50*, 713-723. doi:10.1016/j.aap.2012.06.023

Zheng, Z., Ahn, S., & Monsere, C. M. (2010). Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention, 42*(2), 626-636. doi:10.1016/j.aap.2009.10.009

CHAPTER 6 PREDICTIVE ANALYSIS ON FREEWAY CRASHES USING LANE-SPECIFIC SIMULATED TRAFFIC DATA

6.1 INTRODUCTION

The prevailing traffic conditions are one of the major contributors to crashes. Thus, crash causation can be better understood by studying real-time traffic data with regard to speed, volume, and density that is collected during the time period leading up to a crash. Studying the appropriate data from the most critical time period supports informed decision-making on effective traffic operational strategies for improving safety.

Inductive loop detectors (ILD) are the most popular type of traffic sensor for real-time crash prediction. However, spatio-temporal issue exists in ILD traffic data that may detriment the validity of RTCPMs developed based on it as detailed in Section 5.1.

The objective of this chapter is to develop a RTCPM using simulated lane-specific traffic data generated from macroscopic traffic simulation while accounting for the spatial-tempo issue. The rest of the chapter is organized as follows: a lane-specific cell transmission model (LSCTM) was developed to instrument a corridor of highway with virtual detector stations on each lane; then, different traffic characteristics across lanes as well as lane-change activities between lanes were simulated and collected; next, a RTCPM was developed using lane-specific simulated traffic data; finally, the lane-specific RTCPM model was compared with models developed from field loop detector data.

6.2 Methodology

Lane-changing activities are closely related to the heterogeneous traffic conditions across lanes. Therefore, a LSCTM is needed to model both discretionary lane-changing (DLC) and mandatory

lane-changing (MLC) activities and their impact on lane-specific traffic flow. The LSCTM proposed in this study was constructed based on the work by Pan et al. (Pan, Lam, Sumalee, & Zhong, 2016).

6.2.1 Lane Change Probability and Minimum Gaps

First, lane-specific fundamental diagrams (FD) were applied over lanes, as opposed to a uniform FD, to model different traffic flow characteristics. A multilane freeway corridor is divided into a series of cell packages, and each cell package includes cells associated with lanes. A lane-specific triangular FD was adopted:

$$q_{i,m}(k) = \begin{cases} v_{f,i,m} \cdot \rho_{i,m}(k), & if\ \rho_{i,m}(k) \le \rho_{c,i,m} \\ w_{c,i,m} \cdot \left(\rho_{J,i,m} - \rho_{i,m}(k)\right), & if\ \rho_{i,m}(k) > \rho_{c,i,m} \end{cases} \quad (6\text{-}1)$$

where $q_{i,m}(k)$ (PCE/mile/h) and $\rho_{i,m}(k)$ (PCE/mile/lane) are the flow and traffic density of cell package i, lane m (denoted as *cell (i,m)* thereafter) during time interval $[k\Delta t, (k+1)\Delta t)$, respectively. $v_{f,i,m}$ (mile/h), $\rho_{c,i,m}$ (PCE/mile/lane), $w_{c,i,m}$ (mile/h), and $\rho_{J,i,m}$ (PCE/mile/lane) denote the free-flow speed, critical density, shockwave speed, and jam density of the FD for cell package i, lane m, respectively.

It was assumed that a DLC happens in order to gain a speed advantage. According to Laval and Daganzo (Laval & Daganzo, 2006), the probability for a vehicle to execute a DLC is defined by:

$$p_{i,m,DLC}^{\beta}(k) = min\left\{1, max\left\{0, \frac{v_i^{\beta}(k) - v_i^{m}(k)}{v_{f,i,m} \cdot \tau} \cdot \Delta t\right\}\right\} \quad (6\text{-}2)$$

where $v_i^{\beta}(k)$ and $v_i^{m}(k)$ are the speed of the adjacent lane and the current lane; τ is the duration of the lane change.

MLC was assumed to take place in order to reach the target turning point (e.g., off-ramp, accident, lane drop). The MLC probability is governed by:

$$p_{i,m,MLC}^{\beta}(k,x) = \begin{cases} exp\left(-\frac{(x(k)-x_c)^2}{(\sigma_{i,m}(k))^2}\right), \text{if } x(k) > x_c \\ 1, \qquad\qquad\qquad \text{if } x(k) \le x_c \end{cases} \tag{6-3}$$

$$\sigma_{i,m}(k) = \alpha_0 + \alpha_1 \cdot N_m^{tm} + \alpha_2 \cdot \bar{\rho}_\beta(k)$$

where $x(k)$ *is the remaining* distance to the target turning point; x_c is a critical distance to the turning point; N_m^{tm} is the number of lanes that the vehicle needs to cross from lane m to terminal lane tm; $\bar{\rho}_\beta(k)$ is the average traffic density of lane β; α_0, α_1, and α_2 are associated parameters. The MLC probability increases as the driver approaches the turning point. All drivers intending to make a MLC make the move once they pass the critical distance.

Both MLC and DLC need to consider the available gaps in the target lane in order to make a safe lane change. The lane change might not happen if the minimum gap is not guaranteed. According to Yang and Koutsopoulos (Yang & Koutsopoulos, 1996) and Pan et al. (Pan et al., 2016), the minimum gap for DLC depends only on the speed difference between the subject lane and the target lane, while the minimum gap for MLC also depends on the remaining distance to the target turning point (e.g., off-ramp, accident, lane drop). Vehicles intending to take a MLC consider the turning point to be remote when the remaining distance is $x(k) > x_r$, and close if the distance is $x(k) < x_c$. The minimum gap for MLC decreases linearly as the vehicle is approaching the turning point when the remaining distance is in the range of x_r and x_r. The minimum gaps for DLC and MLC from lane m to lane β, $g_{m,DLC}^{\beta}(k)$ (feet) and $g_{m,MLC}^{\beta}(k)$, are governed by:

$$g_{m,DLC}^{\beta}(k) = c_l \cdot \lfloor v_m(k) - v_\beta(k) \rfloor + c_f \cdot \lfloor v_\beta(k) - v_m(k) \rfloor + g_{min} \tag{6-4}$$

$$g_{m,MLC}^{\beta}(t) =$$

$$\begin{cases} c_l \cdot \lfloor v_m(k) - v_\beta(k) \rfloor + c_f \cdot \lfloor v_\beta(k) - v_m(k) \rfloor + g_{min}, & if\ x(t) > x_r \\ \left(c_l \cdot \lfloor v_m(k) - v_\beta(k) \rfloor + c_f \cdot \lfloor v_\beta(k) - v_m(k) \rfloor\right) \cdot \frac{x(k) - x_c}{x_r - x_c} + g_{min}, & if\ x_c \le x(k) \le x_r \\ g_{min}, & if\ x(k) < x_c \end{cases} \quad (6\text{-}5)$$

where the symbol $\lfloor z \rfloor$ is defined by:

$$\lfloor z \rfloor = \begin{cases} z, if\ z > 0 \\ 0, if\ z \le 0 \end{cases} \quad (6\text{-}6)$$

g_{min} denotes the minimum safe gap, and c_l and c_f are constants related to the extra lead gap and extra lag gap.

The speed obtained from the triangular FD which is adopted in Pan et al. (Pan et al., 2016) may not be appropriate for modeling the lane-changing traffic (del Castillo, 2012; Jin, 2010; Zhong, Pan, Sumalee, & Lam, 2014). A smooth function for estimating the speed (del Castillo, 2012; Jin, 2010) is adopted here:

$$v_{i,m}(k) = v_{f,i,m}\left\{1 - exp\left[1 - exp\left(\frac{w_{c,i,m}}{v_{f,i,m}}\left(\frac{\rho_{i,m}(t)}{\rho_{J,i,m}} - 1\right)\right)\right]\right\} \quad (6\text{-}7)$$

The average gap between successive vehicles on lane m of cell package i can be calculated as:

$$G_{i,m}(k) = \frac{5280 \cdot l_i - \rho_{i,m}(k) \cdot l_i \cdot L_p}{\rho_{i,m}(k) \cdot l_i} \quad (6\text{-}8)$$

where l_i (mile) is the length of cell package i, and L_p is the vehicle length of a passenger car. This equation calculates the average gap size between successive vehicles that are in lane m of cell package i by excluding the space they occupy.

The minimum gaps for DLC and MLC and the average gap are then normalized using the following equations:

$$\tilde{g}_{m,DLC}^{\beta}(k) = \frac{g_{m,DLC}^{\beta}(k)}{L_p} \quad (6\text{-}9)$$

$$\tilde{g}^{\beta}_{m,MLC}(k) = \frac{g^{\beta}_{m,MLC}(k)}{L_p} \tag{6-10}$$

$$\tilde{G}_{i,m}(k) = \frac{G_{i,m}(k)}{L_p} \tag{6-11}$$

6.2.2 Sending Function by Movement

Consider a multilane freeway segment with three lanes, lane $m-1$, m, and $m+1$. As shown in Figure 6-1, flows from all three lanes can merge into Cell i, lane m or *cell* (*i, m*), and the flow from *cell* (*i, m*) can diverge into three lanes. The sending and receiving functions can be determined by considering the merging and diverging flows.

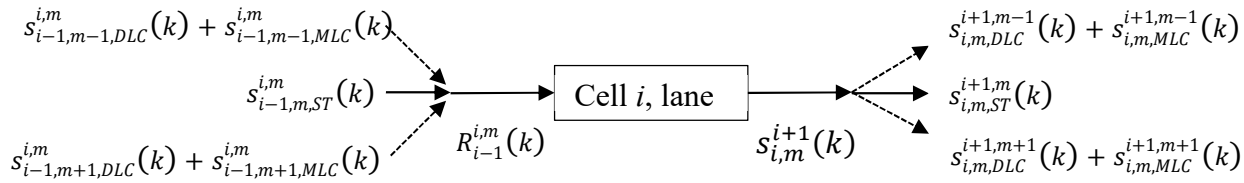


Figure 6-1 Merging and diverging of traffic flows of different movements.

The sending function, which denotes the flow intending to leave *cell* (*i,m*), is defined by:

$$s^{i+1}_{i,m}(k) = \begin{cases} v_{f,\mathrm{i},m} \cdot \rho_{i,m}(k), & if\ \rho_{i,m}(t) \le \rho_{i,c,m} \\ Q_{i,m}, & if\ \rho_{i,m}(t) > \rho_{i,c,m} \end{cases} \tag{6-12}$$

The sending function consists of both straight-moving flow and lane-changing flow (DLC and MLC), and as defined by:

$$s^{i+1}_{i,m}(k) = s^{i+1,m}_{i,m,ST}(k) + \sum_{\beta=m-1,m+1}(s^{i+1,\beta}_{i,m,DLC}(k) + s^{i+1,\beta}_{i,m,MLC}(k)) \tag{6-13}$$

where $\beta = m-1, m+1$ denotes the two adjacent lanes to lane m; $s^{i+1,m}_{i,m,ST}(k)$ is the straight-moving flow; $s^{i+1,\beta}_{i,m,DLC}(k)$ and $s^{i+1,\beta}_{i,m,MLC}(k)$ are the sending functions for DLC flow and MLC flow, respectively, which can be obtained by Equation 6-14 and 6-15. $\rho^{MLC}_{i,m}(k)$ is the density of

the MLC demand of *cell* (*i, m*). Note that the MLC demand includes the demand that stay in the terminal lane and ramp exiting traffic. It is assumed that the MLC demand will not make DLC when $x(k) \le x_r$ as approaching the target turning point is a higher priority than gaining a speed advantage.

$$s_{i,m,MLC}^{i+1,\beta}(k) = \begin{cases} v_{f,\mathrm{i},m} \cdot \rho_{i,m}^{MLC}(k) \cdot p_{i,m,MLC}^{\beta}(k,x), if\ \tilde{g}_{m,MLC}^{\beta}(k) \le \tilde{G}_{i,m}(k); \\ 0, \qquad\qquad if\ \tilde{g}_{m,MLC}^{\beta}(k) > \tilde{G}_{i,m}(k) \end{cases} \tag{6-14}$$

$$s_{i,m,DLC}^{i+1,\beta}(k) = \begin{cases} v_{f,\mathrm{i},m} \cdot \rho_{i,m}(k) \cdot p_{i,m,DLC}^{\beta}(k), if\ \tilde{g}_{m,DLC}^{\beta}(k) \le \tilde{G}_{i,m}(k) \& x(k) > x_r; \\ v_{f,\mathrm{i},m} \cdot \left(\rho_{i,m}(k) - \rho_{i,m}^{MLC}(k)\right) \cdot p_{i,m,DLC}^{\beta}(k), \\ \quad if\ \tilde{g}_{m,DLC}^{\beta}(k) \le \tilde{G}_{i,m}(k) \& x(k) \le x_r; \\ 0, \qquad\qquad if\ \tilde{g}_{m,DLC}^{\beta}(k) > \tilde{G}_{i,m}(k) \end{cases} \tag{6-15}$$

The sending function for straight-moving flow is the remaining portion of the sending function after excluding DCL and MLC flows. Straight-moving flow is defined by:

$$s_{i,m,ST}^{i+1,m}(k) = s_{i,m}^{i+1}(k) - \sum_{\beta=m-1,m+1}\left\{s_{i,m,DLC}^{i+1,\beta}(k) + s_{i,m,MLC}^{i+1,\beta}(k)\right\} \tag{6-16}$$

6.2.3 Receiving Function and Flow Propagation

The receiving function, which denotes the flow intending to enter *cell* (*i, m*), is defined by:

$$R_{i-1}^{i,m}(k) = \begin{cases} Q_{i,m}, & if\ \rho_{i,m}(t) \le \rho_{i,c,m} \\ w_{c,\mathrm{i},m} \cdot \left(\rho_{J,\mathrm{i},m} - \rho_{i,m}(k)\right), & if\ \rho_{i,m}(t) > \rho_{i,c,m} \end{cases} \tag{6-17}$$

The lane-changing flow from adjacent lanes needs to compete with the straight-moving flow in the target lane in order to enter the target lane. Consider that a straight-moving vehicle would occupy a space of *one* PCE length, and a lane-changing vehicle would occupy $\tilde{g}_{i-1,\beta,DLC}^{i,m}(k)$ or

$\tilde{g}_{i-1,\beta,MLC}^{i,m}(k)$ PCE lengths. The sending function is converted to $U_{i-1}^{i,m}(k)$ by accounting for different movements which require varying amounts of space:

$$U_{i-1}^{i,m}(k) = s_{i-1,m,ST}^{i,m}(k) + \sum_{\beta=m-1,m+1}\{\tilde{g}_{i-1,\beta,DLC}^{i,m}(k) \cdot s_{i-1,\beta,DLC}^{i,m}(k) + \tilde{g}_{i-1,\beta,MLC}^{i,m}(k) \cdot s_{i-1,\beta,MLC}^{i,m}(k)\} \quad (6\text{-}18)$$

The flow that is actually received by *cell (i, m)* from its current and adjacent lanes is defined by:

$$q_{i-1,m,ST}^{i,m}(k) = \begin{cases} s_{i-1,m,ST}^{i,m}(k), & if\ U_{i-1}^{i,m}(k) \le R_{i-1}^{i,m}(k) \\ \frac{R_{i-1}^{i,m}(k)}{U_{i-1}^{i,m}(k)} \cdot s_{i-1,m,ST}^{i,m}(k), & if\ U_{i-1}^{i,m}(k) > R_{i-1}^{i,m}(k) \end{cases} \quad (6\text{-}19)$$

$$q_{i-1,\beta,DLC}^{i,m}(k) = \begin{cases} s_{i-1,m,DLC}^{i,m}(k), & if\ U_{i-1}^{i,m}(k) \le R_{i-1}^{i,m}(k) \\ \frac{R_{i-1}^{i,m}(k)}{U_{i-1}^{i,m}(k)} \cdot s_{i-1,m,DLC}^{i,m}(k), & if\ U_{i-1}^{i,m}(k) > R_{i-1}^{i,m}(k) \end{cases} \quad (6\text{-}20)$$

$$q_{i-1,\beta,MLC}^{i,m}(k) = \begin{cases} s_{i-1,m,MLC}^{i,m}(k), & if\ U_{i-1}^{i,m}(k) \le R_{i-1}^{i,m}(k) \\ \frac{R_{i-1}^{i,m}(k)}{U_{i-1}^{i,m}(k)} \cdot s_{i-1,m,MLC}^{i,m}(k), & if\ U_{i-1}^{i,m}(k) > R_{i-1}^{i,m}(k) \end{cases} \quad (6\text{-}21)$$

where $q_{i-1,m,ST}^{i,m}(k)$, $q_{i-1,\beta,DLC}^{i,m}(k)$, and $q_{i-1,\beta,MLC}^{i,m}(k)$ denote straight-moving, DLC, and MLC flows received by *cell (i, m)*, respectively.

The density of *cell (i, m)* evolves over time based on the following flow conservation equation:

$$\rho_{i,m}(k+1) = \rho_{i,m}(k) + \frac{\Delta t}{l_i}\left\{q_{i-1,m,ST}^{i,m}(k) + \sum_{\beta=m-1,m+1}\left(q_{i-1,\beta,DLC}^{i,m}(k) + q_{i-1,\beta,MLC}^{i,m}(k)\right) - q_{i,m,ST}^{i+1,m}(k) - \sum_{\beta=m-1,m+1}\left(q_{i,m,DLC}^{i+1,\beta}(k) + q_{i,m,MLC}^{i+1,\beta}(k)\right)\right\} \quad (6\text{-}22)$$

The density of the MLC demand of *cell (i, m)* evolves over time based on the MLC flow that enters and leaves *cell* (*i, m*), as defined by 6-23.

$$\rho_{i,m}^{MLC}(k+1) = \rho_{i,m}^{MLC}(k) + \frac{\Delta t}{l_i}\left\{\sum_{\beta=m\pm1} q_{i-1,\beta,MLC}^{i,m}(k) - \sum_{\beta=m\pm1} q_{i,m,MLC}^{i+1,\beta}(k)\right\} \quad (6\text{-}23)$$

When the MLC demand increases due to a downstream bottleneck, a driver stops and waits for the chance to take the MLC if necessary. If the MLC demand needs to take the off-ramp, it is assumed that the portion that cannot reach the terminal lane at the off-ramp proceeds rather than stops or slows down as stopping or abruptly slowing down traffic on freeway is a hazard. An additional rule is added to guide the MLC demand movement near the off-ramp. As shown in Figure 6-2, an off-ramp is located at the end of *cell* (*J, m+1*). It is assumed that the MLC demand executes MLC only at the end of its current cell. Curved arrows in Fig. 2 represent MLC movements, which show that the MLC demand from *cell (J-1, m)* can cross one lane to *cell (J, m+1)* and exit. The MLC demand from *cell (J-2, m-1)* has to go to *cell (J-1, m)* first, and then to *cell (J, m+1)* to exit. However, the MLC demand from *cell (J-1, m-1)* can only go to *cell (J, m)* and cannot leave the freeway via the off-ramp. The MLC demand in *cell (J, m)* is assumed to proceed to the next cell.

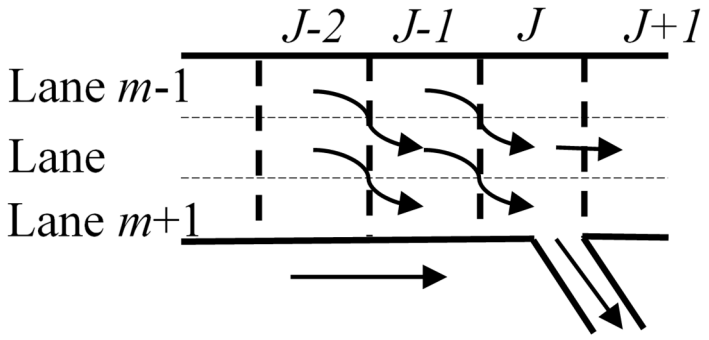


Figure 6-2 MLC movement near the off-ramp

6.3 Case Study

A case study was conducted to apply the artificial lane-specific traffic data and model freeway crashes. The study site and related data presented in Chapter 4 are used in this case study. More details about the data could be referred to Section 4.4.

6.3.1 LSCTM Setup and Calibration

The corridor was divided into 41 uniform 0.1-mile long virtual cells for CTM simulation. A virtual detector station was instrumented at the boundaries of cells, so there were 42 virtual detector stations. The spacing between consecutive virtual stations is 0.1 mile. The off-ramp was located at the end of the 17th cell, while the on-ramp was located at the beginning of the 26th cell. Similar to physical detector stations, virtual stations were expected to capture traffic conditions at locations closer to the crash site.

The calibration method detailed in Section 4.4.2 was used to develop lane-specific FDs based on the three-year traffic data collected from seven physical detector stations. Table 6-1 presents the FD parameters by lane. The first column, "Cell", represents the cells that have the same FD.

Table 6-1 Calibrated Fundamental Diagrams

Cell	Median Lane			Middle Lane			Shoulder Lane		
	v_f	Q	w_c	v_f	Q	w_c	v_f	Q	w_c
1-5	68.6	2385.6	15.9	67.3	2372.7	14.0	65.4	2353.8	9.3
6-12	69.1	2390.5	13.1	66.9	2369.4	12.3	69.4	2393.7	10.8
13-17	67.1	2370.7	18.3	67.8	2378.2	8.5	65.0	2350.4	10.9

18-25	60.7	2306.7	8.3	54.3	1932.0	10.2	58.1	2281.3	10.8
26-31	60.5	2304.8	10.6	62.7	2327.0	11.8	58.5	2284.8	10.1
32-39	61.2	2312.4	10.7	60.5	2305.3	15.4	52.2	2221.9	18.6
40-41	64.8	2316.0	12.9	56.6	2256.0	11.5	58.5	1932.0	9.7

6.3.2 LSCTM Simulation

Due to the data limitation, other parameters including α_0, α_1, α_2, τ, c_l, c_f and g_{min} are set to be -55.9, 726.9/lane, 33.7 mile/PCE, 3s, 1.32 feet*h/mile, 1.32 feet*h/mile, 37.7 feet based on previous literature (Laval & Daganzo, 2006; Pan et al., 2016; Yang & Koutsopoulos, 1996) without calibration. The PCE length, L_p, is set to be 20 feet. Due to the cell length, x_r *and* x_c are set to be the length of integer cells. Since the MLC demand behavior is different from the median lane and the middle lane, x_c is 0.1 mile for the middle lane and 0.2 mile for the median lane. x_r is 1.0 mile for the middle lane and 1.1 mile for the median lane (Pan et al., 2016).

The simulation time step in CTM needs to be chosen so that the Courant–Friedrichs–Lewy (CFL) condition (Courant, Friedrichs, & Lewy, 1967) can be met. A vehicle cannot travel across more than one cell during one simulation step in the CFL condition (i.e., $v_i * \Delta t \leq l_i$) where v_i is the free-flow speed, Δt is the simulation time step, and l_i is the cell length. The simulation time is set to be 3s since it cannot exceed τ, as pointed out by Laval and Daganzo (Laval & Daganzo, 2006).

Traffic conditions in the 0-5-min period before a crash are necessary for crash modeling. Traffic conditions for this time period can be simulated by collecting density data from seven mainline physical detector stations for the 15-min period before the crash occurrence. The data is then interpolated to obtain the initial densities of all cells by lane. The initial density of the MLC demand is set to be 0. The in-flow and out-flow in the 0-15-min period before the crash time are

collected from the first and the last physical detector stations. The average off-ramp flow during the same period collected from the off-ramp detector station is the entering MLC demand of the corridor during this 15-min period. The proportion of the entering MLC demand distributed over three lanes is set to be (1/3, 1/3, 1/3). It is reasonable to assume that MLC activities are likely to be evenly distributed across lanes at the beginning of the corridor which is 1.77 miles from the off-ramp and much larger than x_r. The LSCTM was run for 15 minutes with all necessary data. The first 10 minutes was considered a warm-up. The data simulated during the last 5 minutes was used as the traffic data for the 0-5-min period before the crash.

6.4 Crash Modeling

A total of 113 crashes remained after crashes that had missing physical detector data were removed. The crash record has location information that can be used to determine the location of the cell where the crash occurred. A total of 2,260 non-crash cases were randomly selected from 1,578,240-min intervals from 2012-2014 at one out of 41 cells, which is a 20:1 non-crash to crash case ratio.

Simulated traffic data were collected from the virtual stations upstream and downstream from the cell location of each crash/non-crash. One virtual upstream station and one virtual downstream station that were 0.2 mi (i.e., two cells) away from the crash cell location were identified as stations from which the simulated traffic data were collected.

The 3-s lane-specific traffic data in the prior 0-5-min period from the two selected virtual stations were aggregated over three lanes for each crash and non-crash case and then converted into variables including the average and standard deviation of flow, speed, and density along with the traffic state variable.

Additional non-traffic variables such as curve presence, ramp presence, and weather condition are included. Table 6-2 presents the candidate variables for developing the RTCPM. Drivers may behave differently in different traffic states, and the crash contributing factors may vary across states; therefore, one crash model was developed for the sub dataset with only one of the four traffic states, and only the significant variables in all four models were kept and combined into a single model for the whole dataset. This method is detailed in Section 4.5. The modeling results are presented in Table 6-3. The model includes indicators for BN, BQ, and CT states (FF is the reference state), as well as interaction terms for traffic states and other variables. The variable FF× $AvgDen_{u,}$ for example, is the average density at the upstream station when the traffic state is FF, which is 0 when the traffic state is not FF.

The modeling results show that all three traffic states are more crash-prone than the FF state. Results also show that contributing factors to the crash occurrence vary across traffic states. In the FF state, the estimates of AvgDenu and Snow are positive while that of On_Ramp is negative. This suggests that crash risk increases as the average density at the upstream station increases in snowy conditions. A higher density equals a smaller headway, which leads to a higher crash potential. The crash probability is higher under snowy conditions in the FF state, but not for other states. The negative On_Ramp sign suggests that a crash is less likely to happen near the on-ramp location. It is plausible that drivers tend to be more cautious when approaching an on-ramp.

Table 6-2 Candidate Variables

Variable	Description
$AvgDen_u$	Average 3-s density at the upstream station (veh/mi)
$AvgSpd_u$	Average 3-s speed at the upstream station (mi/h)
$StdDen_u$	Standard deviation of 3-s density at the upstream station (veh/mi)
$StdSpd_u$	Standard deviation of 3-s speed at the upstream station (mi/h)
$AvgDen_d$	Average 3-s density at the downstream station (veh/mi)
$AvgSpd_d$	Average 3-s speed at the downstream station (mi/h)
$StdDen_d$	Standard deviation of 3-s density at the downstream station (veh/mi)
$StdSpd_d$	Standard deviation of 3-s speed at the downstream station (mi/h)
FF	1 = if the location is in the free-flow state; 0 = otherwise
BN	1 = if the location is in the bottleneck front state; 0 = otherwise
BQ	1 = if the location is in the back-of-queue state; 0 = otherwise
CT	1 = if the location is in the congestion state; 0 = otherwise
Curve	1 = Horizontal curve section; 0 = otherwise
OnRamp	1 = an on-ramp between upstream and downstream stations; 0 = otherwise
OffRamp	1 = an off-ramp between upstream and downstream stations; 0 = otherwise
Rain	1 = if the weather is rainy; 0 = otherwise
Snow	1 = if the weather is snowy; 0 = otherwise

Table 6-3 Modeling Results

Variable	Estimate	Standard Error	P-value
Intercept	-4.154	0.380	<0.001
BN	2.208	0.611	<0.001
BQ	1.042	0.506	0.040
CT	3.305	0.766	<0.001
FF× $AvgDen_u$	0.043	0.015	0.004
FF× Snow	1.088	0.493	0.027
FF× On_Ramp	-2.019	1.014	0.046
BQ× $StdSpd_d$	0.133	0.049	0.007
CT× $AvgDen_u$	0.008	0.004	0.035
CT× $AvgSpd_d$	-0.034	0.011	0.003

The positive sign for BQ× $StdSpd_d$ suggests that for the BQ state, crash risk increases with a higher standard deviation of speed at the downstream station. This is logical, as more fluctuation in speed would lead to a higher crash potential. In the CT state, crash risk increases when the average density increases at the upstream station; crash risk decreases when the average speed increases at the downstream station. The higher average density increases the crash risk due to the smaller headway. The higher speed at the downstream station leads to less risk of rear-end crashes at the upstream station.

Validation results were used to plot receiver operating characteristic (ROC) curves for the two models, as shown in Figure 6-3. The ROC curve is a plot of sensitivity against 1-specificity for different thresholds of predicted crash risk. The sensitivity represents the proportion of

correctly predicted crash cases among all crash cases (prediction accuracy), while specificity represents the proportion of correctly predicted non-crash cases among all non-crash cases. 1-specificity is the proportion of incorrectly predicted non-crash cases among all non-crash cases, also referred to as the false alarm rate. A higher sensitivity along with a lower 1-specificity is preferred. The Area Under Curve (AUC) value represents the total prediction accuracy. A higher AUC value is favored. AUC values are 0.66 for Model V and 0.81 for Model P, respectively. The Model V's AUC is lower than Model P's, suggesting that simulated traffic data from uniformly and closely spaced virtual stations did not perform better.

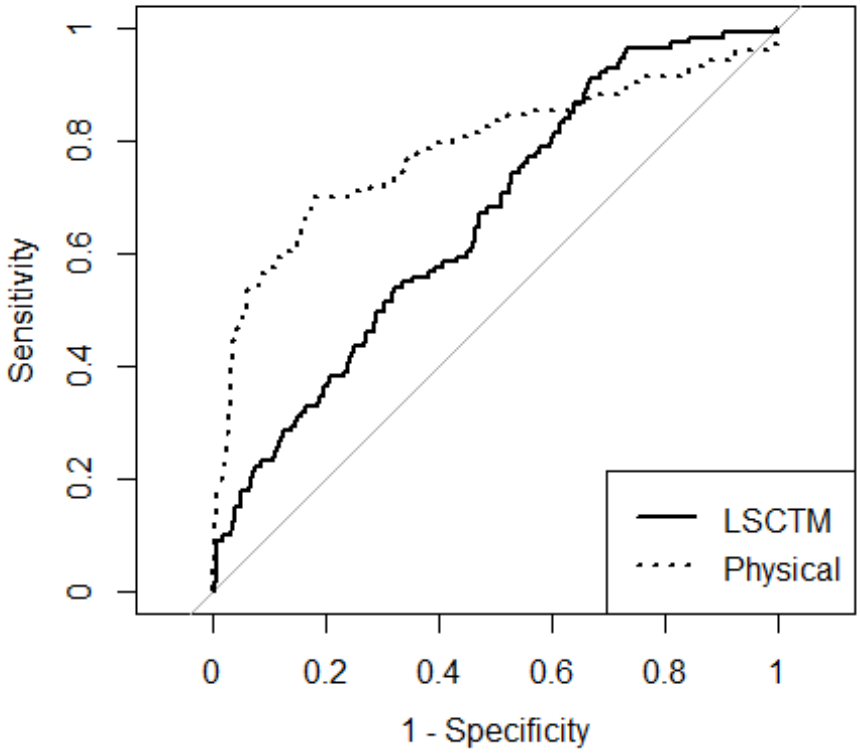


Figure 6-3 ROC curves for models with different data sources.

The less desirable performance of Model P could be due to the inaccurate data simulated by the LSCTM. Note that although the LSCTM operated with lane-specific FDs calibrated from

filed data, some important parameters were borrowed from other studies rather than fine-tuning filed data. Parameters can be carefully calibrated for satisfactory performance when all information is available. With accurate simulated lane-specific data, a RTCPM could be developed following the method detailed in Section 5.6.

6.5 Conclusions

A novel approach for addressing the spatial discrepancy issues that exist in most real-time crash prediction studies has been proposed. A LSCTM was developed to simulate both DLC and MLC activities. The method for developing the RTCPM used simulated traffic data and was demonstrated through a case study. Although the RTCPM developed from artificial data did not outperform the RTCPM developed using physical data, this study presents a viable alternative to utilizing macroscopic traffic simulation for safety analysis and evaluation.

Variables related to the heterogenous traffic between lanes have been utilized in previous studies (Lee, Abdel-Aty, & Hsia, 2006; Lee, En, Young-Jin, & Abdel-Aty, 2009; Xu, Wang, Liu, Wang, & Bao, 2018) to analyze crash probability; these studies could be extended using the lane-specific artificial traffic data. Moreover, the LSCTM can provide artificial lane-changing activities which could be very valuable for analyzing crashes related to lane changes.

6.6 References

Courant, R., Friedrichs, K., & Lewy, H. (1967). On the partial difference equations of mathematical physics. *IBM journal of Research and Development, 11*(2), 215-234.

del Castillo, J. M. (2012). Three new models for the flow–density relationship: derivation and testing for freeway and urban data. *Transportmetrica, 8*(6), 443-465.

Jin, W.-L. (2010). A kinematic wave theory of lane-changing traffic flow. *Transportation Research Part B: Methodological, 44*(8-9), 1001-1021.

Laval, J. A., & Daganzo, C. F. (2006). Lane-changing in traffic streams. *Transportation Research Part B: Methodological, 40*(3), 251-264.

Lee, C., Abdel-Aty, M., & Hsia, L. (2006). Potential real-time indicators of sideswipe crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 41-49.

Lee, C., En, P., Young-Jin, P., & Abdel-Aty, M. A. (2009). *Effects of Lane-Change and Car-Following-Related Traffic Flow Parameters on Crash Occurrence by Lane.* Paper presented at the Transportation Research Board 88th Annual Meeting.

Pan, T., Lam, W. H., Sumalee, A., & Zhong, R. (2016). Modeling the impacts of mandatory and discretionary lane-changing maneuvers. *Transportation Research Part C: Emerging Technologies, 68*, 403-424.

Xu, C., Wang, Y., Liu, P., Wang, W., & Bao, J. (2018). Quantitative risk assessment of freeway crash casualty using high-resolution traffic data. *Reliability Engineering & System Safety, 169*, 299-311.

Yang, Q., & Koutsopoulos, H. N. (1996). A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies, 4*(3), 113-129.

Zhong, R., Pan, T., Sumalee, A., & Lam, W. (2014). A cell transmission model with lane changing by lane-based fundamental diagram, assimilating lane speed observations and incorporation of uncertainty.

CHAPTER 7 CRASH PREDICTION AND PREVENTION APPLICATION

7.1 Introduction

A performance assessment tool is indispensable to evaluate the effectiveness of intervening strategies and promote the research findings from well-developed RTCPMs. A crash prediction and prevention application (CPPA) that combines both the RTCPM and the performance assessment tool can help detect crash-prone traffic conditions, distribute crash warnings, and evaluate traffic control countermeasures before their deployment.

Variable speed limit (VSL) is a traffic control technique that is used to increase mobility and reduce crash risks on freeway mainlines. Unlike typical static speed limit signs, the VSL dynamically posts a speed limit based on current traffic, weather, traffic safety level or other conditions. Although the VSL is mainly designed to improve mobility, its effect on safety has also been demonstrated. VSL has been reported to reduce the crash risks by 10-80% (Abdel-Aty, Cunningham, Gayah, & Hsia, 2008; Abdel-Aty, Dilmore, & Dhindsa, 2006; Abdel-Aty, Pande, Lee, Gayah, & Santos, 2007; Allaby, Hellinga, & Bullock, 2007; Choi & Oh, 2016; Hellinga & Mandelzys, 2011; Lee & Abdel-Aty, 2008; Lee, Hellinga, & Saccomanno, 2006; Li, Li, Liu, Wang, & Xu, 2014; Li, Liu, Wang, & Xu, 2014; Li, Liu, Xu, & Wang, 2016). Due to the effectiveness of VSL in reducing the crash risk, a CPPA which aims to evaluate both the safety and mobility impacts of VSL is developed in this chapter.

7.2 CPPA Development

The CPPA is developed based on a RTCPM within the CTM environment. The RTCPM developed in Chapter 5 which is based on traffic data simulated by traditional CTM outperformed the counterpart developed using ILD data, while the model developed in Chapter 6

which is based on traffic data simulated by LSCTM fails to provide superior performance than that developed using ILD data. Therefore, the RTCPM developed using CTM simulated data in Chapter 5 is selected as the model in CPPA, and the traditional CTM is applied as the simulation environment.

Figure 7-1 presents the process of CPPA. The application consists of a crash prediction module and a crash prevention module. The crash prediction module takes the real-time data as the input. It first simulates the traffic in the future 5-min period using CTM and predicts the crash risk for that period based on simulated traffic data. If the predicted crash risk exceeds the pre-specified threshold, the crash prevention module will be activated. Several candidate TCS alternatives are considered to reduce the crash risk. Each TCS alternative will be simulated in CTM to produce what traffic conditions would be in the future 5-min period if that TCS is deployed. The predicted crash risk is estimated based on the simulated traffic data, and the safety impacts of that TCS are evaluated. The optimal TCS is chosen based on established criteria.

In the real-time crash prediction module, traffic conditions during the future 5-min period need to first be simulated using CTM. The initial densities of all cells were estimated with densities from the seven physical stations at the current moment. The flow inputs, including in-flow, q_{in}, off-ramp flow, r, and on-ramp flow f (as shown in Figure 5-3) in the future 5-min period were required for CTM simulation and were estimated using the k-nearest neighbor (k-NN) approach. The k-NN approach has been applied in a number of studies to forecast traffic flow rates and has shown promising results (Clark, 2003; Habtemichael & Cetin, 2016; Oswald, Scherer, & Smith, 2001; Smith, Williams, & Oswald, 2002).

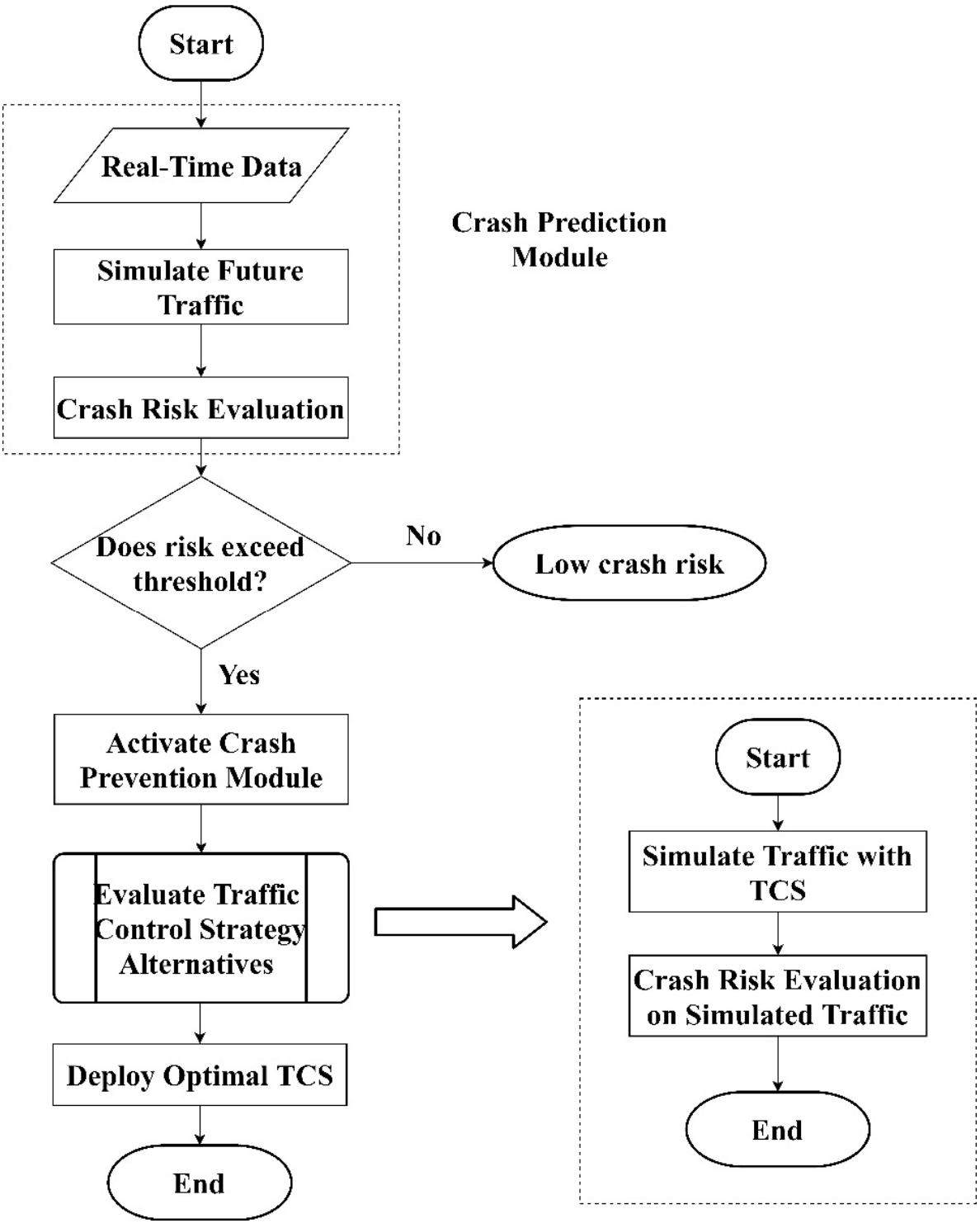


Figure 7-1 Process of the crash prediction and prevention application (CPPA).

The past 30 minutes was considered as the most recent time period. Flows in the recent time period were considered as the subject flow set. All flow sets during the same time period from last 90 days were considered as candidate flow sets and were matched with the subject flow set. The ten matches with ten smallest distances were selected. The distance is determined by:

$$D(\boldsymbol{X^m}, \boldsymbol{Y}) = \sqrt{\sum_{i=1}^{30}(x_i^m - y_i)^2}, \text{m} = 1, \ldots, 90 \quad (7\text{-}1)$$

where $\boldsymbol{X^m} = (x_1^m, \ldots, x_{30}^m)$ is the *m*th candidate flow set of 30 1-minute flow points; $\boldsymbol{Y} = (y_1, \ldots, y_{30})$ represents the subject flow set. The flow in the future 5-min period is calculated as the weighted average of flows in the next 5-min period for those matched flow sets by:

$$\boldsymbol{Y^F} = \frac{1}{10}\sum_{k=1}^{10}\frac{(D_k)^2}{\sum_{k=1}^{10}(D_k)^2}\boldsymbol{X^{k,F}} \quad (7\text{-}2)$$

where $\boldsymbol{Y^F} = (y_1^F, \ldots, y_5^F)$ represents the estimated flow set in the future 5-min period, D_k is the *k*th smallest distance for *k*th nearest matched flow sets among those 10 nearest matched sets, and $\boldsymbol{X^{k,F}} = (x_1^{k,F}, \ldots, x_5^{k,F})$ is the flow set in the next 5-min period for *k*th nearest matched flow sets.

After the required flows are estimated, they are used to run the CTM to simulate traffic in the future 5-min period. Simulated traffic is then used to predict the crash risk of each cell. The 0.2-mi distance setting shows better crash prediction performance, and is therefore applied to data collection and crash prediction. Simulated traffic data for each cell is collected from its upstream and downstream virtual stations, both of which are 0.2 mi away, and is then converted into variables as presented in Table 5-6. The predicted crash risk of Cell *i* is estimated as

$$p_i = \frac{e^{\pi}}{1+e^{\pi}} \quad (7\text{-}3)$$

$$\pi = -4.406 + 1.990 * BN + 1.764 * CT + 0.452 * (FF \times StdTsdDen_d)$$
$$+0.903 * (FF \times StdTsdSpd_d) - 1.049 * (FF \times OnRamp) + 1.146 * (FF \times Snow)$$
$$+0.530 * (BQ \times StdTsdDen_d) + 3.111 * (BQ \times Curve) + 0.00824 * (CT \times AvgDen_u)$$

Crash-prone traffic conditions are detected when the predicted crash probability exceeds an established threshold. If crash-prone conditions are detected, the crash prevention module will be activated. The safety impacts of various TCS are then evaluated. The optimal traffic control strategy is then deployed to improve the safety condition.

The proposed CPPA was applied to the study site for demonstration. Figure 7-2 presents the layout of VSL signs along the study corridor. Eight coordinated VSL signs are marked from VSL 1 to VSL 8 and all spaces between adjacent VSL signs are 0.50 mi. Each 0.50-mi spacing consists of five uniform 0.10-mi cells, so there are 35 cells between VSL 1 and VSL 8.

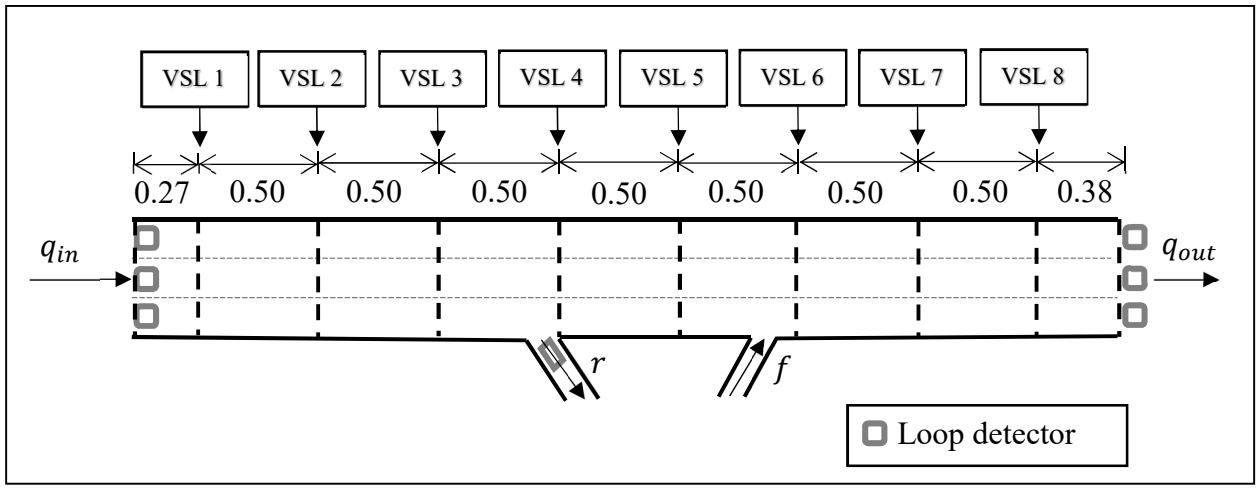


Figure 7-2 Layout of VSL signs along the corridor.

The VSL control strategy proposed in this study was to gradually reduce the posted speed limits of activated VSL signs until a target speed drop was achieved. When the predicted crash probability of one cell in the future 5-min interval exceeds the pre-specified threshold, the nearest upstream VSL sign will be activated. The pre-specified crash probability threshold was set to be 0.0427 because it provided desirable classification performance with the maximum summation of sensitivity and specificity.

Several parameters need to be decided to develop an effective VSL control, including target speed drop, speed change rate, and maximum speed difference between adjacent VSL signs. Two target speed drop alternatives were proposed: 10 MPH speed drop and 20 MPH speed drop. The target speed limit would be 55 MPH with a 10 MPH speed drop and 45 MPH with a 20 MPH speed drop, with an initial speed limit of 65 MPH. The speed change rate determines how fast the VSL sign should change the posted speed limit. A large speed change rate may introduce significant traffic disturbances, whereas a small speed change rate could fail to achieve the target speed limit in a reasonable time period. VSL signs were coordinated to create smooth speed changes between consecutive links. The maximum speed difference between adjacent VSL signs needs to be satisfied. The speed change rate was set to be 10 MPH per 30 s, meaning that the posted speed limit reduces by 10 MPH and stays for 30 s until the next speed change. The maximum speed difference between consecutive VSL signs was set to be 10 MPH. The values for these two parameters have been proven to produce a satisfactory performance for the VSL control (Li, Li, et al., 2014).

Once the crash prevention module is initiated, the proposed VSL strategy with two speed drop alternatives would be simulated in the CTM for 5 min, and then simulated traffic would be used to assess the safety effects and mobility effects. The safety effect is measured as

$$R = \sum_{i=1}^{I} r_i \tag{7-4}$$

$$r_i = \begin{cases} p_i - p_{thre}, & if\ p_i > p_{thre} \\ 0, & if\ p_i \le p_{thre} \end{cases}$$

where R is the crash risk of the corridor, r_i is the crash risk of Cell i, p_i is the predicted crash probability of Cell i and can be estimated using Equation 7-3 given the simulated traffic flow, p_{thre} is the threshold of predicted crash probability for crash classification, which is 0.0427. The mobility effects are measured by the Total Travel Time (TTT).

The proposed CPPA was tested on the 113 crash cases that were used for developing crash prediction models. Five minutes before each crash occurrence was equivalent to the "current moment"; the 0-5-min interval before its crash time was equivalent to the "future 5-min period"; the 30-min interval before the "current moment" was equivalent to the recent time period. The flows were estimated using the k-NN approach and were then applied to simulate the traffic in the "future 5-min period". The crash risk of each cell was re-predicted using Equation 7-3 based on the simulated traffic. The crash prevention module was activated when the crash risk of any cell exceeded the threshold. One control strategy would be deployed among three alternatives: 1) Non-activated VSL, 2) VSL control with 10 MPH drop, and 3) VSL control with 20 MPH drop. The non-activated VSL strategy would not change the traffic conditions and therefore would not change the crash risk. The control strategy that can provide the smallest crash risk would be deployed.

The effectiveness of the crash prevention module was evaluated based on the relative change in R, TTD, and TTT. The relative change in the three measures is estimated by

$$\Delta M = \frac{\sum_{k=1}^{K} M_{k,CPS} - \sum_{k=1}^{K} M_{k,Non}}{\sum_{k=1}^{K} M_{k,Non}} \times 100\% \quad (7\text{-}5)$$

where ΔM is the percentage of relative change in one measure (i.e., R, or TTT), $M_{k,CPS}$ is the measure of case k with the crash prevention module, and $M_{k,Non}$ is the measure of case k without the crash prevention module.

The testing results showed that among the 113 cases, 104 triggered the crash prevention module and different control strategies were then deployed. Table 7-1 shows the safety and mobility effects by deployed control strategy. It shows that the VSL control strategy was not activated for 59 out of 104 cases as it did not lower the crash risk, so ΔR and ΔTTT remained the same for these cases due to unchanged traffic. The other 45 (37+8) cases yielded improved

safety level with the VSL control being deployed. Specifically, the VSL with 10 MPH drop was deployed for 37 cases and decreased the crash risk by 26.9% while increasing the TTT by only 7.2%; the VSL with 20 MPH drop was deployed for 8 cases and decreased the crash risk by 10.5% while increasing the TTT by only 2.5%. On average, the crash prevention module reduced crash risk by 21.2% for the total 104 cases. Mobility was only slightly compromised, as TTT increased by 5.8%. In general, the proposed CPPA proves to be promising in improving safety without sacrificing mobility.

Table 7-1 Safety and Mobility Effects by Deployed Control Strategy

Control Strategy	**Count**	ΔR	ΔTTT
Non-activated VSL	59	0%	0%
VSL: 10 MPH Drop	37	-26.9%	7.2%
VSL: 20 MPH Drop	8	-10.5%	2.5%
Total	104	-21.2%	5.8%

7.3 Conclusions

A crash prediction and prevention application (CPPA) based on simulated traffic data was proposed to detect crash-prone conditions and help select the desirable TCS for crash prevention. The proposed application was tested in a case study with VSL strategies for demonstration, and results showed that the proposed crash prediction and prevention method could effectively detect crash-prone conditions and evaluate the safety and mobility impacts of various VSL alternatives before their deployment.

The crash prediction and prevention method proposed in this study could be applied in ATIS to detect crash-prone traffic conditions, distribute crash warnings, and evaluate traffic

control countermeasures before their deployment. Further improvements of CTM or equivalent simulation models will help to improve the current method. The CPPA could be applied in the LSCTM environment to detect crash-prone conditions with lane-specific traffic data and evaluate the effectiveness of TCS by lane.

7.4 References

Abdel-Aty, M., Cunningham, R., Gayah, V., & Hsia, L. (2008). Dynamic Variable Speed Limit Strategies for Real-Time Crash Risk Reduction on Freeways. *Transportation Research Record: Journal of the Transportation Research Board, 2078*, 108-116. doi:10.3141/2078-15

Abdel-Aty, M., Dilmore, J., & Dhindsa, A. (2006). Evaluation of variable speed limits for real-time freeway safety improvement. *Accident Analysis & Prevention, 38*(2), 335-345.

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., & Santos, C. D. (2007). Crash Risk Assessment Using Intelligent Transportation Systems Data and Real-Time Intervention Strategies to Improve Safety on Freeways. *Journal of Intelligent Transportation Systems, 11*(3), 107-120. doi:10.1080/15472450701410395

Allaby, P., Hellinga, B., & Bullock, M. (2007). Variable Speed Limits: Safety and Operational Impacts of a Candidate Control Strategy for Freeway Applications. *IEEE Transactions on Intelligent Transportation Systems, 8*(4), 671-680. doi:10.1109/TITS.2007.908562

Choi, S., & Oh, C. (2016). Proactive Strategy for Variable Speed Limit Operations on Freeways Under Foggy Weather Conditions. *Transportation Research Record: Journal of the Transportation Research Board, 2551*, 29-36. doi:10.3141/2551-04

Clark, S. (2003). Traffic prediction using multivariate nonparametric regression. *Journal of transportation engineering, 129*(2), 161-168.

Habtemichael, F. G., & Cetin, M. (2016). Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C: Emerging Technologies, 66*, 61-78.

Hellinga, B., & Mandelzys, M. (2011). Impact of Driver Compliance on the Safety and Operational Impacts of Freeway Variable Speed Limit Systems. *Journal of transportation engineering, 137*(4), 260-268. doi:10.1061/(ASCE)TE.1943-5436.0000214

Lee, C., & Abdel-Aty, M. (2008). Testing effects of warning messages and variable speed limits on driver behavior using driving simulator. *Transportation Research Record: Journal of the Transportation Research Board*(2069), 55-64.

Lee, C., Hellinga, B., & Saccomanno, F. (2006). Evaluation of variable speed limits to improve traffic safety. *Transportation Research Part C: Emerging Technologies, 14*(3), 213-228.

Li, Z., Li, Y., Liu, P., Wang, W., & Xu, C. (2014). Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers. *Accident Analysis & Prevention, 72*, 134-145. doi:10.1016/j.aap.2014.06.018

Li, Z., Liu, P., Wang, W., & Xu, C. (2014). Development of a Control Strategy of Variable Speed Limits to Reduce Rear-End Collision Risks Near Freeway Recurrent Bottlenecks.

IEEE Transactions on Intelligent Transportation Systems, 15(2), 866-877. doi:10.1109/TITS.2013.2293199

Li, Z., Liu, P., Xu, C., & Wang, W. (2016). Optimal Mainline Variable Speed Limit Control to Improve Safety on Large-Scale Freeway Segments: Optimal mainline variable speed limit. *Computer-Aided Civil and Infrastructure Engineering, 31*(5), 366-380. doi:10.1111/mice.12164

Oswald, R. K., Scherer, W. T., & Smith, B. L. (2001). Traffic flow forecasting using approximate nearest neighbor nonparametric regression. *Center for Transportation Studies, University of Virginia*.

Smith, B. L., Williams, B. M., & Oswald, R. K. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies, 10*(4), 303-321.

CHAPTER 8 CONCLUSIONS, CONTRIBUTIONS AND FUTURE RESEARCH

Crashes can be accurately predicted through reliable data sources and rigorous statistical models; and prevented through data-driven, evidence-based traffic control strategies. Both predictive analysis and analysis on causal effects of traffic variables of real-time crashes are instrumental to crash prediction and a better understanding of the mechanism of crash occurrence. However, the research on the latter analysis is very limited for real-time crash prediction; and the conventional predictive analysis using inductive loop detector data has accuracy issues related to inconsistently and distantly spaced loop detectors. The effectiveness of traffic control strategies for improving safety performance cannot be measured and compared without an appropriate traffic simulation application. This dissertation is an attempt to address these research gaps.

Chapter 3 of the dissertation conducts the analysis to assess the causal effect of speed variation on crash occurrence using the crash data and ILD data on a 4.15-mile long corridor on I-94 East in Wisconsin in 2012-2014. As a rigorous analysis method to estimate the causal effect, the propensity score based model is applied to generate samples with similar covariate distributions in both high- and low-speed variation groups of all cases. Under this setting, the confounding effects are removed, and the causal effect of speed variation can be obtained. Upstream and downstream speed variations are first converted into binary treatments − high upstream speed variation (HUSV) and high downstream speed variation (HDSV) − based on cutoff values. Then, the selected variables are included in the propensity model for treated and control groups. The propensity score for each case is estimated based on the propensity model. A weighted sample is generated using the inverse probability of treatment weighting (IPTW) method, from which the causal effect of HUSV and HDSV between the treated and control group

can be impartially estimated. Sensitivity analysis on the cutoff value of speed variation has been performed to test the consistency of the findings. The results show that the causal effect of neither treatment is significant. Hence, it is concluded with high confidence that speed variation is not one of the causes for a crash. In the future, the propensity score based analysis can be extended to other real-time traffic variables and environmental factors, from which crash causation prediction models can be developed.

Chapter 4 conducts a predictive analysis on lane-changing related crashes using lane-specific traffic data collected from three ILD stations near a crash location. The real-time traffic data for the two lanes – the vehicle's lane (subject lane) and the lane to which that a vehicle intends to change (target lane) – are more closely related with lane-change related crashes, as opposed to congregated traffic data for all lanes. Lane-change related crash data are obtained from a 62-mile long freeway in Wisconsin in 2012 and 2013. One-minute traffic data from the 5 to 10-minute interval prior to the crashes are extracted from an immediate upstream detector station and two immediate downstream stations from the crash location. Weather information is collected from a major historical weather database. A matched case-control logistic regression is used for analysis. It is found that the following factors significantly affect the probability of a lane-change related crash, including average flow into the target lane at the first downstream station, the flow ratio at the second downstream station, and snow condition. The average speed in the target lane at the first downstream station also contributes to the occurrence of lane-change crashes during snowy conditions.

Chapter 5 conducts a predictive analysis on real-time crashes using simulated traffic data. The purpose of using simulated traffic data rather than real data is to mitigate the temporal and spatial issues of detector data. Crash cases and non-crash cases are collected from a 4.15-mile

long corridor on I-94 in Wisconsin in 2012-2014. The cell transmission model (CTM), a macroscopic simulation model, is employed to instrument the corridor with a uniform and close layout of virtual detector stations that measure traffic data when physical stations are not available. Traffic flow characteristics at the crash site are simulated by CTM 0-5 minutes prior to a crash. Crash prediction models are developed using the binary logistic regression with traffic flow characteristics of simulated traffic data. As a comparison, crash models are developed from physical detectors. The prediction performance of several crash prediction models shows that the simulated traffic data can improve the prediction performance by accounting for the spatial-tempo issue of ILD data. It is also found that the 0.2-mi distance setting is better than the 0.5-mi distance setting for collecting simulated traffic data regarding the distance from the cell location to its virtual upstream/downstream stations. The crash prediction method can be used for detecting crash-prone conditions where the predicted crash probability exceeds a predetermined threshold value.

Chapter 6 presents a novel approach to modeling freeway crashes using lane-specific simulated traffic data. The new model can not only account for the spatio-temporal issues of detector data but also account for heterogeneous traffic conditions across lanes using a lane-specific cell transmission model (LSCTM). The LSCTM illustrates both discretionary lane-changing (DLC) and mandatory lane-changing (MLC) activities. A case study is performed to demonstrate the method for modeling freeway crashes. Although the models developed from the simulated traffic does not outperform the models with actual traffic data, this new approach presents a viable alternative for utilizing traffic simulation models for safety analysis and evaluation. It is worth noting that the challenge of using traffic simulation lies in model calibration. Uncalibrated parameters of LSCTM are different from actual field values, rendering

inaccurate artificial data. Nevertheless, crash prediction performance will improve despite these challenges when traffic simulation produces more realistic traffic data. Future research will be focused on identifying the efficient and effective ways to calibrate traffic parameters in a LSCTM that are essential to the improvement of RTCPM.

Chapter 7 develops a crash prediction and prevention application (CPPA) based on simulated traffic data to detect crash-prone conditions and to help select the desirable traffic control strategies for crash prevention. The proposed application is tested in a case study with VSL strategies, and results show that the proposed crash prediction and prevention method could effectively detect crash-prone conditions and evaluate the safety and mobility impacts of various VSL alternatives before their deployment. In the future, the application will be more user-friendly and can provide both online traffic operations support as well as offline evaluation of various traffic control operations and methods.

The contributions of this dissertation are summarized as follows:

1. This dissertation demonstrates the need for a propensity score based analysis to obtain causal effects of real-time traffic factors on crash occurrence. Based on the causal effects of traffic factors, more effective countermeasures can be deployed to mitigate crash risk.
2. The dissertation identifies the crash-prone traffic patterns related to lane-changing crashes. The identified crash contributing factors can help traffic operators select traffic control and management countermeasures to proactively mitigate lane-change related crashes.
3. This dissertation identifies the spatial-tempo issues of ILD data and proposes a macroscopic traffic simulation model − a cell transmission model − to generate traffic

data for developing RTCPMs. RTCPMs developed from the simulated traffic data have consistent and comparable performance with different ILD station layouts.

4. This dissertation proposes a LSCTM to provide lane-specific simulated traffic data for crash modeling. The availability of lane-specific traffic characteristics offers new opportunities for modeling more specific crash types such as lane-changing crashes, and evaluating active traffic management (e.g., managed lanes, smart lanes).
5. A crash prediction and prevention application is proposed. This application can monitor real-time crash risk and evaluate traffic control strategies. Agencies can apply this application to detect crash-prone traffic conditions, distribute crash warnings, and evaluate traffic control countermeasures before their deployment.

CURRICULUM VITAE

Zhi Chen

Education

University of Wisconsin-Milwaukee Milwaukee, WI
Ph.D., Civil and Environmental Engineering, GPA: 3.96/4.00 Aug 2018
- Dissertation: Advanced Quantitative Methods for Imminent Detection of Crash Prone Conditions and Safety Evaluation

South Dakota State University Brookings, SD
M.S., Transportation Engineering in Civil Engineering, GPA: 3.79/4.00 Jan 2015
- Thesis: Highway Safety Manual Modification on Intersection Crashes and Multivariate Model Development of Crash Types

Southeast University Nanjing, China
B.S., Transportation Engineering in Civil Engineering, GPA: 3.40/4.00 Jun 2012

Research Experience

University of Wisconsin-Milwaukee 2015-Present
Graduate Researcher
- Performed novel analysis to model crashes with real-time simulated traffic data to account for issues in traditional real-time crash prediction studies and involved handling 12 million traffic flow records
- Developed crash count prediction models for rural two-lane two-way segments in Wisconsin incorporating spatial effects
- Investigated the intrinsic and indirect relationships between contributing factors and crash outcomes by incorporating the knowledge of crash mechanism into crash modeling using structural equation modeling (SEM)
- Proposed and evaluated improved validity tests on the inductive loop detector traffic data
- Composed a literature review of statistical models for the crash count and severity prediction based on over 290 studies
- Composed a literature review of validity tests on inductive loop detector traffic data
- Designed a survey and summarized 54 respondents' applications of and opinions on the Wisconsin WisTransportal V-SPOC data (V-SPOC system records the traffic data collected from all inductive loop detectors in Wisconsin)

South Dakota State University 2012-2015
Graduate Researcher
- Modeled lane-change related crashes using lane-specific real-time traffic data and identified contributing factors related with traffic on the two lanes involved in lane-change activities
- Developed a multivariate Poisson-lognormal model to model multiple crash types simultaneously

- Calibrated Highway Safety Manual (HSM) predictive models for all intersection types in South Dakota
- Calibrated Highway Safety Manual (HSM) predictive models for and evaluated roadway designs of local and tribal two-way two-lane roadways in South Dakota
- Provided guidance on how to calibrate Interactive Highway Safety Design Model (IHSDM) models for South Dakota Department of Transportation

Projects

University of Wisconsin-Milwaukee 2015-Present
Graduate Researcher

- Lead Researcher, Developing Faulty Loop Detection and Diagnosis Tools for Improving V-SPOC Data Quality, 05/2017 – present
- Assistant Researcher, Identifying Highly Correlated Variables Relating to the Potential Causes of Reportable Wisconsin Traffic Crashes, 03/2017 – 07/2017

South Dakota State University 2012-2015
Graduate Researcher

- Lead Researcher, Calibration of Highway Safety Manual Predictive Methods for State and Local Rural Highways, 08/2013 – 01/2015
- Lead Researcher, Evaluating Local and Tribal Rural Road Design with Interactive Highway Safety Design Model (IHSDM), 10/2012 – 09/2013

Teaching Experience

University of Wisconsin-Milwaukee 09/2016 – 05/2017
Graduate Teaching Assistant

- Worked as a teaching assistant for CE 280: Computer Based Engineering Analysis for 82 undergraduate students
- Presented 3 lectures, designed materials for student labs and held weekly office hours
- Graded homework, lab assignments and exams

Publications

1. **Chen, Zhi**, Xiao Qin. “A Novel Method for Imminent Crash Prediction and Prevention.” Accident Analysis & Prevention (2018).
2. Qin, Xiao, **Zhi Chen**, and M. Razaur Rahman Shaon. "Developing jurisdiction-specific SPFs and crash severity portion functions for rural two-lane, two-way intersections." Journal of Transportation Safety & Security (2018).
3. **Chen, Zhi**, Xiao Qin, Renxin Zhong, Pan Liu, Yang Cheng. "Predicting Imminent Crash Risk with Simulated Traffic from Distant Sensors." Transportation Research Record: Journal of the Transportation Research Board (forthcoming, 2018).
4. Shaon, M. Razaur Rahman, Xiao Qin, and **Zhi Chen**. " An Exploration of Contributing Factors Related to Driver Errors on Highway Segments." Transportation Research Record: Journal of the Transportation Research Board (forthcoming, 2018).

5. **Chen, Zhi**, Xiao Qin, and M. Razaur Rahman Shaon. "Modeling Lane-change Related Crashes with Lane-specific Real-time Traffic and Weather Data." Journal of Intelligent Transportation Systems (2017).
6. Qin, Xiao, M. Razaur Rahman Shaon, and **Zhi Chen**. "Developing Analytical Procedures for Calibrating the Highway Safety Manual Predictive Methods." Transportation Research Record: Journal of the Transportation Research Board 2583 (2016): 91-98.

Conference Papers

1. **Chen, Zhi**, Xiao Qin, and M. Razaur Rahman Shaon. "A Spatial Hurdle Crash Prediction Model: Analyzing the Correlation Between Dual States." Presented at Transportation Research Board 97th Annual Meeting, Washington D.C. 2018.
2. **Chen, Zhi**, Xiao Qin. "Using Simulated Traffic Data for Real-Time Crash Prediction." Presented at Transportation Research Board 96th Annual Meeting, Washington D.C. 2017.
3. Qin, Xiao, **Chen Zhi**, and Kimberly Vachal. "Calibration of Highway Safety Manual Predictive Methods for Rural Local Roads." Transportation Research Board 93rd Annual Meeting. No. 14-1053. 2014.