2019

Arterial-level Real-time Safety Evaluation in the Context of Proactive Traffic Management

Jinghui Yuan
University of Central Florida

ARTERIAL-LEVEL REAL-TIME SAFETY EVALUATION IN THE CONTEXT OF PROACTIVE TRAFFIC MANAGEMENT

by

JINGHUI YUAN

B.S., Central South University, China, 2013

M.S., Tongji University, China, 2016

A dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

in the Department of Civil, Environmental and Construction Engineering

in the College of Engineering and Computer Science

at the University of Central Florida

Orlando, Florida

Summer Term

2019

Major Professor: Mohamed Abdel-Aty

ABSTRACT

In the context of pro-active traffic management, real-time safety evaluation is one of the most important components. Previous studies on real-time safety analysis mainly focused on freeways, seldom on arterials. With the advancement of sensing technologies and smart city initiative, more and more real-time traffic data sources are available on arterials, which enables us to evaluate the real-time crash risk on arterials. However, there exist substantial differences between arterials and freeways in terms of traffic flow characteristics, data availability, and even crash mechanism. Therefore, this study aims to deeply evaluate the real-time crash risk on arterials from multiple aspects by integrating all kinds of available data sources. First, Bayesian conditional logistic models (BCL) were developed to examine the relationship between crash occurrence on arterial segments and real-time traffic and signal timing characteristics by incorporating the Bluetooth, adaptive signal control, and weather data, which were extracted from four urban arterials in Central Florida. Second, real-time intersection-approach-level crash risk was investigated by considering the effects of real-time traffic, signal timing, and weather characteristics based on 23 signalized intersections in Orange County. Third, a deep learning algorithm for real-time crash risk prediction at signalized intersections was proposed based on Long Short-Term Memory (LSTM) and Synthetic Minority Over-Sampling Technique (SMOTE). Moreover, in-depth cycle-level real-time crash risk at signalized intersections was explored based on high-resolution event-based data (i.e., Automated Traffic Signal Performance Measures (ATSPM)). All the possible real-time cycle-level factors were considered, including traffic volume, signal timing, headway and occupancy, traffic variation between upstream and downstream detectors, shockwave characteristics, and weather conditions. Above all, comprehensive real-time safety evaluation algorithms were

developed for arterials, which would be key components for future real-time safety applications (e.g., real-time crash risk prediction and visualization system) in the context of pro-active traffic management.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor Dr. Mohamed Abdel-Aty, for his unwavering support and guidance throughout my research work at UCF. I have been extremely lucky to have an advisor who cared so much about my work and life. I would also like to extend my sincere thanks to the committee members, Dr. Naveen Eluru, Dr. Samiul Hasan, Dr. Liqiang Wang, and Dr. Qing Cai for their valuable advice and suggestions. In addition, I'd like to acknowledge Dr. Ling Wang, Dr. Jaeyoung Lee, and Dr. Yina Wu for their insightful suggestions. Also, I gratefully acknowledge the help of all my colleagues. It was fantastic to have the opportunity to work with you, especially the three gentlemen in ENG2-215.

Special thanks to my wife for her relentless support. Without her love and encouragements, I could achieve nothing. I am also grateful to my parents for their unparalleled support and encouragements.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

AADT	Annual Average Daily Traffic
ADT	Average Daily Traffic
AFR	Average Flow Ratio
AOG	Arrive on Green
AOGR	Arrival on Green Ratio
AOY	Arrive on Yellow
AOYR	Arrive on Yellow Ratio
ATSPM	Automated Traffic Signal Performance Measures
AUC	Area under ROC Curve
AVI	Automatic Vehicle Identification
BCI	Bayesian Confidential Interval
BCL	Bayesian Conditional Logistic
BGR	Brooks-Gelman-Rubin
BRPCL	Bayesian Random Parameters Conditional Logistic
BRPL	Bayesian Random Parameters Logistic
DIC	Deviance Information Criterion
DMS	Dynamic Message Sign
FARS	Fatality Analysis Reporting System
FCD	Floating Car Data
FPR	False Positive Rate

GPS	Global Positioning System
GPU	Graphic Processing Unit
LCD	Local Climatological Data
LD	Loop and Radar Detector
LOS	Level of Service
LSTM	Long Short-Term Memory
MCMC	Markov Chain Monte Carlo
MIC	Maximal Information Coefficient
MVDS	Microwave Vehicle Detection System
NCDC	National Climate Data Center
NOAA	National Oceanic Atmospheric Administration
OAFR	Overall Average Flow Ratio
PDO	Property Damage Only
POG	Percent of Green
POY	Percent of Yellow
PR	Platoon Ratio
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RTMS	Remote Traffic Microwave Sensor
S4A	Signal Four Analytics
SMOTE	Synthetic Minority Over-Sampling Technique

TPR	True Positive Rate
UAV	Unmanned Aerial Vehicle
V2I	Vehicle to Infrastructure
VSL	Variable Speed Limit

CHAPTER 1: INTRODUCTION

1.1 Overview

Urban arterials play a critical role in the road network system as they provide the high-capacity network for travel within urban areas. Meanwhile, urban arterials are suffering from serious traffic safety issues. Take Florida as an example, over 51% of crashes have occurred on urban arterials in 2014. Substantial efforts have been made by previous researchers to reveal the relationship between crash frequency on urban arterials and all the possible contributing factors such as roadway geometric, traffic characteristics, etc. (El-Basyouny and Sayed, 2009; Gomes, 2013; Greibe, 2003; Wang et al., 2015b). However, these studies were conducted based on static and highly aggregated data (e.g., Annual Average Daily Traffic (AADT), annual crash frequency). These aggregated data limit the reliability of the study findings simply because they are averages and cannot reflect the real conditions at the time of crash occurrence.

With the rapid development of traffic surveillance system and detection technologies, real-time traffic data are not only available on freeways and expressways but also on urban arterials (including road segments and intersections). During the past decade, an increasing number of studies have investigated the crash likelihood on freeways by using real-time traffic and weather data (Abdel-Aty et al., 2004; Abdel-Aty et al., 2012; Ahmed et al., 2012a; Lee et al., 2003; Oh et al., 2001; Xu et al., 2013a; Xu et al., 2013b; Yu and Abdel-Aty, 2014; Yu et al., 2014; Zheng et al., 2010). However, little research has been conducted on the real-time safety analysis of urban arterials (Mussone et al., 2017; Theofilatos, 2017; Theofilatos et al., 2017; Yuan et al., 2018a). Among those research, they are mainly focused on the road segments (Theofilatos, 2017; Theofilatos et al., 2017; Yuan et al., 2018a) rather than the signalized intersections (Mussone et al., 2017).

Theofilatos (2017) was the first to investigate crash likelihood and severity by exploiting real-time traffic and weather data collected from urban arterials. He found that both the variation in occupancy and logarithm of the coefficient of variation of flow are positively associated with crash occurrence. However, the traffic parameters were aggregated to 1-hour interval, which might be too large to capture the short-term traffic status prior to crash occurrence. In terms of signalized intersection, Mussone et al. (2017) examined the factors which may affect the crash severity level at intersections based on real-time traffic flow and environmental characteristics, and they found that the real-time traffic flow characteristics have a relevant role in the prediction of crash severity. However, they didn't consider the crash likelihood at intersections, which means that the effects of real-time traffic flow and environmental characteristics on the crash likelihood at intersections are still unclear. Moreover, it is worth noting that the crash risk of urban arterials might be highly influenced by signal operation, which has never been examined in real-time safety analysis.

In recent years, with the rapid development of connected vehicle technologies, it is feasible for us to implement highly efficient real-time proactive traffic management strategies on urban arterials, e.g., dynamic message sign to show the real-time crash risk for the downstream segments and intersections, and individual variable speed limit to provide driver with the optimal speed advisory through vehicle-to-infrastructure (V2I) communication. In this context, more efficient and reliable real-time crash risk predictive algorithms for arterials are required.

1.2 Data Summary

In general, there are five types of data sources that were utilized in this study.

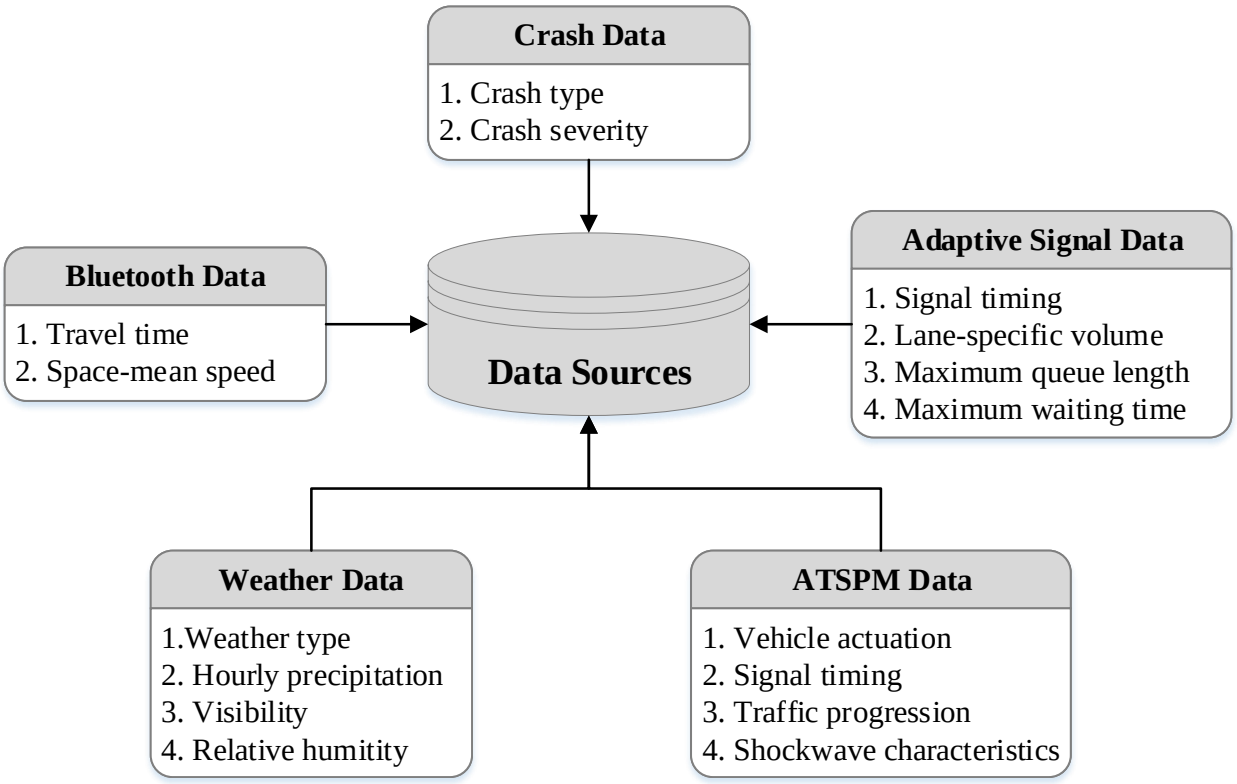


Figure 1-1: Available Data Sources

1) **Crash Data:** Signal Four Analytics (S4A) provides detailed crash information, including crash time, coordinates, severity, type, weather condition, etc.
2) **Bluetooth Data:** Bluetooth data provides the *travel time* and *space-mean speed* of the detected vehicle for each segment. Bluetooth detectors can only detect the vehicles equipped with Bluetooth device which is working at discoverable mode.

   The *space-mean speed* of each vehicle on a specific segment is calculated as the segment length divided by the travel time of each detected vehicle on the segment based on the detection data of two Bluetooth detectors located at the two contiguous intersections.
3) **Adaptive Signal Data:** The adaptive signal control system at signalized intersection is operated based on the video detectors installed on the approaches, which can detect the real-time queue length, maximum waiting time, and traffic volume by movement. This

system archives the *real-time signal phasing*, *queue length*, *waiting time*, and *15-minute aggregated lane-specific traffic volume* data.

4) **Automated Traffic Signal Performance Measures (ATSPM) Data:** ATSPM data archives all kinds of event data generated by signal controllers and loop detectors installed at intersections in a very high resolution (0.1 second). Every event generated by signal controllers or loop detectors is recorded in sets of four bytes per event: two bytes for the timestamp of when the event occurred, one byte for event code type, and one byte for event parameter (for signifying detector numbers and phases). The event code is important for determining the type of reported activity, which could be phase initiation or termination, detection on/off, etc. Based on the ATSPM data, both *signal timing* and *lane-specific vehicle count* variables could be calculated. Also, the real-time *shockwave characteristics* could be estimated based on the detector activation and signal controller events.
5) **Weather Data:** Weather data were collected from the National Climate Data Center (NCDC), which archives weather data from nationwide weather stations operated by the National Oceanic Atmospheric Administration (NOAA). In this study, four weather related variables (*weather type, hourly precipitation, visibility, and relative humidity*) were collected from the nearest airport weather station.

1.3 Research Objectives

The primary objective of this dissertation is to evaluate the real-time crash risk on arterials by incorporating all the available data sources. In this context, four specific objectives have been achieved in this study:

(1) Investigating the relationship between crash occurrence and real-time traffic, signal timing, and weather characteristics on arterial segments;

(2) Identifying all the possible contributing factors for intersection approach-level real-time crash risk;

(3) Improve the predictive performance of real-time crash risk prediction by utilizing advanced deep learning algorithms and oversampling method;

(4) Modeling real-time crash risk at the cycle-level for signalized intersections with the consideration of shockwave characteristics.

The first objective has been achieved in Chapter 3 by the following sub-tasks:

a) The concept of real-time safety analysis on urban arterials by considering microscopic traffic and signal timing characteristics is demonstrated;

b) Two kinds of new data sources (Bluetooth and adaptive signal control data) are introduced to real-time safety analysis;

c) Bayesian random parameters logistic (BRPL) and Bayesian random parameters conditional logistic models (BRPCL) are developed to compare with the Bayesian conditional logistic model (BCL);

d) The relationships between real-time crash occurrence and real-time traffic and signal characteristics on urban arterials are preliminarily revealed.

The second objective has been achieved in Chapter 4 by the following sub-tasks:

e) Examining the real-time crash risk at signalized intersections based on the disaggregated data from multiple sources;

f) Intersection and intersection-related crashes were collected and then divided into two types, i.e., within intersection crashes and intersection entrance crashes. Bayesian conditional logistic models were developed for these two kinds of crashes;
g) Matched case-control design with a control-to-case ratio of 4:1 was employed to select the corresponding non-crash events for each crash event.

The third objective has been achieved in Chapter 5 by the following sub-tasks:

h) Predicting the real-time crash risk at signalized intersections by using multilayer LSTM recurrent neural network, which is designed for sequence modeling, and they can consider the time series characteristics automatically;
i) Real-world unbalanced dataset was collected for every minute by incorporating real-time traffic, signal, and weather data. Also, both the approach-level and intersection-level geometric characteristics were included into the algorithm;
j) To train the algorithm without losing any non-crash information, the synthetic minority over-sampling technique (SMOTE) was employed in this study to generate a balanced training dataset. In comparison, a traditional conditional logistic model was developed based on the matched case-control dataset with the control-to-case ratio of 10:1.

The fourth objective has been achieved in Chapter 6 by the following sub-tasks:

k) Identifying the exact signal cycle where every crash has occurred based on the high-resolution event based ATSPM dataset;
l) modeling real-time crash risk at the cycle-level for signalized intersections with the consideration of shockwave characteristics;

m) determining the best undersampling strategy while calibrating real-time crash risk prediction models for signalized intersections.

1.4 Dissertation Organization

The dissertation is organized as follows: Chapter 2 presents a thorough literature review, which includes aggregated arterial safety analysis, real-time crash risk analysis and vehicle/driver-level crash risk evaluation. Chapter 3 investigates the real-time crash risk on arterial segments by utilizing multiple data sources. Followed by chapter 4, where approach-level real-time crash risk was evaluated for signalized intersections. Chapter 5 proposes a deep learning algorithm for real-time crash risk prediction at signalized intersections based on Long Short-Term Memory (LSTM) and Synthetic Minority Over-Sampling Technique (SMOTE). Chapter 6 reveals the relationship between real-time crash occurrences and cycle-level characteristics at signalized intersection approaches. Chapter 7 summarizes the overall dissertation and proposes a set of recommendations for future studies.

CHAPTER 2: LITERATURE REVIEW

2.1 Aggregated Arterial Safety Analysis

2.1.1 Arterial Segments

A number of studies have explored the effects of various road geometric design and traffic characteristics on arterial safety based on aggregated data. As to road geometric design, high crash frequency was found to be associated with high intersection density (Bonneson and McCoy, 1997; El-Basyouny and Sayed, 2009; Wang and Yuan, 2017; Wang et al., 2016a; Wang et al., 2018) and access density (Bonneson and McCoy, 1997; Wang and Yuan, 2017; Wang et al., 2016a). The number of lanes was found to be positively correlated with crash occurrence (El-Basyouny and Sayed, 2009; Gomes, 2013; Wang et al., 2015b). In addition, an increased segment length (El-Basyouny and Sayed, 2009; Wang et al., 2015b) and decreased lane width (Yanmaz-Tuzel and Ozbay, 2010) tend to increase the crash frequency.

In terms of traffic related contributing factors, traffic volume and travel speed have been found to be significantly associated with the crash frequency on arterials. Traffic volume (represented by AADT, hourly volume, etc.) has been widely demonstrated to be positively correlated with crash frequency (El-Basyouny and Sayed, 2009; Gomes, 2013; Wang et al., 2015b). While the safety effects of travel speed are not consistent among existing studies, many studies suggested that higher average speed tends to increase the crash frequency (Aarts and Van Schagen, 2006; Elvik, 2009; Nilsson, 2004; Taylor et al., 2002), as higher speed increases the drivers' overall

stopping distance which may in turn increase the probability of crash occurrence (Wang et al., 2013). However, some researchers found that the average speed is negatively associated with crash frequency (Baruya, 1998; Stuster, 2004).

Moreover, Pei et al. (2012) evaluated the relationship between speed and crash risk with respect to distance and time exposure, they found that the correlation between speed and crash risk is positive when distance exposure (i.e., vehicle kilometers travelled) is considered, but negative when time exposure (i.e., vehicle hours travelled derived by multiplying traffic volume by average travel time) is used. Wang et al. (2015b) utilized the Floating Car Data (FCD) to calculate average speeds during peak and off-peak hours, and then developed crash prediction models for peak and off-peak separately. The model results indicated that average travel speed was not significantly related to crash frequency during the off-peak period, however, during the peak period, a significant positive relationship between average speed and crash frequency was demonstrated. More recently, Imprialou et al. (2016) proposed a new condition-based approach to aggregate the crashes according to the similarity of their pre-crash traffic and geometric conditions, and then compared it with the traditional segment-based aggregation approach. The results showed that average speed was significantly positively associated with crash occurrence in the condition-based model, while the relationship was found to be negative in the segment-based model. In conclusion, the inconsistent findings of the safety effects of travel speed might be caused by the inaccuracy of data aggregation, as the aggregated data

cannot represent the actual traffic circumstance when the crashes have occurred. At this point, more disaggregated real-time analysis should be conducted for urban arterials to figure out the underlying relationship between crash occurrence and traffic characteristics.

2.1.2 Signalized Intersections

Signalized intersection safety analysis has been a critical research topic during past decades. Substantial efforts have been made by previous researchers to reveal the relationship between crash frequency of signalized intersections and all the possible contributing factors such as roadway geometric, signal control, and traffic characteristics, etc. (Abdel-Aty and Wang, 2006; Chin and Quddus, 2003; Guo et al., 2010; Poch and Mannering, 1996; Wang et al., 2009; Wang et al., 2006). However, these studies were conducted based on static and highly aggregated data (e.g., Annual Average Daily Traffic (AADT), annual crash frequency). These aggregated data limit the reliability of the study findings simply because they are averages and cannot reflect the real conditions at the time of crash occurrence.

More specifically, nearly all the traffic volume related variables were found to have significant positive effects on the crash frequency at signalized intersections, including total entering ADT (Abdel-Aty and Wang, 2006; Chin and Quddus, 2003; Guo et al., 2010; Poch and Mannering, 1996), right-turn ADT (Chin and Quddus, 2003; Poch and Mannering, 1996), left-turn ADT (Poch and Mannering, 1996), total ADT on major road (Dong et al., 2014; Wang et al., 2009), total ADT on minor road (Dong et al., 2014; Wang et al., 2009), left-turn ADT on major road

(Guo et al., 2010), through ADT on minor road (Guo et al., 2010). However, Guo et al. (2010) found that the through ADT on major road and the left-turn ADT on minor road are significantly negatively associated with the crash frequency at signalized intersections. Moreover, Wang et al. (2009) investigated the relationship between LOS and safety at signalized intersections. They found that LOS D is a desirable level which is associated with less total crashes, rear-end and sideswipe crashes, as well as right-angle and left-turn crashes. Xie et al. (2013) investigated the safety effect of corridor-level travel speed, they found that the high-speed corridor may results in more crashes at the signalized intersections. Similarly, the speed limit of the corridor was found to be significantly positively correlated with the crash frequency of the signalized intersections (Abdel-Aty and Wang, 2006; Dong et al., 2014; Guo et al., 2010; Poch and Mannering, 1996; Wang et al., 2009).

With respect to the geometric design, number of lanes, median width, and intersection sight distance et al. were found to have significant effects on the crash frequency of signalized intersections. More specifically, the number of lanes was found to be positively correlated with the crash frequency of signalized intersections (Abdel-Aty and Wang, 2006; Dong et al., 2014; Guo et al., 2010; Poch and Mannering, 1996). Median width and intersection sight distance was also found to have positive effect on the crash frequency(Chin and Quddus, 2003). Moreover, Abdel-Aty and Wang (2006) found that the existence of exclusive right-turn lanes could significantly decrease the crash frequency.

In terms of signal control characteristics, the adaptive signal control was found to have significant lower crash frequency than the pre-timed signal control (Chin and Quddus, 2003). The number of phases was found to be positively associated with the crash frequency of signalized intersections (Chin and Quddus, 2003; Poch and Mannering, 1996; Xie et al., 2013). The left-turn protection could significantly improve the safety performance of the signalized interaction (Abdel-Aty and Wang, 2006; Chin and Quddus, 2003; Poch and Mannering, 1996). However, Abdel-Aty and Wang (2006) found that the left-turn protection on minor roadway tends to increase the crash frequency of signalized intersection. Surprisingly, Guo et al. (2010) found that the coordinated intersections are more unsafe than the isolated ones. They explained it as the travel speed is higher for coordinated intersections because of the green wave, which may result in more crashes.

2.2 Real-time Crash Risk Analysis

Real-time crash risk analysis has been widely adopted to reveal crash occurrence precursors by investigating the differences in traffic conditions between crash and non-crash events. As crash risk analysis is a typical binary classification problem, the most commonly used methods are the matched case-control logistic models (Abdel-Aty and Pande, 2005; Abdel-Aty et al., 2004; Ahmed and Abdel-Aty, 2012; Xu et al., 2012; Zheng et al., 2010), Bayesian logistical models (Ahmed et al., 2012a; Shi and Abdel-Aty, 2015; Wang et al., 2017a; Wang et al., 2015a; Yu et al., 2014), Bayesian random effect logistic models (Shi and Abdel-Aty, 2015; Yu et al., 2016),

Bayesian random parameter logistic models (Shi and Abdel-Aty, 2015; Xu et al., 2014; Yu and Abdel-Aty, 2014; Yu et al., 2017). Besides, several approaches of data mining such as neural networks (Abdel-Aty and Pande, 2005; Abdel-Aty et al., 2008), support vector machines (Yu and Abdel-Aty, 2013; Yu and Abdel-Aty, 2014), and Bayesian networks (Hossain and Muromachi, 2012; Sun and Sun, 2015) were also applied to evaluate the real-time crash risk.

In order to identify the crash-prone conditions, huge efforts have been made to investigate the relationship between real-time crash risk and various traffic parameters and weather-related variables. Generally, the average speed was found to be negatively correlated with crash likelihood (Abdel-Aty et al., 2012; Ahmed et al., 2012a, b; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2016). The speed variation in the form of speed standard deviation or coefficient of speed variation was found to have significant positive effects on crash occurrence (Abdel-Aty et al., 2004; Abdel-Aty et al., 2012; Ahmed et al., 2012a, b; Ahmed and Abdel-Aty, 2012; Xu et al., 2012; Zheng et al., 2010). Intuitively, higher traffic volume contributes to higher crash risk (Roshandel et al., 2015). Moreover, several studies (Hossain and Muromachi, 2012; Shi and Abdel-Aty, 2015) reported that the congestion index is positively correlated with crash occurrence. With respect to weather related variables, adverse weather is usually associated with increased crash risk (Ahmed et al., 2012a; Xu et al., 2013a).

In summary, all the above real-time safety analyses were focused on the freeways, while urban arterials have seldom been analyzed. Theofilatos (2017) was the first to investigate crash likelihood and severity by exploiting real-time traffic and weather data collected from urban arterials. He found that both the variation in occupancy and logarithm of the coefficient of variation of flow are positively associated with crash occurrence. However, the traffic parameters were aggregated to 1-hour interval, which might be too large to capture the short-term traffic status prior to crash occurrence. Moreover, it is worth noting that the crash risk of urban arterials might be highly influenced by signal operation, while it has never been examined in real-time safety analysis.

2.3 Vehicle/Driver-Level Crash Risk Evaluation

Previous real-time crash risk predictions were conducted based on time and space, i.e., predicting the crash risk at a specific location during a specific time period. These predictions were mainly based on the real-time traffic characteristics, and there is no any driving behavior/vehicle kinematic characteristics were incorporated into those prediction. Therefore, the predicted high crash risk information cannot specify to vehicle/driver level. However, different drivers may have different response even toward the same traffic condition. For example, given the same dangerous traffic condition, if all the drivers are conservative/safe driver, there may not have any crash, the potential crash may always cause by the aggressive drivers. Therefore, more accurate warning information at vehicle/driver level might be more

helpful to alleviate the crash risk. Also, as the rapid development on V2I technology, it is possible for us to deliver more accurate information to a specific driver/vehicle in the future. Above all, a vehicle/driver level real-time crash risk prediction algorithm should be proposed to incorporate the real-time traffic and the antecedent vehicle kinematic characteristics (driving behavior) simultaneously.

In recent years, there are many studies tried to evaluate driving behavior by using different indicators, e.g., acceleration, braking events, lateral acceleration, yaw rate, and speed profile etc. Jun et al. (2011) evaluated the differences in observed speed patterns between crash-involved and crash-not-involved drivers through various potential speed metrics created from longitudinally-measured GPS speed data. They found that at most times, drivers who had crash experiences tended to drive at higher speeds than crash-not-involved drivers except in freeway travels during AM peak hours. Crash-involved drivers also showed higher tendencies of non-compliance with the posted speed limit.

Bagdadi (2013) developed a new method based on critical jerk to identify safety critical braking events during car driving based on 637 near crashes extracted from the VTTI naturalistic driving dataset, and then compared it with the longitudinal acceleration measure. The findings show that the critical jerk method performed approximately 1.6 times higher overall success rate than the longitudinal acceleration measure. In addition, a positive correlation was found between driver's safety critical braking event and crash involvement.

Simons-Morton et al. (2013) defined a kinematic risky driving when they exceeded the following thresholds: longitudinal deceleration/hard braking ($\leq -0.45g$), longitudinal acceleration/rapid starts ($\geq 0.35g$), lateral negative/left turn ($\leq -0.50g$) and lateral positive/right turn ($\geq 0.50g$) acceleration, and yaw rate (± 6 degrees per second). They found that the kinematic risky driving was best characterized as two classes, a higher-risk and a lower-risk class.

Wang et al. (2015c) proposed a new measure, i.e., driving volatility score, which was defined as the percentage of time when the driver's acceleration or vehicular jerk goes beyond the typical driving thresholds (acceleration or vehicular jerk bands). They found that younger drivers exhibit higher volatility in driving, and ten-year increase in driver age is associated with a decrease of 0.57 in volatility scores.

Eboli et al. (2016) proposed a methodology for analyzing driving behavior by considering kinematic parameters such as speed and longitudinal and lateral accelerations as the elements that can best explain if driver adopts a safe driving or not. They proposed a theoretical domain to distinguish the safe driving and unsafe driving, and then validated by the real test data on a rural two-lane road in Southern Italy.

Zhu et al. (2017) employed a Bayesian Network model to investigate the relationships between GPS driving observations, individual driving behavior, individual driving risks, and individual crash frequency. They incorporated the contextual features, such as road conditions

surrounding the vehicle of interest and dynamic traffic flow information, as well as the non-contextual data such as instantaneous driving speed and the acceleration/deceleration of a vehicle. The findings indicate that drivers who drive at a speed faster than others or much slower than the speed limit at the ramp, and with more rapid acceleration or deceleration on freeways are more likely to be involved in crash events.

2.4 Summary

Substantial efforts have been made by previous researchers to reveal the relationship between crash frequency on urban arterials and all the possible contributing factors. However, these studies were conducted based on static and highly aggregated data, which may limit the reliability of the findings simply because they are averages and cannot reflect the real conditions at the time of crash occurrence. Also, most of the previous real-time studies have been applied to freeways and seldom to arterials. Those bare studies on the real-time safety analysis on urban arterials were based on one-hour aggregated traffic parameters prior to crash occurrence, which might not be truly “real-time” as the traffic flow are likely to differ within one hour.

Moreover, there is no previous research that focused on real-time safety analysis at signalized intersections. The conflicting traffic movements at signalized intersection are temporally

separated by traffic signals, and signal timing plays a very important role in the intersection safety, especially when the adaptive signal control technology was widely adopted on major urban arterials. However, the safety effect of real-time signal status has never been considered, while improper signal timing may result in dangerous situation. Therefore, the relationship between real-time signal timing and intersection safety need to be further investigated.

It is worth noting that the previous real-time crash risk prediction models were evaluated based on artificially balanced test data, while these evaluation results can hardly represent the prediction performance in real-world application. Also, no research has been conducted for real-time crash risk prediction by using LSTMs, which were proved to have very good performance on a large variety of time series sequence learning problems. However, real-time crash risk prediction is a typical time series related sequential prediction process, and the impacts of long-term and short-term traffic data might be quite different, which could be captured by LSTM efficiently.

Furthermore, considering that the traffic flow at signalized intersections presents cyclical characteristics, which are temporally interrupted by signal timing. Therefore, the data preparation for real-time crash risk prediction at signalized intersections should be based on the signal cycle rather than a predefined fixed time interval (i.e., 5 minutes), and cycle-level real-time crash risk analysis should be conducted while considering the cyclical characteristics of the traffic flow at signalized intersections.

CHAPTER 3: UTILIZING BLUETOOTH AND ADAPTIVE SIGNAL CONTROL DATA FOR URBAN ARTERIALS SAFETY ANALYSIS[1]

3.1 Introduction

Urban arterials play a critical role in the road network system as they provide the high-capacity network for travel within urban areas as well as the access to roadside activities. Meanwhile, urban arterials are suffering from serious traffic safety issues. Take Florida as an example, over 51% of crashes occurred on urban arterials in 2014. Substantial efforts have been made by previous researchers to reveal the relationship between crash frequency on urban arterials and all the possible contributing factors such as roadway geometric, traffic characteristics, etc. (El-Basyouny and Sayed, 2009; Gomes, 2013; Greibe, 2003; Wang et al., 2015b). However, these studies were conducted based on static and highly aggregated data (e.g., Annual Average Daily Traffic (AADT), annual crash frequency).

Recently, an increasing number of studies investigated the crash likelihood on freeways by using real-time traffic and weather data (Abdel-Aty et al., 2004; Abdel-Aty et al., 2012; Ahmed et al., 2012a; Lee et al., 2003; Oh et al., 2001; Xu et al., 2013a; Xu et al., 2013b; Yu and Abdel-Aty, 2014; Yu et al., 2014; Zheng et al., 2010). However, little research has been conducted on real-time safety analysis of urban arterials (Theofilatos, 2017; Theofilatos et al., 2017; Yuan and Abdel-Aty, 2018), although the real-time traffic and weather data are available on many

[1] This chapter has been published in Transportation Research Part C (https://doi.org/10.1016/j.trc.2018.10.009)

major arterials. This might be due to the substantial difference in traffic flow characteristics, data availability, and even crash mechanism between urban arterials and freeways, thus it is inappropriate to simply transfer the same research framework from freeways to urban arterials. More specifically, the interrupted traffic flow on urban arterials is highly influenced by the traffic signals (Cai et al., 2014; Wang et al., 2017b), which is quite different from the uninterrupted flow on freeways. Therefore, the crash risk on urban arterials might be associated with not only real-time traffic flow characteristics but also real-time signal timing, which has not been considered in previous research (Theofilatos, 2017; Theofilatos et al., 2017). Moreover, those pioneering studies on the real-time safety analysis of urban arterials were based on one-hour aggregated traffic parameters prior to crash occurrence, which might not be truly "real-time" as the traffic flow are likely to differ within one hour.

In terms of real-time traffic data, most of the previous studies were based on inductive loop detectors (ILDs) (Abdel-Aty et al., 2008; Abdel-Aty et al., 2012; Zheng et al., 2010). ILDs are the most commonly used sensors in traffic management, however, there are some inherent problems with it, such as high failure rates and difficulty with maintenance, especially for arterials. Recently, several studies tried to conduct real-time safety analysis for freeways based on the traffic data collected from nonintrusive detectors, such as automatic vehicle identification system (AVI) (Ahmed et al., 2012a, b; Ahmed and Abdel-Aty, 2012) and remote traffic microwave sensor (RTMS) (Ahmed and Abdel-Aty, 2013; Shi and Abdel-Aty, 2015).

AVI is used mainly for toll collection and travel time estimation while RTMS is mostly used for operation and incident management. The speed data collected from different detectors are quite different, AVI and Bluetooth detectors measure space mean speed, whereas RTMS and ILDs measure time mean speed. As to the data availability, AVI and RTMS are usually available on freeways, and the possible available real-time traffic data on urban arterials are ILDs, Bluetooth, and floating car data (FCD). To the best of our knowledge, there is no real-time safety analysis has been carried out using traffic data from Bluetooth detectors.

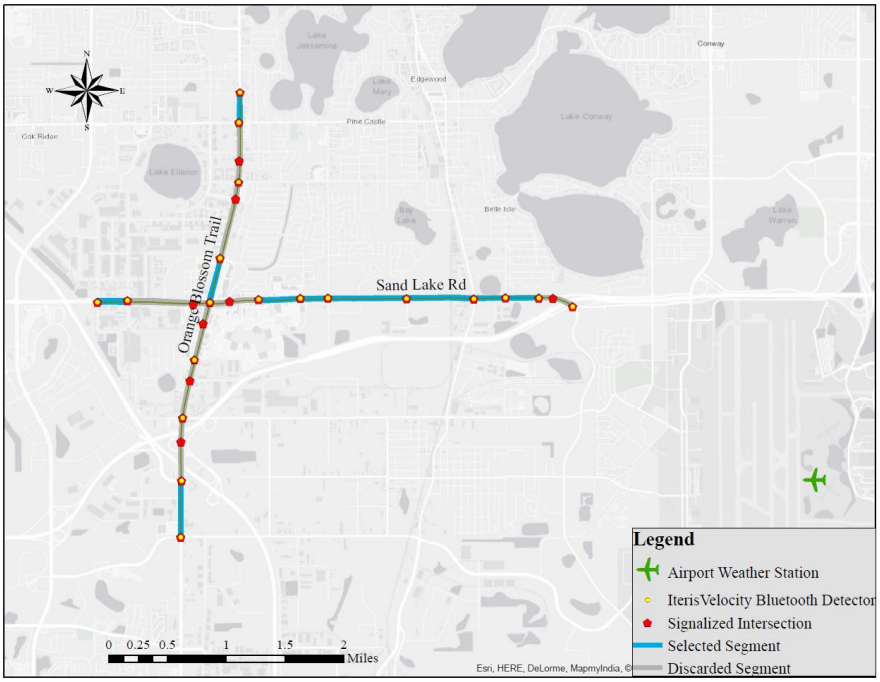
Above all, this study aims to investigate the relationship between crash occurrence on urban arterials and real-time traffic, signal timing, and weather characteristics by utilizing data from multiple sources, i.e., Bluetooth, weather, and adaptive signal control datasets. The main contributions of this chapter include:

(1) The concept of real-time safety analysis on urban arterials by considering microscopic traffic and signal timing characteristics is demonstrated.
(2) Two kinds of new data sources (Bluetooth and adaptive signal control data) are introduced to real-time safety analysis.
(3) Bayesian random parameters logistic (BRPL) and Bayesian random parameters conditional logistic models (BRPCL) are developed to compare with the Bayesian conditional logistic model (BCL).

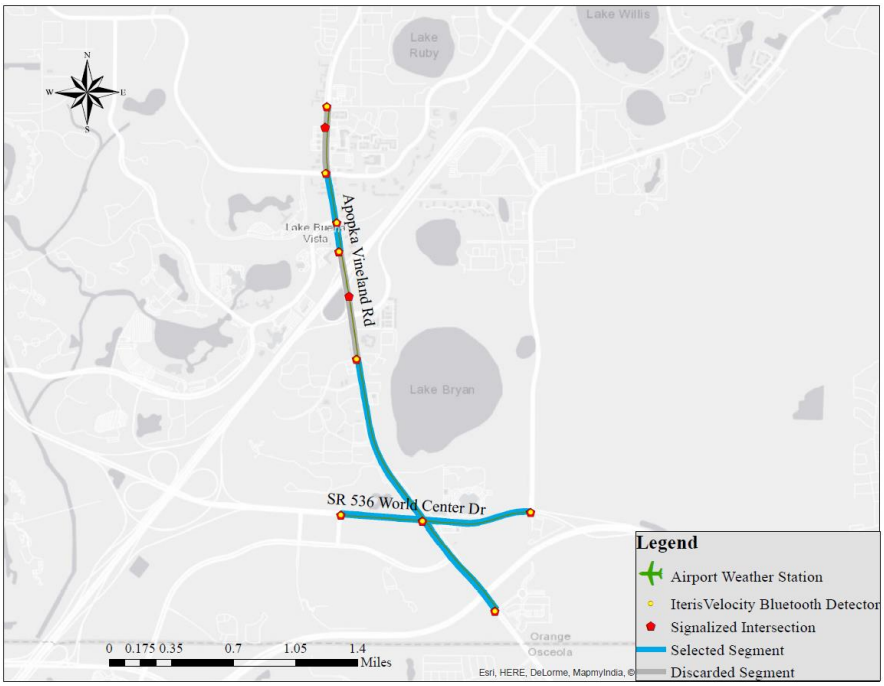
(4) The relationships between real-time crash occurrence and real-time traffic and signal characteristics on urban arterials are preliminarily revealed.

3.2 Data Preparation

The roads chosen are four urban arterials in Orlando, Florida, as shown in Figure 3-1. Initially, 72 road segments in both directions were considered in this study, the road segment here mentioned is defined as the segment between adjacent intersections. A total of four datasets were used: (1) crash data from March, 2017 to December, 2017 provided by Signal Four Analytics (S4A); (2) travel speed data collected by 23 IterisVelocity Bluetooth detectors installed at 23 intersections; (3) signal timing and 15-minute interval traffic volume provided by 23 adaptive signal controllers; (4) weather characteristics collected by the nearest airport weather station.



(a) Sand Lake Road and Orange Blossom Trail



(b) Apopka Vineland Road and SR 536 World Center Drive

Figure 3-1: Selected Four Urban Arterials

S4A provides detailed crash information, including crash time, coordinates, severity, type, weather condition, etc. In terms of the crash time information, there are three kinds of time information for each crash, i.e. time of crash occurrence, time reported, and time dispatched. Only the time of crash occurrence was utilized in this study, and the difference between this crash time and the actual crash time is supposed to be within 5 minutes since there exist several efficient and accurate technologies for the police officer to identify the accurate time of crash occurrence, e.g. closed-circuit television cameras and mobile phones.

First, all crashes occurred on the selected arterials from March 2017 to December 2017 were collected. After that, based on the attributes of "Type of Intersection" and "First Harmful Event Relation to Junction", all the intersection and intersection-related crashes were excluded. Meanwhile, all the crashes that occurred under the influence of alcohol and drugs were excluded. After these filtering processes, a total of 523 crashes remained and these crashes were assigned to the corresponding road segments.

Matched case-control design was employed in this study to explore the effects of traffic, signal, and weather-related variables while eliminating the effects of other confounding factors through the design of study. First, all the crash events were collected, and for each selected crash, several confounding factors, i.e., segment ID, time of day, and day of the week, were selected as matching factors. Therefore, a group of non-crash events could be identified by using these matching factors and then a specific number of non-crash events could be randomly

selected from this group of non-crash events for every crash (Figure 3-2). The number of non-crash events m corresponding to a crash event is preferred to be fixed in the entire analysis. As stated in Hosmer Jr et al. (2013), the value of m was commonly chosen from one to five. In addition, Abdel-Aty et al. (2004) found that there is no significant difference when m changing from one to five. Therefore, the control-to-case ratio of 4:1 was adopted in this study, which is consistent with previous research (Abdel-Aty et al., 2008; Ahmed and Abdel-Aty, 2013; Ahmed et al., 2012b; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2016; Zheng et al., 2010). Consequently, 4 non-crash events from the same road segment, time of day, and day of week were extracted for each crash event. Besides, these non-crash events were extracted only when there are no crashes occurring within 3 hours before or after the non-crash event on the same road segment.

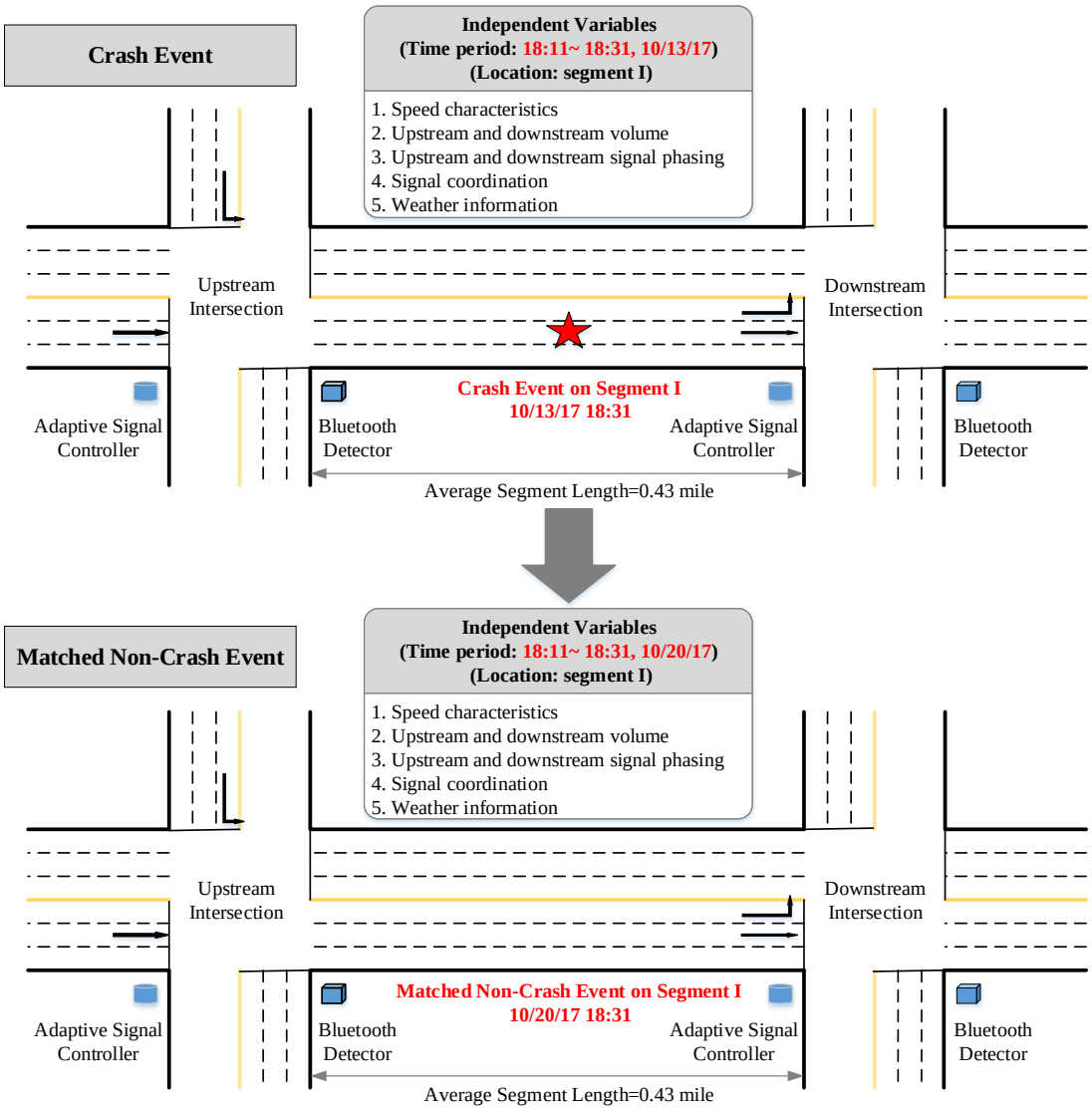


Figure 3-2: Illustration of Matched Case-Control Design

Bluetooth data provides the travel time and space-mean speed of the detected vehicle for each segment. Bluetooth detectors can only detect the vehicles equipped with Bluetooth device and the device is working at discoverable mode. The space-mean speed of each vehicle on a specific segment is calculated as the segment length divided by the travel time of each detected vehicle on the segment based on the detection data of two Bluetooth detectors located at the two contiguous intersections (Gong et al., 2019b). The procedure of Bluetooth data collection is

illustrated in Figure 3-3. In order to mitigate the impact of signal delay, the vehicle-level travel speed data were filtered by the algorithm which only keeps the data sample within 75% of the interquartile range of the preceding 15 samples on the same segment, this filtering algorithm could filter out those biased samples which might be highly influenced by the signal delay.

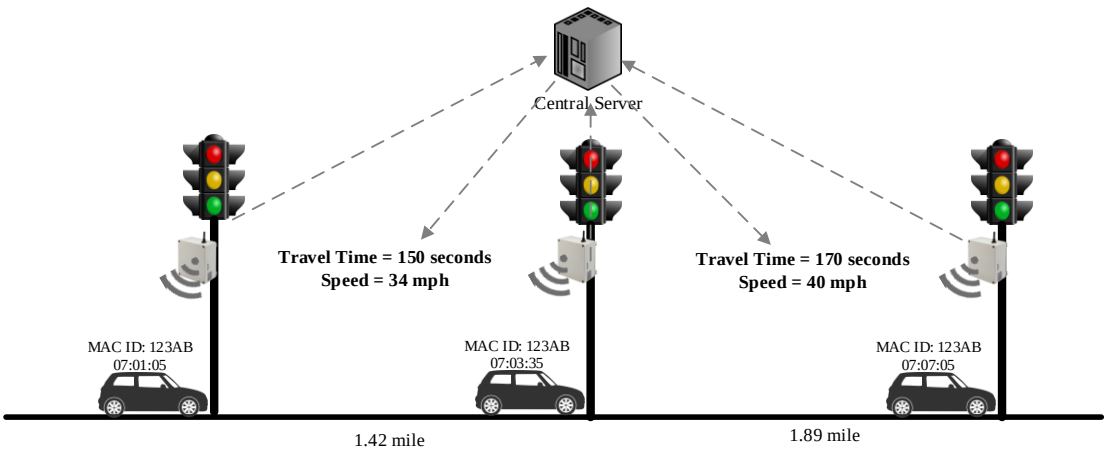


Figure 3-3: Illustration of Bluetooth Data Collection

If there is no Bluetooth detector on one of the contiguous intersections (Figure 3-4), the travel speed on the segment will be decreased after including the intersection delay, thus, all the segments with missing Bluetooth detector on either contiguous intersection were deleted. Consequently, only 32 road segments were selected for data collection (Figure 3-1).

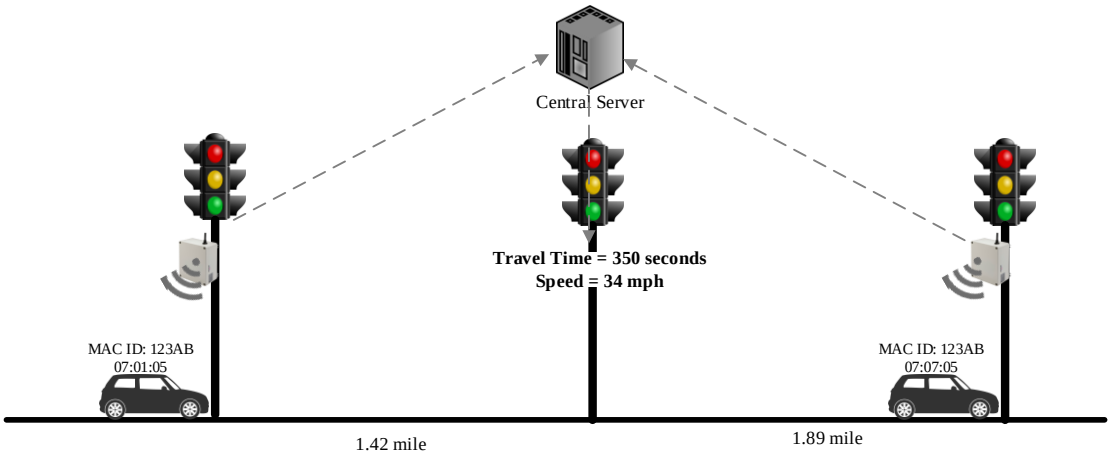


Figure 3-4: Illustration of Excluded Bluetooth Segment

It is worth noting that the Bluetooth overall sampling rate is 6.05%, which is higher than the threshold suggested by the previous studies (Chen and Chien, 2000; Long Cheu et al., 2002b), which stated that a floating car sample of just 3% of the vehicle population is sufficient for a 95% confidence level in travel time and speed estimates. The real-time travel speed data were extracted for a period of 20 minutes (divided into four 5-minute time slices) prior to crash occurrence. For example, if a crash occurred on segment-15 at 15:00, the corresponding travel speed data from 14:40 to 15:00 were extracted and named as time-slices 1, 2, 3, and 4. The distribution histogram of the 5-minute Bluetooth sample frequency is shown in Figure 3-5, if the number of vehicles that are detected within any time slice is lower than 2 (17.12%), then the corresponding crashes were excluded. Finally, a total of 273 crashes were used in the analysis.

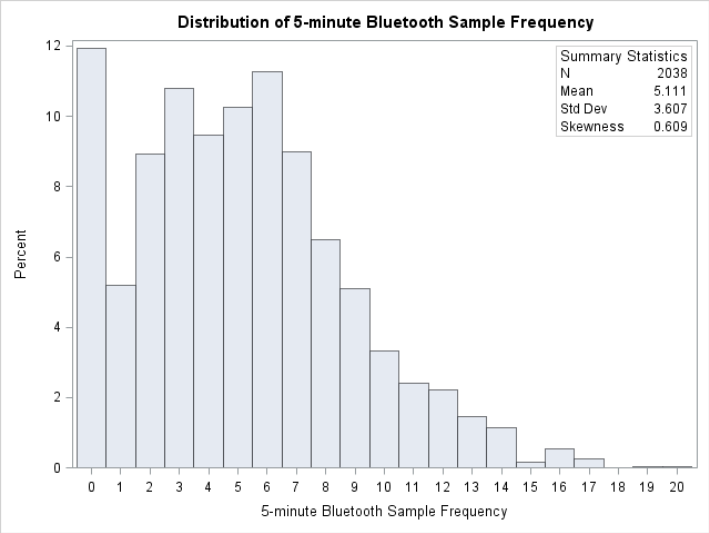


Figure 3-5: Distribution of 5-minutes Bluetooth Sample Frequency

The adaptive signal control system at a signalized intersection is operated based on the video detectors installed on the approaches, which can detect the real-time queue length, maximum waiting time, and traffic volume by movement. This system archives the real-time signal phasing, queue length, waiting time, and 15-minute aggregated traffic volume data. Since the right-turn vehicles are unprotected at the intersection, the traffic volume data only include the through and left-turn vehicles. As shown in Figure 3-2, the upstream volume of the segment consists of the through and left-turn traffic volume coming from the upstream intersection, while the downstream volume of the segment consists of the through and left-turn traffic volume approaching into the downstream intersection. Since the archived volume data are aggregated by 15 minutes, therefore, the traffic volume during 5-mintue interval was proportionally calculated based on the assumption that the traffic volume within 15-minute interval are evenly distributed.

The 5-minute through green ratio for the contiguous upstream and downstream intersections were collected for the period of 4 time slices prior to the reported crash time. Also, the 5-minutes signal coordination between the contiguous upstream and downstream intersections was collected. As shown in Figure 3-6, the signal coordination is the total maximum bandwidth ("windows" of green for traveling platoons) between the upstream and downstream signals during the periods of each time-slice prior to the reported crash time. The ideal offset, which is calculated by the segment length divided by the corresponding speed limit, was adopted to represent the offset between the upstream and downstream intersections.

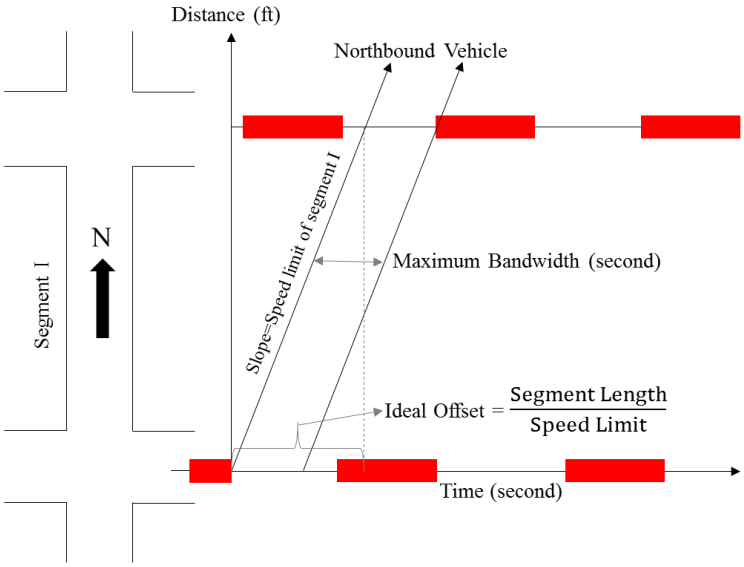


Figure 3-6: Illustration of maximum bandwidth and signal coordination

Two weather related variables (rainy weather indicator and visibility) were collected from the nearest airport weather station, which is located at the Orlando international airport (as shown in Figure 3-1). Since the weather data is not recorded continuously, once the weather condition

changes and reaches a preset threshold, a new record will be added to the archived data. Therefore, for each specific crash, based on the reported crash time, the closest weather record prior to the crash time has been extracted and used as the crash time weather condition, which is identical for four time slices.

In order to validate the weather data collected by the airport weather station with the weather condition reported in the crash report. The weather type information of each crash event collected from two data sources was selected to conduct a cross table analysis. The weather type information reported in the crash report including clear (76.44%), cloudy (13.29%), and rain (10.27%), which were converted into a binary variable (rainy and normal) to compare with the rainy weather indicator collected by the airport weather station. The results indicated that the accuracy ((True positive + True negative)/Total sample size) of weather station is 92%.

The final dataset includes 1365 observations (273 crash events and 1092 non-crash events), which were then divided into training (80%: 218 crash events) and validation (20%: 55 crash events) datasets. The summary statistics of the final dataset for all the traffic, signal, and weather-related variables are as shown in Table 3-1.

Table 3-1: Summary of Variables Descriptive Statistics (Crash and Non-crash Events)

Variables	Time Slice	Description	Mean	Std dev.	Min	Max
Crash_count	-	Number of crashes for each segment	9.10	7.50	1.00	29.00
Avg_Speed	1	Average speed within 5-minute interval (mph)	25.91	10.18	4.88	55.00
	2		26.07	10.01	4.00	56.00
	3		26.40	10.05	4.00	58.00
	4		26.21	9.84	4.33	59.33
Std_Speed	1	Speed standard deviation within 5-minute interval (mph)	9.86	5.20	0.00	30.41
	2		10.06	5.02	0.00	31.01
	3		10.00	5.12	0.00	36.77
	4		10.11	5.22	0.00	28.54
Up_Vol	1	Number of vehicles coming from the upstream intersection within 5-minute interval	108.85	53.55	0.00	346.67
	2		109.00	53.81	0.00	346.67
	3		108.25	53.04	0.00	316.67
	4		107.87	54.54	0.00	491.33
Down_Vol	1	Number of vehicles approaching into the downstream intersection within 5-minute interval	123.28	56.81	0.00	869.33
	2		123.38	56.05	0.00	869.33
	3		122.85	56.54	0.00	869.33
	4		122.96	55.76	0.00	557.33
Up_Vol_LT	1	Number of left turn vehicles coming from the upstream intersecting road segment within 5-minute interval (Figure 2)	10.19	18.15	0.00	146.67
	2		10.14	18.00	0.00	134.93
	3		10.18	18.09	0.00	142.67
	4		10.11	17.80	0.00	142.67
Down_Vol_LT	1	Number of left turn vehicles approaching into the downstream intersection within 5-minute interval (Figure 2)	16.12	14.76	0.00	118.33
	2		16.09	15.25	0.00	149.67
	3		16.14	15.59	0.00	149.67
	4		16.14	15.79	0.00	149.67
Up_Green_Ratio	1	The percentage of green time for through vehicle in the upstream intersection within 5-minute interval (%)	47.87	18.32	4.00	100.00
	2		46.96	17.57	12.67	94.67
	3		47.11	18.14	8.33	100.00
	4		47.78	17.96	10.67	100.00
Down_Green_ratio	1	The percentage of green time for through vehicle in the downstream intersection within 5-minute interval (%)	46.97	18.66	9.00	100.00
	2		47.11	18.76	7.67	93.67
	3		46.17	17.99	6.67	100.00
	4		46.76	18.85	7.67	92.33
Signal_Coordination	1	Total bandwidth divided by the upstream green time within 5-minute interval	0.66	0.29	0.00	1.00
	2		0.65	0.29	0.00	1.00
	3		0.65	0.29	0.00	1.00
	4		0.65	0.29	0.00	1.00
Rainy	-	Binary variable for rainy weather indicator (0 for normal and 1 for rainy)	0.05	0.21	0.00	1.00
Visibility	-	Visibility (mile)	9.79	1.09	1.00	10.00

3.3 Methodology

As crash risk analysis is a typical binary classification problem (crash and non-crash), logistic regression model would be the most basic and preferable method. However, since the matched case control design was employed in this study to select the non-crash events rather than the random sample method, which means that the selected non-crash events and the corresponding crash event are within the same stratum. Therefore, conditional logistic regression, which is also known as matched-case control regression, should be more appropriate for this study, which is in line with previous research (Abdel-Aty and Pande, 2005; Abdel-Aty et al., 2004; Ahmed and Abdel-Aty, 2012; Xu et al., 2012; Zheng et al., 2010). In this study, four Bayesian conditional logistic models were developed for the four time slices separately.

Furthermore, many previous research found that random parameters model performs much better than fixed parameters model (Shi and Abdel-Aty, 2015; Xu et al., 2014; Yu and Abdel-Aty, 2014; Yu et al., 2017). Therefore, Bayesian random parameters logistic model and Bayesian random parameters conditional logistic model were also employed based on the best time slice dataset to compare with the Bayesian conditional logistic model. Bayesian approach, which treats the parameters as random variable and incorporates prior knowledge to estimate the posterior distribution of parameters, was adopted in this study. It was claimed that the Bayesian approach provided better fit and reduced uncertainty for parameter estimations than the frequentist approach (Ahmed et al., 2012b).

3.3.1 Bayesian Conditional Logistic Model

Suppose that there are *N* strata with 1 crash (y_{ij}=1) and *m* non-crashes (y_{ij}=0) in stratum *i*, *i*=1, 2, …, *N* and *j*=0,1,2, …, *m*. Let p_{ij} be the probability that the *j*th observation in the *i*th stratum is a crash. This crash probability could be expressed as:

$$y_{ij} \sim Bernoulli(p_{ij}) \quad (3\text{-}1)$$

$$logit(p_{ij}) = \alpha_i + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{Kij} \quad (3\text{-}2)$$

Where α_i denotes the effects of matching variables on crash likelihood for ith stratum; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$ is the vector of regression coefficients for K independent variables, and all the $\boldsymbol{\beta}$ coefficients are set up with non-informative priors as following normal distributions (0, 1E-6); $\boldsymbol{X_{ij}} = (X_{1ij}, X_{2ij}, \ldots, X_{Kij})$ is the vector of K independent variables.

In order to take the stratification in the analysis of the observed data, the stratum-specific intercept α_i is considered to be nuisance parameters. Suppose the observation y_{i0} is a crash, and $y_{ij}, j = 1, 2, \ldots, m$ are non-crashes, then the conditional likelihood for the *i*th stratum would be expressed as (Hosmer Jr et al., 2013):

$$l_i(\boldsymbol{\beta}) = \frac{\exp(\sum_{k=1}^{K} \beta_k X_{ki0})}{\sum_{j=0}^{m} \exp(\sum_{k=1}^{K} \beta_k X_{kij})} \quad (3\text{-}3)$$

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

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} l_i(\boldsymbol{\beta}) \qquad (3\text{-}4)$$

Since the full conditional likelihood is independent of stratum-specific intercept α_i, thus Eq. (3-2) cannot be used to estimate the crash probabilities. However, the estimated $\boldsymbol{\beta}$ coefficients are the log-odd ratios of corresponding variables and can be used to approximate the relative risk of an event. Furthermore, the log-odds ratios can also be used to develop a prediction model under this matched case-control analysis. Suppose two observation vectors $\boldsymbol{X_{i1}} = (X_{1i1}, X_{2i1}, \dots, X_{Ki1})$ and $\boldsymbol{X_{i2}} = (X_{1i2}, X_{2i2}, \dots, X_{Ki2})$ from the ith stratum, the odds ratio of crash occurrence caused by observation vector $\boldsymbol{X_{i1}}$ relative to observation vector $\boldsymbol{X_{i2}}$ could be calculated as:

$$\frac{p_{i1}/(1-p_{i1})}{p_{i2}/(1-p_{i2})} = \exp[\sum_{k=1}^{K} \beta_k (X_{ki1} - X_{ki2})] \qquad (3\text{-}5)$$

The right-hand side of Eq. (3-5) is independent of α_i and can be calculated using the estimated $\boldsymbol{\beta}$ coefficients. Thus, the above relative odds ratio could be utilized for predicting crash occurrences by replacing $\boldsymbol{X_{i2}}$ with the vector of the independent variables in the *i*th stratum of non-crash events. One may use simple average of each variable for all non-crash observations within the stratum. Let $\bar{\boldsymbol{X}}_{\boldsymbol{i}} = (\bar{X}_{1i}, \bar{X}_{2i}, \dots, \bar{X}_{Ki})$ denote the vector of mean values of non-crash events of the k variables within the ith stratum. Then the odds ratio of a crash relative to the non-crash events in the *i*th stratum could be approximated by:

$$\frac{p_{i1}/(1-p_{i1})}{p_{\bar{i}}/(1-p_{\bar{i}})} = \exp[\sum_{k=1}^{K} \beta_k (X_{ki1} - \bar{X}_{ki})] \quad (3\text{-}6)$$

3.3.2 Bayesian Random Parameters Logistic Model

Suppose the crash occurrence has the outcomes y_i=1 (crash event) and y_i=0 (non-crash event) with respective probability p_i and 1-p_i, i=1, 2,…, *N(m+1). N* and *m* represent the number of strata and the number of control events within each stratum, separately. *N(m+1)* indicates the total number of observations. The random parameters logistic regression can be expressed as follows:

$$y_i \sim Bernoulli(p_i) \quad (3\text{-}7)$$

$$logit(p_i) = \beta_{0i} + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \cdots + \beta_{Ki}X_{Ki} \quad (3\text{-}8)$$

$$\beta_{ki} = \beta_k + \varphi_{ki}, \qquad k = 0,1,2,\dots,K \quad (3\text{-}9)$$

$$\varphi_{ki} \sim N(0, \sigma_k^2) \quad (3\text{-}10)$$

Where β_{0i} is the random intercept for the ith observation; $\boldsymbol{\beta_i} = (\beta_{1i}, \beta_{2i}, \dots, \beta_{Ki})$ is the vector of K random coefficients for the ith observation; $\boldsymbol{X_{ij}} = (X_{1ij}, X_{2ij}, \dots, X_{Kij})$ is the vector of K independent variables for the ith observation; φ_{ki} is a randomly distributed term to account for the heterogeneity across observations; all the β_k coefficients are set up with non-informative priors as following normal distributions (0, 1E-6), and all the σ_k^2 are specified to be inverse-gamma priors as $\sigma_b^2 \sim \text{Inverse} - \text{gamma}(0.001, 0.001)$.

3.3.3 Bayesian Random Parameters Conditional Logistic Model

Suppose the crash occurrence has the outcomes y_{ij}=1 (crash event) and y_{ij}=0 (non-crash event) with respective probability p_{ij} and 1-p_{ij}. The definitions of *i* and *j* are the same with Eq. (3-1). The random parameters conditional logistic regression can be expressed as follows:

$$y_{ij} \sim Bernoulli(p_{ij}) \quad (3\text{-}11)$$

$$logit(p_{ij}) = \alpha_i + \beta_{1i}X_{1ij} + \beta_{2i}X_{2ij} + \cdots + \beta_{Ki}X_{Kij} \quad (3\text{-}12)$$

$$\beta_{ki} = \beta_k + \varphi_{ki}, \qquad k = 0,1,2,\dots,K \quad (3\text{-}13)$$

$$\varphi_{ki} \sim N(0, \sigma_k^2) \quad (3\text{-}14)$$

Where α_i is the random intercept term for the ith stratum; $\boldsymbol{\beta_i} = (\beta_{1i}, \beta_{2i}, \dots, \beta_{Ki})$ is the vector of K random coefficients for the *i*th stratum; $\boldsymbol{X_{ij}} = (X_{1ij}, X_{2ij}, \dots, X_{Kij})$ is the vector of K independent variables for the jth observation in the ith stratum; φ_{ki} is a randomly distributed term to account for the heterogeneity across strata; The main difference between random parameters logistic model and random parameters conditional logistic model is that the estimation of random parameters logistic model is based on classical likelihood function while random parameters conditional logistic model is based on the stratified conditional likelihood function (as shown in Eq. (3-4)). All the β_k coefficients are also set up with non-informative

priors as following normal distributions (0, 1E-6), and all the σ_k^2 are specified to be inverse-gamma priors as $\sigma_b^2 \sim \text{Inverse} - \text{gamma}(0.001, 0.001)$.

Bayesian Inference and Model Comparisons

Bayesian inference was employed in this study. For each model, three chains of 20,000 iterations were set up in WinBUGS (Lunn et al., 2000), the first 5,000 iterations were excluded as burn-in, the latter 15,000 stored iterations were set to estimate the posterior distribution. Convergence was evaluated using the built-in Brooks-Gelman-Rubin (BGR) diagnostic statistic (Brooks and Gelman, 1998).

The Deviance Information Criterion (DIC) can be used to compare complex models by offering a Bayesian measure of model fitting and complexity (Spiegelhalter et al., 2002). DIC is defined as:

$$DIC = \overline{D(\theta)} + p_D \qquad (3\text{-}15)$$

Where $D(\theta)$ is the Bayesian deviance of the estimated parameter, and $\overline{D(\theta)}$ is the posterior mean of $D(\theta)$. $\overline{D(\theta)}$ can be viewed as a measure of model fit, while p_D denotes the effective number of parameters and indicates the complexity of the models. Models with smaller DIC are preferred. Very roughly, difference of more than 10 might definitely rule out the model with the higher DIC (Spiegelhalter et al., 2003).

In terms of model goodness-of-fit, the AUC value which is the area under Receiver Operating Characteristic (ROC) curve was also adopted. The ROC curve illustrates the relationship between the true positive rate (sensitivity) and the false alarm rate (1–specificity) of model classification results based on a given threshold from 0 to 1. It is worth noting that the classification results of Bayesian random parameters logistic model are based on the predicted crash probabilities, which lie in the range of 0 to 1, while the classification result of Bayesian conditional logistic model and Bayesian random parameters conditional logistic model are based on the predicted odds ratio, which may be larger than 1. In order to be consistent with the other two models, all the odds ratios predicted by Bayesian conditional logistic model were divided by the maximum odds ratio to create adjusted odds ratios. Later, the adjusted odds ratios were used to create the classification result based on different threshold from 0 to 1. In this study, AUC values were calculated using R package pROC (Robin et al., 2011).

3.4 Modeling Results

This section discusses the modeling results of the Bayesian conditional logistic models based on four time slices datasets, followed by the model comparisons between Bayesian conditional logistic model, Bayesian random parameters logistic model, and Bayesian random parameters conditional logistic model based on the same dataset.

Four models based on 4 time-slice datasets are presented in Table 3-2. The model comparison results based on training and validation AUC values indicate that the slice 2 model (5-10 minute interval) performs the best, followed by the slice 1 (0-5 minute interval) model. However, based on slice 1 model, there would be no spare time to implement any proactive traffic management strategy to prevent the possible crash occurrence. Moreover, as Golob and Recker (Golob et al., 2004) mentioned that there may exist 2.5 min difference between the exact crash time and reported crash time, thus the slice 1 model was treated as a reference. On the other hand, slice 2 model performs the best in terms of the number of significant variables. Finally, the slice 2 model was selected to conduct further interpretation and model comparison.

Table 3-2: Model Results of Bayesian Conditional Logistic Regression Models based on Different Time Slices

Parameter	Slice 1		Slice 2		Slice 3		Slice 4	
	Mean (95% BCI)	Odds Ratio	Mean (95% BCI)	Odds Ratio	Mean (95% BCI)	Odds Ratio	Mean (95% BCI)	Odds Ratio
Avg_speed	**-0.049 (-0.071, -0.029)**	**0.952 (0.931, 0.971)**	**-0.025 (-0.048, -0.004)**	**0.975 (0.953, 0.996)**	-	-	-	-
Up_Vol_LT	**0.024 (0.007, 0.044)**	**1.024 (1.007, 1.045)**	**0.024 (0.005, 0.044)**	**1.024 (1.005, 1.045)**	**0.024 (0.006, 0.045)**	**1.024 (1.006, 1.046)**	**0.036 (0.014, 0.06)**	**1.037 (1.014, 1.062)**
Down_GreenRatio	-	-	**-0.042 (-0.075, -0.011)**	**0.959 (0.928, 0.989)**	-	-	-	-
Rainy	**0.551 (0.02374, 1.065)***	**1.735 (1.024, 2.901)**	**0.667 (0.055, 1.274)**	**1.948 (1.057, 3.575)**	**0.682 (0.037, 1.322)**	**1.978 (1.038, 3.751)**	**0.72 (0.078, 1.341)**	**2.054 (1.081, 3.823)**
Training AUC	0.6150		0.6210		0.5451		0.5507	
Validation AUC	0.6081		0.6169		0.5300		0.5476	

*Note: Mean (95% BCI) values marked in bold are significant at the 0.05 level; Mean (95% BCI) values marked in bold and noted by * are significant at the 0.1 level.*

Based on the estimation results in the slice 2 model, four variables were found to be significantly associated with the crash occurrence on urban arterials: (1) the negative coefficient (-0.025) of average speed indicates that higher average speed tends to decrease the crash risk, which is consistent with other studies (Abdel-Aty et al., 2012; Ahmed et al., 2012a, b; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2016). This could be explained as the traffic condition with higher average speed, which represents more smooth traffic flow, could have better safety performance. Similarly, congestion index was found to have positive effect on crash likelihood (Hossain and Muromachi, 2012; Shi and Abdel-Aty, 2015), which means that the congested traffic condition is expected to have higher crash risk. The odds ratio of 0.975 means that when other variables held constant, one-unit increase in the average speed would decrease the odds of crash occurrence by 2.5%; (2) the upstream left-turn volume from the intersecting road segment was found to be positively correlated with crash likelihood, which might be explained in that more vehicles from the intersecting road segment left turning into the subject segment may result in more lane change behavior, which may lead to more conflicts with through vehicles. The odds ratio of 1.024 indicates that one-unit increase in upstream left-turn volume would lead to an increase of 2.4% in the odds of crash occurrence; (3) downstream green ratio was found to have negative effect on crash risk, and the odds ratio of 0.959 indicates that one percentage increase in downstream green ratio would decrease the odds of crash occurrence by 4.1%; (4) rainy indicator has a positive effect, the odds ratio of 1.948 means that the odds of crash occurrence under rainy

condition is 94.8% higher than normal conditions, which is in line with previous studies (Ahmed et al., 2012a).

Furthermore, both Bayesian random parameters logistic model and Bayesian random parameters conditional logistic model were developed based on time slice 2 dataset. In order to improve the model performance of the Bayesian random parameters conditional logistic model, 15 ($\sum_{i=0}^{3} \frac{4!}{i!(n-i)!}$) combinations of fixed and random variables were developed to compare the model results, Table 3-3 shows the model performance of the 15 random parameter combinations.

Table 3-3: Model Performance of Different Random Parameter Combinations

Model Type	Fixed Variables	Training AUC	Validation AUC
4 random variables	-	0.6217	0.6196
3 random and 1 fixed variables	Rainy	0.6211	0.6155
	Down_GreenRatio	0.6202	0.6126
	Up_Vol_LT	0.6216	0.6232
	Avg_speed	0.6208	0.6134
2 random and 2 fixed variables	Avg_speed & Rainy	0.6206	0.614
	Avg_speed & Up_Vol_LT	0.6208	0.6246
	Avg_speed & Down_GreenRatio	0.6209	0.6163
	Up_Vol_LT & Down_GreenRatio	0.622	0.6232
	Up_Vol_LT & Rainy	0.6215	0.6163
	Down_GreenRatio & Rainy	0.6208	0.6157
1 random and 3 fixed variables	Avg_speed & Up_Vol_LT & Down_GreenRatio	0.6213	0.6164
	Avg_speed & Up_Vol_LT & Rainy	0.6216	0.6164
	Avg_speed & Down_GreenRatio & Rainy	0.6202	0.6119
	Up_Vol_LT & Down_GreenRatio & Rainy	0.6207	0.6158

Since all the modeling results of these 15 combinations will be too much to present, only the best model (i.e. fix “Up_Vol_LT” and “Down_GreenRatio”, while randomize the other two variables) was presented in Table 3-4. Both the AUC and DIC values indicate that the Bayesian random parameters conditional logistic model performs better than the Bayesian conditional logistic model, which verified that introducing random parameters could improve model performance. However, in the Bayesian random parameters logistic model, the upstream left-turn volume and downstream green ratio are insignificant, and this model has the lowest AUC value and the highest DIC value among the three models, these indicate that without considering the stratified data structure of the matched case-control dataset may significantly deteriorate the model performance.

Table 3-4: Model Comparison Results based on Time Slice 2

Parameter	Bayesian conditional logistic regression		Bayesian random parameters logistic model		Bayesian random parameters conditional logistic model	
	Mean (95% BCI)	Hazard Ratio	Mean (95% BCI)	Hazard Ratio	Mean (95% BCI)	Hazard Ratio
Intercept	-	-	**-1.514 (-2.35, -0.607)**	-	-	-
Standard deviation	-	-	*0.074 (0.021, 0.19)*	-	-	-
Avg_speed	**-0.025 (-0.048, -0.004)**	0.975 (0.953, 0.996)	**-0.023 (-0.041, -0.006)**	0.977 (0.96, 0.994)	**-0.027 (-0.051, -0.006)**	0.973 (0.95, 0.994)
Standard deviation	-	-	*0.012 (0.009, 0.017)*	-	*0.044 (0.018, 0.091)*	-
Up_Vol_LT	**0.024 (0.005, 0.044)**	1.024 (1.005, 1.045)	**0.009 (-0.002, 0.021)**	1.009 (0.998, 1.021)	**0.025 (0.004, 0.047)**	1.025 (1.004, 1.048)
Standard deviation	-	-	*0.017 (0.012, 0.024)*	-	-	-
Down_GreenRatio	**-0.042 (-0.075, -0.011)**	0.959 (0.928, 0.989)	**-0.007 (-0.017, 0.003)**	0.993 (0.983, 1.003)	**-0.045 (-0.076, -0.013)**	0.956 (0.927, 0.987)
Standard deviation	-	-	*0.009 (0.007, 0.011)*	-	-	-
Rainy	**0.667 (0.055, 1.274)**	1.948 (1.057, 3.575)	**0.797 (0.102, 1.436)**	2.219 (1.107, 4.204)	**0.591 (0.082, 1.224)***	1.806 (1.085, 3.401)
Standard deviation	-	-	*0.070 (0.021, 0.17)*	-	*0.283 (0, 0.543)*	-
DIC	682.290		1179.610		676.674	
Training AUC	0.6210		0.5748		0.6220	
Validation AUC	0.6169		0.5714		0.6232	

*Note: Mean (95% BCI) values marked in bold are significant at the 0.05 level; Mean (95% BCI) values marked in bold and noted by * are significant at the 0.1 level; The value in italic are the standard deviation of the corresponding parameter distribution.*

3.5 Conclusion and Discussion

This study investigated the crash risk on urban arterials based on real-time data from multiple sources, including travel speed provided by Bluetooth detectors, traffic volume and signal phasing extracted from adaptive signal controllers, and weather data collected by the airport weather station. Matched case-control design with a control-to-case ratio of 4:1 was applied to collect data for crash and non-crash events. Four Bayesian conditional logistic models were developed separately for four 5-minute interval datasets (20-minute window prior to the reported crash time). In terms of AUC values, the model estimation results indicated that slice 2 (5-10 minute) model performs the best, followed by the slice 1 (0-5 minute) model. Considering that the implementation of proactive traffic management strategy may need some time in advance to possible crash occurrence, and there may exists error between the reported and actual crash times (Golob et al., 2004), slice 1 model was disregarded and slice 2 model was selected to conduct further analysis.

The results of the slice 2 model indicate that the average speed, upstream left-turn volume, downstream green ratio, and rainy indicator are significantly associated with the crash risk on urban arterials. In general, these finding are consistent with previous studies, in which the average speed was found to have significant negative impact on crash occurrence (Abdel-Aty et al., 2012; Ahmed et al., 2012a, b; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2016), while adverse weather (Ahmed et al., 2012a; Xu et al., 2013a) were found to be positively correlated with crash likelihood. In terms of the effect of traffic volume, only the upstream left-turn volume was found to have significant effect on crash likelihood, which indicates that more vehicles from the intersecting road segment left turning into the subject segment may

increase the crash risk on the segment. This is quite different from the findings on freeways, which showed that the total upstream volume has significant positive impact on crash occurrence (Shi and Abdel-Aty, 2015; Yu et al., 2017; Yu et al., 2016).

It is worth noting that the downstream green ratio was found to be negatively associated with crash occurrence, this could be explained as the higher downstream green ratio could efficiently reduce the percentage of stop-and-go traffic, which may increase the safety performance. Surprisingly, the speed standard deviation is insignificant, this could be explained in that the average number of vehicles detected by the Bluetooth detector within 5-minute interval is about 6, which might be too small to capture the variation in speed.

Compared with the previous research on the real-time safety analysis of urban arterials (Theofilatos, 2017), they found that the 1 hour variation in both occupancy and volume were significantly associated with crash likelihood, which is quite different from our study. This might be explained in that the 1 hour aggregated traffic parameters can hardly represent the actual short-term traffic status such as speed and volume prior to crash occurrence, while it can capture the variation in traffic flow. This comparison implies that the traffic parameters should be aggregated based on more appropriate time interval, which can not only represent the short-term traffic status but also capture the variation in traffic flow characteristics.

Furthermore, the Bayesian random parameters logistic and Bayesian random parameters conditional logistic models were developed and compared with the Bayesian conditional logistic model based on the time slice 2 dataset. The results indicate that the Bayesian random parameters logistic model which ignored the stratified structure of the matched-case-control dataset performs the worst, which verifies that the stratified structure of the matched-case-control dataset should be

considered in the modeling process. Moreover, the Bayesian random parameters conditional logistic model performs better than the Bayesian conditional logistic model, which demonstrates the advantage of random parameters model.

From the application point of view, the outcome of this study could be implemented from several aspects. The most straightforward application is to apply this algorithm to develop an arterial real-time crash risk prediction system. The real-time prediction results could be fed into the implementation of proactive traffic management strategies (e.g., variable speed limit), which can efficiently mitigate the crash risk in advance of the potential crash occurrence. Also, the real-time prediction results could be provided to drivers to assist with the route choice decisions. Furthermore, the real-time crash prediction results could be delivered to the drivers through connected-vehicle technology to provide crash risk warning information. In addition, the arterial real-time crash risk prediction system could be integrated with the real-time crash prediction on freeways. Therefore, an integrated arterial/freeway active traffic management strategy could be employed to proactively mitigate the safety of the road network.

However, the validation AUC value of 0.6232 implies that the model is still not ready to be applied to the real-time crash risk prediction and active traffic management system. It is worth noting that this study could be considered as a pioneering but early stage investigation of real-time safety analysis on urban arterials and that its major contribution is to demonstrate the concept of applying Bluetooth and adaptive signal control data to predict real-time crash risk on urban arterials. Even though, the current estimation results could still provide some insights for traffic engineers to understand the relationship between crash risk and real-time traffic characteristics and weather conditions on arterials.

As this is the first attempt to investigate the real-time crash risk on urban arterials based on 5-minute aggregated data, there are still plenty of room for further improvements: (1) in order to achieve more accurate vehicle-level travel time and speed, the vehicle delay at intersections should be excluded from the travel time. In this context, high-resolution vehicle trajectory data would be preferable rather than Bluetooth data. (2) The current study focused on the safety effect of the traffic and signal status during different 5-minute intervals prior to the crash occurrence. Therefore, the exact signal status at the time of crash occurrence has not been considered. More disaggregate analyses, e.g., 1-min level or even signal cycle level analysis, should be conducted when higher resolution data are available. (3) As the Bluetooth data only provide the speed of the segment, it cannot distinguish the lane specific travel speed. In the future, lateral speed difference should be considered when more microscopic data are available. (4) The signal timing characteristics were incorporated as several independent variables, which is relatively simple and superficial. More integrated analysis should be conducted to reveal the intrinsic relationship between signal timing and real-time crash risk on urban arterials. (5) This study only focused on the total crashes, while different crash types and crash severity could be considered in future research.

CHAPTER 4: APPROACH-LEVEL REAL-TIME CRASH RISK ANALYSIS FOR SIGNALIZED INTERSECTIONS[2]

4.1 Introduction

Intersections are among the most dangerous roadway facilities due to the complex traffic conflicting movements and frequent stop-and-go traffic. Take Florida as an example, nearly 26% of crashes happen at or influenced by intersections (including signalized and non-signalized) in 2014. Moreover, signalized intersections are generally large intersections with higher traffic volume, therefore, the safety status of signalized intersection would be even more complicated. Safety analysis for signalized intersection has been a critical research topic during past decades. Substantial efforts have been made by previous researchers to reveal the relationship between crash frequency of signalized intersections and all the possible contributing factors such as roadway geometric, signal control, and traffic characteristics, etc. (Abdel-Aty and Wang, 2006; Cai et al., 2018a; Cai et al., 2018b; Chin and Quddus, 2003; Guo et al., 2010; Lee et al., 2017; Liu et al., 2018; Wang et al., 2009; Wang et al., 2006).

More specifically, nearly all the traffic volume related variables were found to have significant positive effects on the crash frequency at signalized intersections, including total entering ADT (Abdel-Aty and Wang, 2006; Chin and Quddus, 2003; Guo et al., 2010; Poch and Mannering, 1996), right-turn ADT (Chin and Quddus, 2003; Poch and Mannering, 1996), left-turn ADT (Poch and Mannering, 1996), total ADT on major road (Dong et al., 2014; Wang et al., 2009), total ADT on minor road (Dong et al., 2014; Wang et al., 2009), left-turn ADT on major road (Guo et al., 2010), through ADT on minor road (Guo et al., 2010). However, Guo et al. (2010) found that the through ADT on major road and the left-turn ADT on minor road are significantly negatively

[2] This chapter has been published in Accident Analysis & Prevention (https://doi.org/10.1016/j.aap.2018.07.031)

associated with the crash frequency at signalized intersections. Moreover, Wang et al. (2009) investigated the relationship between LOS and safety at signalized intersections. They found that LOS D is a desirable level which is associated with less total crashes, rear-end and sideswipe crashes, as well as right-angle and left-turn crashes. Xie et al. (2013) investigated the safety effect of corridor-level travel speed, they found that the high-speed corridor may results in more crashes at the signalized intersections. Similarly, the speed limit of the corridor was found to be significantly positively correlated with the crash frequency of the signalized intersections (Abdel-Aty and Wang, 2006; Dong et al., 2014; Guo et al., 2010; Poch and Mannering, 1996; Wang et al., 2009).

With respect to the geometric design, number of lanes, median width, and intersection sight distance et al. were found to have significant effects on the crash frequency of signalized intersections. More specifically, the number of lanes was found to be positively correlated with the crash frequency of signalized intersections (Abdel-Aty and Wang, 2006; Dong et al., 2014; Guo et al., 2010; Poch and Mannering, 1996). Median width and intersection sight distance was also found to have positive effect on the crash frequency(Chin and Quddus, 2003). Moreover, Abdel-Aty and Wang (2006) found that the existence of exclusive right-turn lanes could significantly decrease the crash frequency.

In terms of signal control characteristics, the adaptive signal control was found to have significant lower crash frequency than the pre-timed signal control (Chin and Quddus, 2003). The number of phases was found to be positively associated with the crash frequency of signalized intersections (Chin and Quddus, 2003; Poch and Mannering, 1996; Xie et al., 2013). The left-turn protection could significantly improve the safety performance of the signalized interaction (Abdel-Aty and Wang, 2006; Chin and Quddus, 2003; Poch and Mannering, 1996). However, Abdel-Aty and

Wang (2006) found that the left-turn protection on minor roadway tends to increase the crash frequency of signalized intersection. Surprisingly, Guo et al. (2010) found that the coordinated intersections are more unsafe than the isolated ones. They explained it as the travel speed is higher for coordinated intersections because of the green wave, which may result in more crashes.

However, these studies were conducted based on static and highly aggregated data (e.g., Annual Average Daily Traffic (AADT), annual crash frequency). These aggregated data limit the reliability of the findings simply because they are averages and cannot reflect the real conditions at the time of crash occurrence. With the rapid development of traffic surveillance system and detection technologies, real-time traffic data are not only available on freeways and expressways but also on urban arterials (including road segments and intersections). During the past decade, an increasing number of studies have investigated the crash likelihood on freeways by using real-time traffic and weather data (Abdel-Aty et al., 2004; Abdel-Aty et al., 2012; Ahmed et al., 2012a; Basso et al., 2018; Lee et al., 2003; Oh et al., 2001; Theofilatos et al., 2018a; Xu et al., 2013a; Xu et al., 2013b; Yu and Abdel-Aty, 2014; Yu et al., 2014; Zheng et al., 2010). It is worth noting that Theofilatos et al. (2018a) investigated crash occurrence by utilizing real-time traffic data while considering that the number of crashes is very few, and they could be considered as rare events. In this context, they compared the model results of different crash to non-crash ratio (1:10 and full sample of non-crash events) by using two different statistical models (bias correction and firth model), respectively. It was found that the two methods have different advantages and disadvantages, and the choice of the most appropriate method depends on several criteria. Also, Basso et al. (2018) developed real-time crash prediction model for urban expressway based on the original unbalanced data, rather than artificially balanced data by using Synthetic Minority Over-

sampling Technique (SMOTE). They claimed that their model performance is among the best in the literature.

However, little research has been conducted on the real-time safety of urban arterials (Theofilatos, 2017; Theofilatos et al., 2017; Yuan et al., 2018a), especially signalized intersections (Mussone et al., 2017). Mussone et al. (2017) examined the factors which may affect the crash severity level at intersection based on real-time traffic flow and environmental characteristics, and they found that the real-time traffic flow characteristics have a relevant role in predicting crash severity. However, they didn't consider the crash likelihood at intersections, which means that the effects of real-time traffic flow and environmental characteristics on the crash likelihood at intersections are still unclear.

Moreover, the conflicting traffic movements at signalized intersection are temporally separated by traffic signals. Therefore, signal timing plays a very important role in the intersection safety, especially when the adaptive signal control technology was widely adopted on major urban arterials. Adaptive signal control technology optimize signal timing plans in real-time, it was found to have significant effects in reducing stops and delays (Khattak et al., 2018a) and improving traffic safety (Chin and Quddus, 2003; Khattak et al., 2018b). However, the safety effect of real-time signal status has never been considered, while improper signal timing may result in dangerous situation. Therefore, the relationship between real-time signal timing and intersection safety need to be further investigated.

On the other hand, with the rapid development of connected vehicle technologies in recent years, it is feasible for us to implement efficient proactive traffic management strategies at intersections, e.g., dynamic message sign (DMS) to show the real-time crash risk for the downstream

intersections, and vehicle-level optimal speed advisory through vehicle-to-infrastructure (V2I) communication (Yue et al., 2018). In this context, an efficient and reliable real-time crash risk predictive algorithm for intersections is required. However, traditional intersection safety analyses were usually conducted by modeling historical crash frequency with geometric, AADT, and static signal control characteristics, which ignore the impacts of real-time traffic environment (e.g., traffic and weather) when crashes occur.

To the best of the authors' knowledge, there have been no studies done on the real-time crash risk at signalized intersections. To bridge this gap, this study aims to investigate the relationship between crash likelihood at signalized intersections and real-time traffic, signal timing, and weather characteristics by utilizing data from multiple sources, i.e., Bluetooth, weather, and adaptive signal control datasets.

4.2 Data Preparation

There are 23 intersections chosen from four urban arterials in Orlando, Florida, as shown in Figure 4-1. A total of four datasets were used: (1) crash data from March, 2017 to March, 2018 provided by Signal Four Analytics (S4A); (2) travel speed data collected by 23 IterisVelocity Bluetooth detectors installed at 23 intersections; (3) signal timing and 15-minute interval traffic volume provided by 23 adaptive signal controllers; (4) weather characteristics collected by the nearest airport weather station.

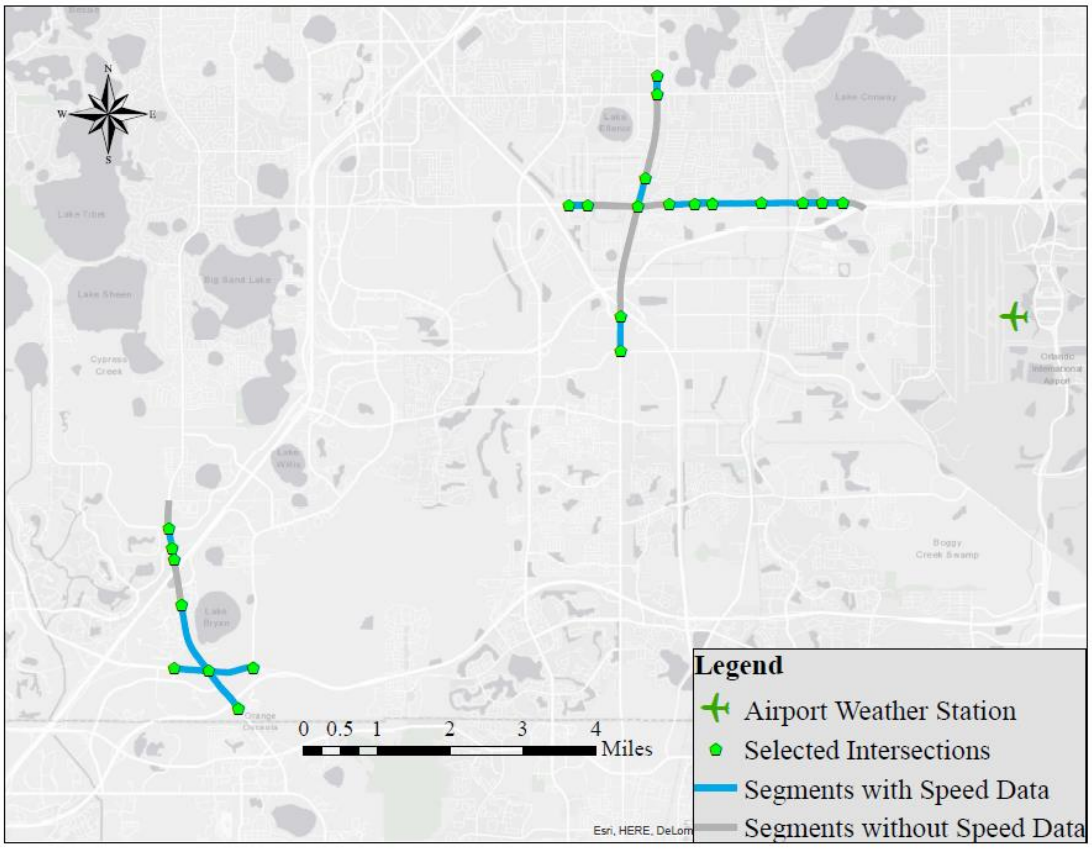


Figure 4-1: Layout of Selected Intersections

S4A provides detailed crash information, including crash time, coordinates, severity, type, weather condition, etc. In terms of the crash time information, there are three kinds of time information for each crash, i.e. time of crash occurrence, time reported, and time dispatched. Only the time of crash occurrence was utilized in this study, and the difference between this recorded crash time and the actual crash time is supposed to be within 5 minutes since there exist several efficient and accurate technologies for the police officer to identify the accurate time of crash occurrence, e.g. closed-circuit television cameras and mobile phones.

First, all crashes occurred at intersection or influenced by intersection (within 250 feet of intersection) from March 2017 to March 2018 were collected. Second, all the single-vehicle crashes and the crashes under the influence of alcohol and drugs were excluded, since these kinds

of crashes are usually not attributed to the real-time traffic and signal characteristics which are the focus of this study. After that, a total of 803 crashes remained and these crashes were divided into three types based on their location, which are within intersection area, intersection entrance area, and intersection exit area, as shown in Figure 4-2. There are 446 (55.54%) crashes that had occurred within intersection, 264 (32.88%) crashes that had occurred in the intersection entrance area, and 93 (11.58%) crashes that had occurred in the intersection exit area. In terms of the sample size, only within intersection crashes and intersection entrance crashes were utilized in this study.

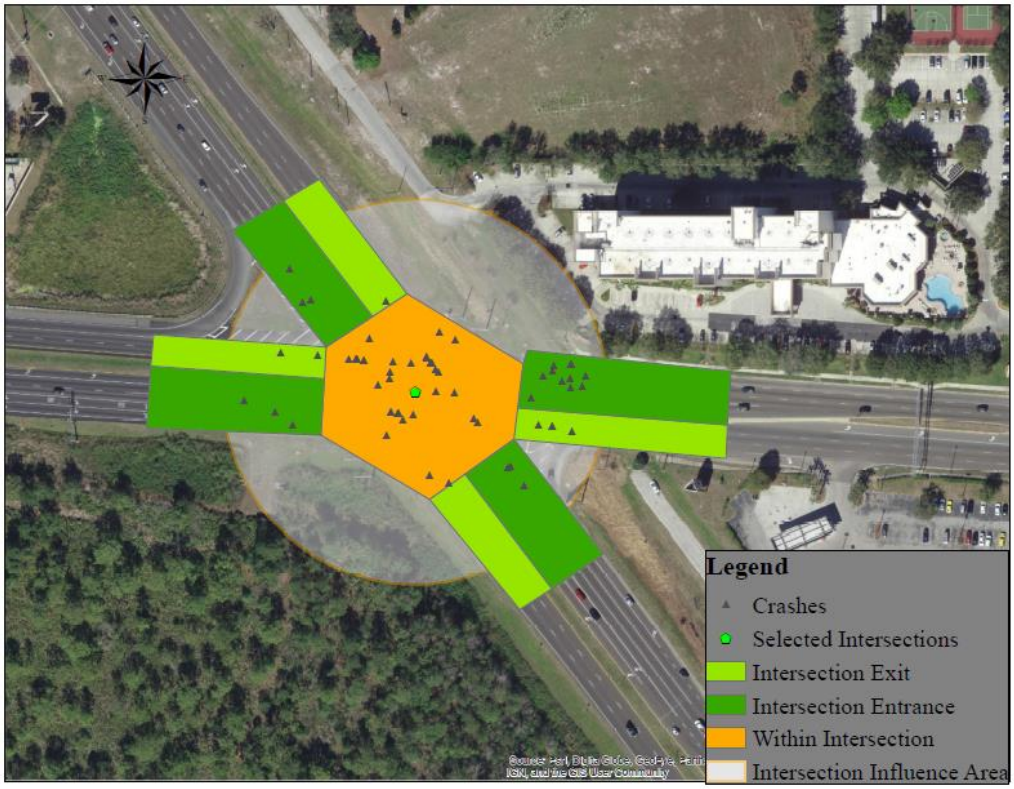


Figure 4-2: Illustration of Three Types of Intersection Crash Location

Before collecting the real-time traffic and signal timing variables for each crash, two preprocess steps were conducted: First, identify the at-fault vehicle travel direction for each crash based on the attribute of "Crash Type Direction", and then rename the approach of at-fault vehicle as "A" approach; Second, retrieve the travel direction of the other three approaches based on the

nomenclature in Figure 4-3, and then rename them as “B”, “C”, and “D” approaches, respectively. After this preprocessing, all the relationship between crash location and intersection approaches were consistent, i.e., the travel approach of at-fault vehicle for all crashes were named as “A” approach and all the other corresponding approaches were named as “B”, “C”, and “D” approaches according to the nomenclature. For the within intersection crash and non-crash events, the real-time traffic and signal timing data were collected from four approaches, while for the intersection entrance crash and non-crash events, only the data from “A” approach were collected.

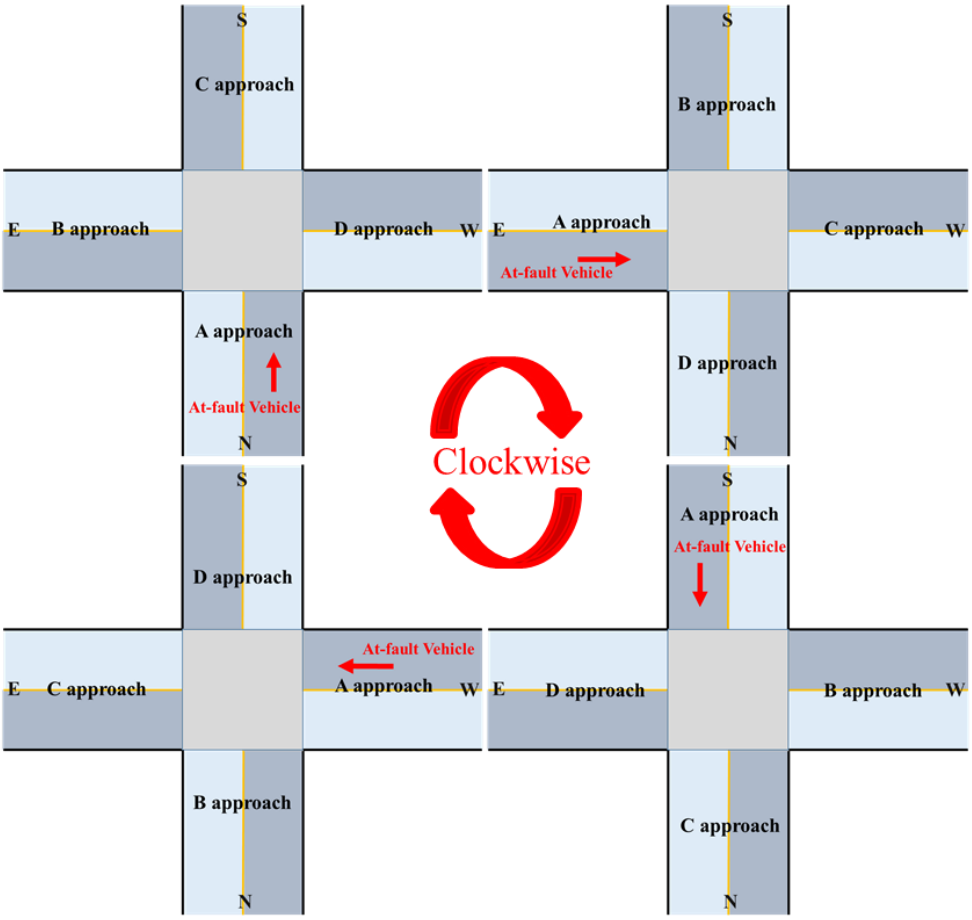


Figure 4-3: The Nomenclature of the Four Approach (“A”, “B”, “C”, and “D”)

Matched case-control design was employed in this study to explore the effects of traffic, signal, and weather-related variables while eliminating the effects of other confounding factors through

the design of study. For each crash, four confounding factors, i.e., intersection ID, crash location type (within intersection or intersection entrance), time of day, and day of week, were selected as matching factors. Therefore, a group of non-crash events could be identified by using these matching factors and then a specific number of non-crash events could be randomly selected from this group of non-crash events for every crash event. The number of non-crash events *m* corresponding to a crash event is preferred to be fixed in the entire analysis. As stated in Hosmer Jr et al. (2013), the value of *m* was commonly chosen from one to five. Moreover, Abdel-Aty et al. (2004) found that there is no significant difference when *m* changing from one to five. Therefore, the control-to-case ratio of 4:1 was adopted in this study, which is consistent with previous studies (Abdel-Aty et al., 2008; Ahmed and Abdel-Aty, 2013; Ahmed et al., 2012b; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2016; Zheng et al., 2010). Consequently, 4 non-crash events from the same intersection, crash location type, time of day, and day of week were randomly selected for each crash event. Figure 4-4 shows an example of the matched case control design for the within intersection crash event. Besides, the non-crash events were selected only when there are no crashes occurring within 3 hours before or after the non-crash event on the same location.

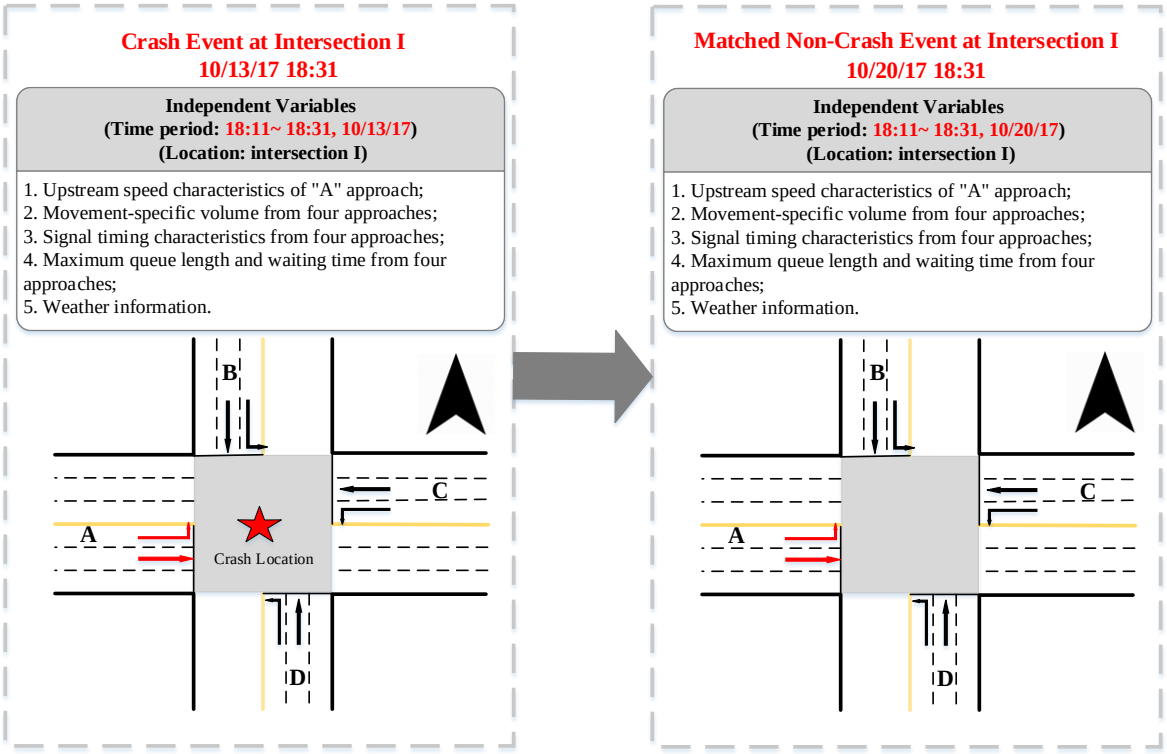


Figure 4-4: Illustration of Matched Case-Control Design for the Within-Intersection Crashes

The real-time traffic and signal timing data for both crash and non-crash events were extracted for a period of 20 minutes (divided into four 5-minute time slices) prior to crash occurrence. For example, if a crash event *i* occurred within intersection at 15:00, the corresponding traffic and signal timing data from 14:40 to 15:00 were extracted and named as time slice 4, 3, 2, and 1, respectively. As shown in Figure 4-5, the traffic and signal timing data collection for different crash location are different. For the within-intersection crashes, all the traffic and signal timing variables from four approaches were collected. However, for the intersection entrance crashes, data were collected only from the "A" approach.

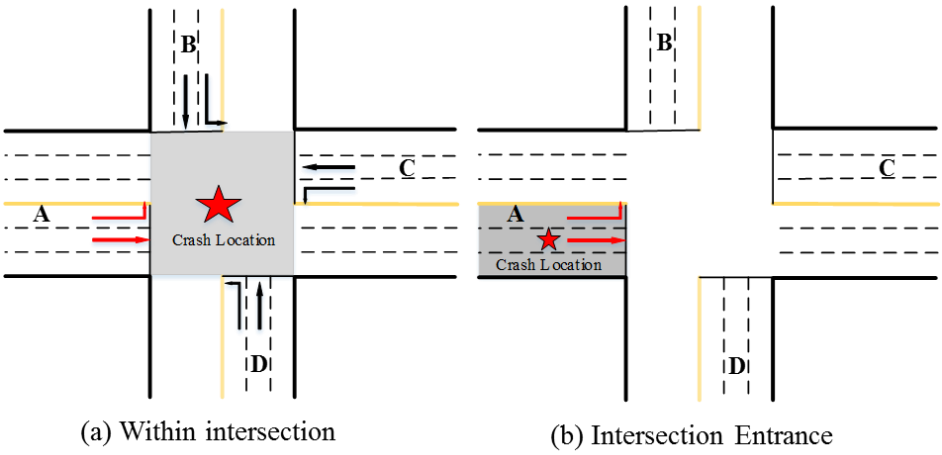


Figure 4-5: Schematic Figure of Crash Location and Data Collection

Speed data were provided by the 23 IterisVelocity Bluetooth detectors, which measure the space-mean speed of a specific segment, as shown in Figure 4-6. Bluetooth detectors can only detect the vehicles equipped with Bluetooth device which is working at discoverable mode. The space-mean speed of each vehicle on a specific segment is calculated as the segment length divided by the travel time of each detected vehicle on the segment based on the detection data of two Bluetooth detectors located at the two contiguous intersections. In this study, speed data, including average speed and speed standard deviation, were only collected for the segment of "A" approach, which represents the traveling segment of the at-fault vehicle. Moreover, since all the Bluetooth detectors are installed on the major arterials, therefore, only the major approaches were provided with the real-time traffic speed data. In this context, all the intersection entrance crashes included in the final datasets were occurred on the major approach, and all the at-fault vehicles of the within intersection crashes were coming from the major approach.

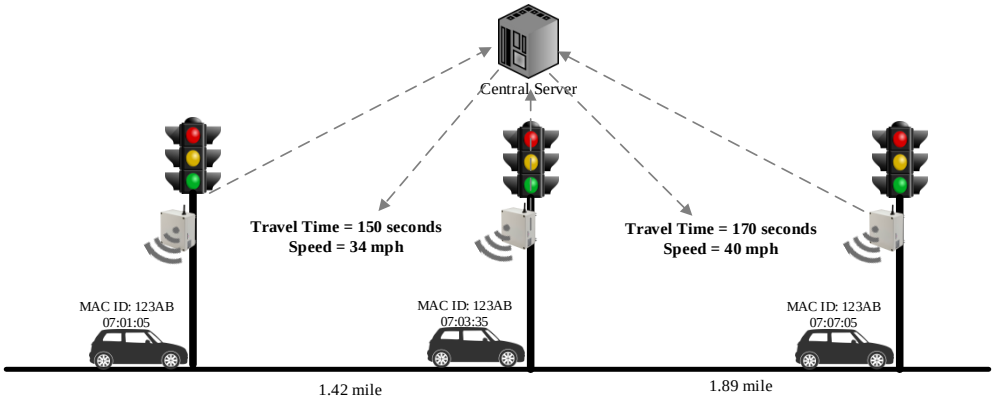


Figure 4-6: Illustration of Bluetooth Data Collection

Adaptive signal controllers archive the real-time signal timing and lane-specific 15-minute aggregate traffic volume data. The lane-specific 15-minute aggregated traffic volume data are collected by the video detectors, which are installed for the adaptive signal controller to detect the real-time volume, queue length and waiting time. Since the right-turn vehicles are unprotected at the intersection, the traffic volume data only include the through and left-turn vehicles. The traffic volume for each time slice (5-minute) was calculated based on the assumption that the traffic volume within 15-minute interval are evenly distributed. Moreover, the variation in traffic flow across lanes in the form of overall average flow ratio (OAFR) were considered in this study. The OAFR was proposed by Lee et al. (2006) to represent a surrogate measure of the lane change frequency within all lanes. The OAFR is calculated as the geometric mean of the modified average flow ratio (AFR) of all lanes, while the modified AFR is calculated as the ratio of the average flow in the adjacent lanes ($i-1, i+1$) to the average flow in the subject lane (i), as shown in Eq. (4-1).

$$AFR_i(t) = \frac{V_{i-1}(t)}{V_i(t)} \times \frac{NL_{i-1,i}(t)}{NL_{i-1,i}(t) + NL_{i-1,i-2}(t)} + \frac{V_{i+1}(t)}{V_i(t)} \times \frac{NL_{i+1,i}(t)}{NL_{i+1,i}(t) + NL_{i+1,i+2}(t)} \quad (4\text{-}1)$$

Where $V_i(t)$ is average flow in the subject lane i during time interval t; $V_{i-1}(t)$ and $V_{i+1}(t)$ are the average flow in the adjacent lanes $i-1$ and $i+1$, respectively during time interval t; $NL_{i-1,i}(t)$ is the number of lane changes from lane $i-1$ to lane i, if lane $i-1$ exists, during time interval t; Similarly, $NL_{i-1,i-2}(t)$, $NL_{i+1,i}(t)$, and $NL_{i+1,i+2}(t)$ represent the number of lane changes from lane $i-1$ to $i-2$, $i+1$ to i, and $i+1$ to $i+2$ during time interval t, respectively. Because the fractions of the number of lane change from lane $i-1$ to lane i and $i-2$, as well as the fractions from lane $i+1$ to lane i and $i+2$, were unknown in this study, they were assumed to be equal, which is in line with Lee et al. (2006).

It is worth noting that the OAFR calculated by Lee et al. (2006) as the geometric mean of the modified average flow ratio (AFR) of all lanes is only appropriate for the segment with lane number greater than 3. If the total lane number is 2, the calculated OAFR will always be 0.5 ($\sqrt[2]{\frac{V_1(t)}{V_2(t)} \times 0.5 \times \frac{V_2(t)}{V_1(t)} \times 0.5}$), no matter with the real flow variation between these two lanes. Therefore, the OAFR in this study was calculated as the arithmetic mean of the modified AFR ($\frac{1}{n}\sum_{i=1}^{n} AFR_i(t)$).

Three weather related variables (weather type, visibility, and hourly precipitation) were collected from the nearest airport weather station, which is located at the Orlando international airport (as shown in Figure 4-1). Since the weather data is not recorded continuously, once the weather condition changes and reaches a preset threshold, a new record will be added to the

archived data. Therefore, for each specific crash, based on the reported crash time, the closest weather record prior to the crash time has been extracted and used as the crash time weather condition, which is identical for four time slices. A cross table was made to validate the weather type information extracted from weather station and the weather condition recorded in the crash report, results indicated that the consistency ((True positive + True negative)/Total sample size) between weather station and crash report is around 92%. Therefore, all the weather information for both crash and non-crash events were extracted from the airport weather station data.

After the above data collection process, the final dataset for the within intersection area includes 470 observations (94 crash events and 376 non-crash events), while the final dataset for the intersection entrance area includes 425 observations (85 crash events and 340 non-crash events). The summary statistics of within intersection and intersection entrance datasets are as shown in Table 4-1 and Table 4-2, separately.

Table 4-1: Summary of Variables Descriptive Statistics for the Within Intersection Area (Crash and Non-crash Events)

Variable	Time Slice	Decription	Crash Events		Non-Crash Events	
			Mean (Std)	(Min, Max)	Mean (Std)	(Min, Max)
Avg_speed	1	Average speed on the upstream segment of "A" approach within 5-minute interval (mph)	25.69 (9.52)	(5.00, 45.57)	26.94 (10.42)	(4.75, 54.00)
	2		27.8 (10.32)	(6.20, 51.67)	27.11 (10.29)	(6.50, 56.00)
	3		27.43 (10.32)	(5.00, 52.00)	27.04 (10.37)	(6.42, 53.00)
	4		26.9 (10.33)	(5.50, 54.00)	27.10 (10.27)	(4.60, 54.75)
Std_speed	1	Speed standard deviation on the upstream segment of "A" approach within 5-minute interval (mph)	10.59 (4.70)	(0.00, 20.92)	9.83 (5.15)	(0.00, 27.58)
	2		9.39 (4.69)	(0.71, 21.21)	10.15 (5.34)	(0.00, 36.77)
	3		10.05 (5.09)	(0.00, 23.33)	10.14 (5.49)	(0.00, 36.06)
	4		10.62 (5.26)	(0.00, 22.19)	10.12 (5.34)	(0.00, 26.87)
A_Vol_LT	1	Left turn volume of "A" approach within 5-minute interval (vehicle)	24.84 (24.14)	(0.00, 133.67)	22.26 (22.48)	(0.00, 186.00)
	2		24.44 (21.93)	(0.00, 125.67)	21.99 (20.96)	(0.00, 186.00)
	3		23.94 (20.84)	(0.00, 125.67)	21.77 (19.67)	(0.00, 177.67)
	4		24.50 (25.00)	(0.00, 192.00)	22.24 (21.48)	(0.00, 177.67)
A_Vol_Th	1	Through volume of "A" approach within 5-minute interval (vehicle)	112.30 (50.86)	(0.00, 298.33)	106.09 (54.43)	(0.00, 481.33)
	2		113.73 (49.08)	(0.00, 298.33)	106.24 (51.07)	(0.00, 404.00)
	3		109.82 (48.26)	(0.00, 259.33)	104.89 (50.61)	(0.00, 369.80)
	4		113.65 (63.11)	(0.00, 416.00)	105.79 (54.53)	(0.00, 405.33)
A_OAFR	1	Overall average flow ratio of "A" approach within 5-minute interval	1.33 (1.22)	(0.94, 11.29)	1.40 (2.25)	(0.94, 38.88)
	2		1.48 (1.79)	(0.95, 11.29)	1.56 (3.09)	(0.94, 37.27)
	3		1.69 (3.26)	(0.94, 29.42)	1.58 (2.7)	(0.94, 30.28)
	4		1.75 (3.38)	(0.94, 29.42)	1.54 (2.64)	(0.94, 30.28)

Variable	Time Slice	Decription	Crash Events		Non-Crash Events	
			Mean (Std)	(Min, Max)	Mean (Std)	(Min, Max)
A_LT_GreenRatio	1	Ratio of left turn green time on "A" approach within 5-minute interval	0.14 (0.07)	(0.02, 0.35)	0.14 (0.08)	(0.02, 0.45)
	2		0.14 (0.07)	(0.03, 0.31)	0.14 (0.08)	(0.02, 0.41)
	3		0.14 (0.07)	(0.03, 0.36)	0.14 (0.07)	(0.02, 0.36)
	4		0.14 (0.08)	(0.02, 0.36)	0.14 (0.08)	(0.01, 0.40)
A_LT_Avg_Green	1	Average length of left turn green phase on "A" approach within 5-minute interval (second)	18.41 (9.49)	(4.00, 46.00)	18.45 (9.74)	(2.50, 50.00)
	2		18.82 (9.16)	(6.40, 41.00)	18.86 (10.67)	(2.00, 67.00)
	3		18.26 (10.05)	(4.50, 57.00)	18.34 (9.61)	(4.00, 61.00)
	4		17.74 (9.18)	(3.50, 47.00)	18.55 (9.66)	(2.00, 64.00)
A_LT_Std_Green	1	Standard deviation of the length of left turn green phase on "A" approach within 5-minute interval (second)	5.90 (6.34)	(0.00, 40.20)	5.90 (5.9)	(0.00, 31.11)
	2		5.91 (5.64)	(0.00, 36.77)	5.13 (5.24)	(0.00, 34.65)
	3		6.52 (6.14)	(0.00, 26.87)	6.69 (6.52)	(0.00, 43.84)
	4		5.28 (4.71)	(0.00, 21.21)	6.32 (5.62)	(0.00, 31.11)
A_LT_Avg_Queue	1	Average left turn queue length at the beginning of left turn green phase on "A" approach (vehicle)	8.39 (6.54)	(1.00, 33.33)	8.90 (7.40)	(0.00, 47.00)
	2		8.56 (6.07)	(0.75, 33.33)	8.74 (7.03)	(0.00, 46.00)
	3		9.58 (7.84)	(0.33, 40.00)	8.59 (6.70)	(0.00, 45.00)
	4		9.03 (7.34)	(0.00, 40.00)	9.10 (7.46)	(0.00, 45.00)
A_LT_Avg_Wait	1	Average left turn maximum waiting time at the beginning of left turn green phase on "A" approach (vehicle)	94.69 (45.35)	(0.50, 167.50)	97.25 (48.07)	(0.00, 266.00)
	2		95.16 (45.79)	(0.50, 179.00)	97.96 (49.07)	(0.00, 241.00)
	3		96.72 (51.78)	(0.40, 279.00)	97.71 (49.14)	(0.00, 246.5)
	4		98.59 (48.58)	(2.50, 169.50)	95.96 (49.48)	(0.00, 284.00)
A_TH_GreenRatio	1	Ratio of through green time on "A" approach within 5-minute interval	0.45 (0.16)	(0.14, 0.86)	0.44 (0.16)	(0.07, 0.88)
	2		0.44 (0.15)	(0.15, 0.85)	0.44 (0.16)	(0.06, 0.92)
	3		0.44 (0.16)	(0.15, 0.85)	0.43 (0.16)	(0.12, 0.84)
	4		0.43 (0.16)	(0.11, 0.90)	0.43 (0.17)	(0.08, 0.89)
A_TH_Avg_Green	1	Average length of through green phase on "A" approach within 5-minute interval (second)	28.88 (17.99)	(11.2, 105.00)	29.42 (19.58)	(9.64, 128.00)
	2		28.66 (17.97)	(9.05, 105.50)	29.47 (21.6)	(7.00, 137.5)
	3		28.67 (19.75)	(11.29, 105.50)	29.32 (20.65)	(8.89, 122.00)
	4		28.11 (17.23)	(8.00, 82.50)	29.09 (19.68)	(9.33, 133.00)
A_TH_Std_Green	1	Standard deviation of the length of through green phase on "A" approach within 5-minute interval (second)	18.26 (12.76)	(0.00, 60.25)	18.78 (13.9)	(0.00, 99.51)
	2		18.21 (11.60)	(0.00, 51.04)	18.24 (13.27)	(0.00, 89.8)
	3		18.12 (12.50)	(0.00, 60.09)	18.19 (12.3)	(0.00, 68.14)
	4		20.04 (14.57)	(0.00, 64.55)	18.81 (15.05)	(0.00, 164.05)
A_TH_Avg_Queue	1	Average through queue length at the beginning of through green phase on "A" approach (vehicle)	12.55 (8.97)	(2.08, 40.00)	12.54 (8.84)	(0.8, 54.00)
	2		11.94 (8.59)	(2.00, 40.00)	12.77 (9.14)	(0.00, 57.00)
	3		12.40 (8.84)	(1.50, 40.00)	12.87 (9.13)	(0.33, 57.00)
	4		12.09 (8.71)	(2.00, 40.00)	12.82 (9.32)	(1.27, 62.00)
A_TH_Avg_Wait	1	Average through maximum waiting time at the beginning of through green phase on "A" approach (vehicle)	36.49 (25.78)	(1.29, 135.50)	36.10 (27.68)	(0.00, 142.00)
	2		35.48 (24.94)	(1.67, 135.50)	36.21 (28.37)	(0.00, 175)
	3		36.56 (24.12)	(0.00, 135.50)	37.06 (28.17)	(0.00, 192.5)
	4		36.41 (25.49)	(4.2, 135.00)	37.27 (29.67)	(0.00, 213.00)
HourlyPrecip	-	Hourly precipitation (1/10 inch)	0.03 (0.10)	(0.00, 0.70)	0.09 (0.65)	(0.00, 8.00)
Visibility	-	Visibility (mile)	9.86 (0.68)	(5.00, 10.00)	9.64 (1.51)	(0.00, 10.00)
WeatherType	-	Weather type: 0 for normal and 1 for adverse weather	0.11 (0.31)	(0.00, 1.00)	0.09 (0.28)	(0.00, 1.00)

Note: due to the limitation of table content, this table only list the "A" approach data. However, the within intersection dataset including the data from four approaches.

Table 4-2: Summary of Variables Descriptive Statistics for the Intersection Entrance Area (Crash and Non-crash Events)

Variable	Time Slice	Decription	Crash Events		Non-Crash Events	
			Mean (Std)	(Min, Max)	Mean (Std)	(Min, Max)
Avg_speed	1	Average speed on the upstream segment of "A" approach within 5-minute interval (mph)	25.61 (8.63)	(7.75, 42.00)	26.77 (9.44)	(5.33, 53.50)
	2		27.16 (9.66)	(5, 45.17)	26.77 (9.24)	(4.83, 56.50)
	3		27.24 (9.88)	(4.75, 50.33)	27.35 (10.05)	(5.17, 55.14)
	4		26.94 (8.82)	(4.00, 47.67)	27.09 (10.11)	(6.00, 57.50)
Std_speed	1	Speed standard deviation on the upstream segment of "A" approach within 5-minute interval (mph)	10.49 (4.96)	(0.00, 25.36)	11.02 (5.06)	(0.00, 28.28)
	2		11.28 (5.04)	(0.53, 24.02)	11.03 (5.14)	(0.58, 31.11)
	3		11.06 (4.9)	(0.96, 24.75)	10.61 (4.81)	(0.58, 25.46)
	4		11.72 (4.75)	(0.00, 23.83)	10.86 (4.88)	(0.00, 29.70)
A_Vol_LT	1	Left turn volume of "A" approach within 5-minute interval (vehicle)	16.46 (12.31)	(0.00, 55.67)	18.30 (13.68)	(0.00, 101.33)
	2		16.24 (12.00)	(0.00, 55.67)	18.47 (13.81)	(0.00, 101.33)
	3		16.14 (11.54)	(0.00, 55.67)	18.09 (13.41)	(0.00, 92.93)
	4		15.74 (10.36)	(0.00, 46.00)	17.84 (13.08)	(0.00, 80.33)
A_Vol_Th	1	Through volume of "A" approach within 5-minute interval (vehicle)	108.67 (64.00)	(0.00, 343.33)	107.18 (62.64)	(0.00, 614.33)
	2		108.49 (63.69)	(0.00, 343.33)	107.28 (61.21)	(0.00, 614.33)
	3		108.28 (64.13)	(0.00, 309.53)	106.2 (55.34)	(0.00, 360.00)
	4		108.1 (63.94)	(0.00, 328.33)	105.98 (55.48)	(0.00, 360.00)
A_OAFR	1	Overall average flow ratio of "A" approach within 5-minute interval	1.74 (3.23)	(0.95, 21.56)	1.61 (3.26)	(0.94, 36.04)
	2		1.94 (3.63)	(0.95, 21.56)	1.64 (3.32)	(0.94, 32.95)
	3		1.77 (3.07)	(0.95, 21.56)	1.65 (3.3)	(0.94, 32.95)
	4		1.51 (2.03)	(0.95, 16.68)	1.65 (3.67)	(0.95, 43.45)
A_LT_GreenRatio	1	Ratio of left turn green time on "A" approach within 5-minute interval	0.13 (0.07)	(0.03, 0.33)	0.13 (0.06)	(0.02, 0.36)
	2		0.12 (0.06)	(0.03, 0.31)	0.12 (0.06)	(0.02, 0.39)
	3		0.13 (0.07)	(0.03, 0.36)	0.13 (0.07)	(0.02, 0.37)
	4		0.12 (0.06)	(0.00, 0.26)	0.13 (0.07)	(0.01, 0.35)
A_LT_Avg_Green	1	Average length of left turn green phase on "A" approach within 5-minute interval (second)	18.41 (9.44)	(4.00, 44.50)	18.27 (9.42)	(3.00, 68.00)
	2		15.88 (7.53)	(5.00, 39.00)	17.46 (8.07)	(4.50, 50.00)
	3		17.82 (9.58)	(5.00, 60.00)	18.02 (8.75)	(5.00, 48.50)
	4		16.39 (7.76)	(1.00, 39.00)	18.21 (8.91)	(3.00, 46.00)
A_LT_Std_Green	1	Standard deviation of the length of left turn green phase on "A" approach within 5-minute interval (second)	4.17 (4.86)	(0.00, 18.50)	4.92 (4.86)	(0.00, 33.94)
	2		5.54 (5.02)	(0.00, 23.52)	6.17 (6.40)	(0.00, 31.82)
	3		5.61 (5.41)	(0.00, 24.04)	5.47 (5.11)	(0.00, 24.75)
	4		4.92 (4.90)	(0.00, 23.33)	5.09 (5.10)	(0.00, 22.63)
A_LT_Avg_Queue	1	Average left turn queue length at the beginning of left turn green phase on "A" approach (vehicle)	8.53 (6.03)	(1.00, 31.00)	8.69 (6.51)	(0.00, 37.00)
	2		7.92 (5.74)	(0.33, 31.00)	8.73 (6.71)	(0.00, 37.00)
	3		8.00 (6.30)	(0.75, 35.00)	8.18 (6.23)	(0.00, 37.50)
	4		7.81 (5.64)	(1.00, 34.00)	8.09 (6.28)	(0.00, 38.00)
A_LT_Avg_Wait	1	Average left turn maximum waiting time at the beginning of left turn green phase on "A" approach (vehicle)	102.36 (45.84)	(10.00, 178.00)	101.86 (46.77)	(0.00, 291.00)
	2		95.15 (46.45)	(4.67, 169.00)	103.11 (45.32)	(0.00, 185.5)
	3		96.13 (45.44)	(4.25, 188.00)	101.72 (42.78)	(0.00, 202.00)
	4		97.65 (43.31)	(5.50, 170.50)	100.31 (43.37)	(0.00, 182.00)
A_TH_GreenRatio	1	Ratio of through green time on "A" approach within 5-minute interval	0.44 (0.17)	(0.08, 0.84)	0.44 (0.18)	(0.11, 0.89)
	2		0.43 (0.18)	(0.12, 0.80)	0.44 (0.20)	(0.08, 1.00)
	3		0.44 (0.19)	(0.15, 0.83)	0.44 (0.19)	(0.09, 0.93)
	4		0.45 (0.18)	(0.08, 0.87)	0.45 (0.18)	(0.11, 0.86)
A_TH_Avg_Green	1	Average length of through green phase on "A" approach within 5-minute interval (second)	31.24 (22.11)	(10.13, 126.00)	29.57 (19.91)	(9.13, 134.00)
	2		30.71 (19.41)	(12.00, 105.50)	31.07 (24.07)	(7.00, 137.5)
	3		29.49 (21.13)	(11.3, 123.00)	28.93 (17.72)	(8.75, 132.00)
	4		29.32 (20.97)	(9.29, 131.00)	30.28 (21.35)	(10.00, 126.00)
A_TH_Std_Green	1	Standard deviation of the length of through green phase on "A" approach within 5-minute interval (second)	21.76 (21.10)	(0.00, 164.05)	21.04 (15.51)	(0.00, 148.49)
	2		21.74 (14.01)	(0.00, 63.52)	22.77 (21.03)	(0.00, 183.14)
	3		18.84 (14.12)	(0.00, 63.02)	20.30 (15.06)	(0.00, 74.08)
	4		18.57 (13.67)	(0.71, 58.29)	20.65 (15.73)	(0.00, 128.69)

Variable	Time Slice	Decription	Crash Events		Non-Crash Events	
			Mean (Std)	(Min, Max)	Mean (Std)	(Min, Max)
A_TH_Avg_Queue	1	Average through queue length at the beginning of through green phase on "A" approach (vehicle)	15.41 (10.82)	(2.00, 72.00)	12.78 (9.77)	(1.33, 99.00)
	2		14.84 (10.49)	(0.00, 72.00)	12.83 (9.91)	(1.33, 99.00)
	3		13.60 (10.38)	(2.50, 74.50)	12.56 (9.87)	(1.31, 99.00)
	4		13.08 (10.43)	(1.45, 77.00)	12.61 (9.95)	(1.43, 99.00)
A_TH_Avg_Wait	1	Average through maximum waiting time at the beginning of through green phase on "A" approach (vehicle)	41.74 (31.75)	(1.60, 155.00)	38.92 (29.42)	(0.00, 140.00)
	2		44.37 (33.46)	(0.00, 144.00)	39.10 (31.11)	(0.00, 171.00)
	3		39 (32.30)	(2.25, 143.00)	38.53 (30.57)	(0.50, 156.00)
	4		37.38 (32.27)	(0.33, 148.50)	38.03 (30.30)	(0.00, 156.00)
HourlyPrecip	-	Hourly precipitation (1/10 inch)	0.06 (0.41)	(0.00, 3.70)	0.11 (0.71)	(0.00, 6.90)
Visibility	-	Visibility (mile)	9.76 (0.92)	(5.00, 10.00)	9.62 (1.56)	(0.00, 10.00)
WeatherType	-	Weather type: 0 for normal and 1 for adverse weather	0.09 (0.29)	(0.00, 1.00)	0.10 (0.30)	(0.00, 1.00)

In order to achieve a preliminary understanding about the difference between crash and non-crash events, the variable of average speed was selected as an example and the probability density distributions were presented in Figure 4-7 (within intersection) and Figure 4-8 (intersection entrance). Both Figure 4-7 and Figure 4-8 indicate that the distribution of average speed before crash events are more likely to be wide-spread than non-crash events, especially during the 5-10 minute interval. This means that the traffic condition before crash event tends to be more diverse than non-crash events, which is consistent with Theofilatos et al. (2018a).

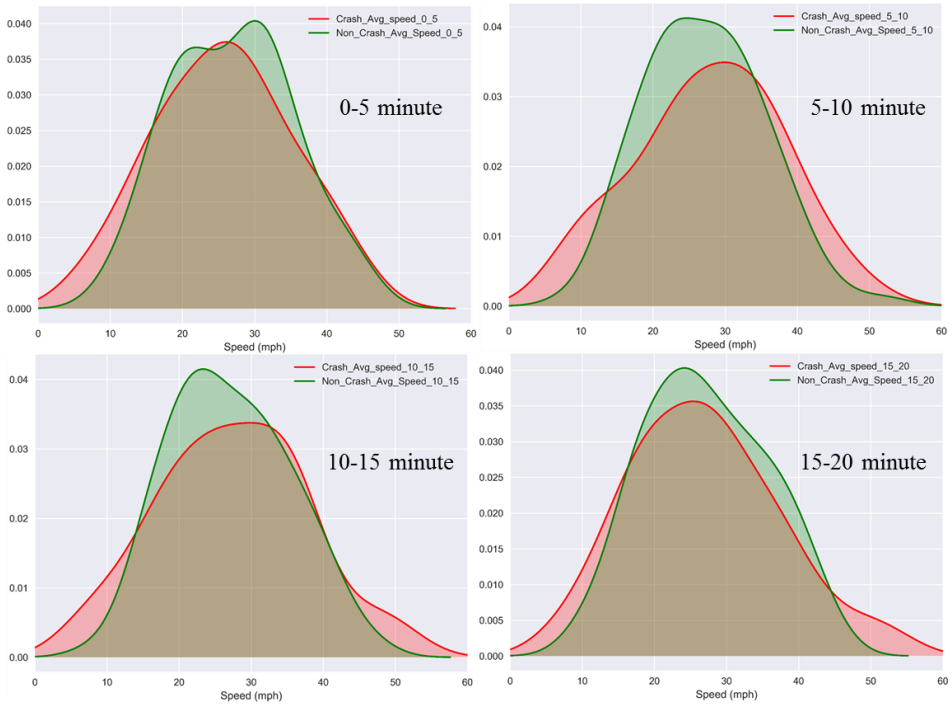


Figure 4-7: Distribution of the Average Speed between Crash and Non-Crash Events among Four Time Slices (Within Intersection).

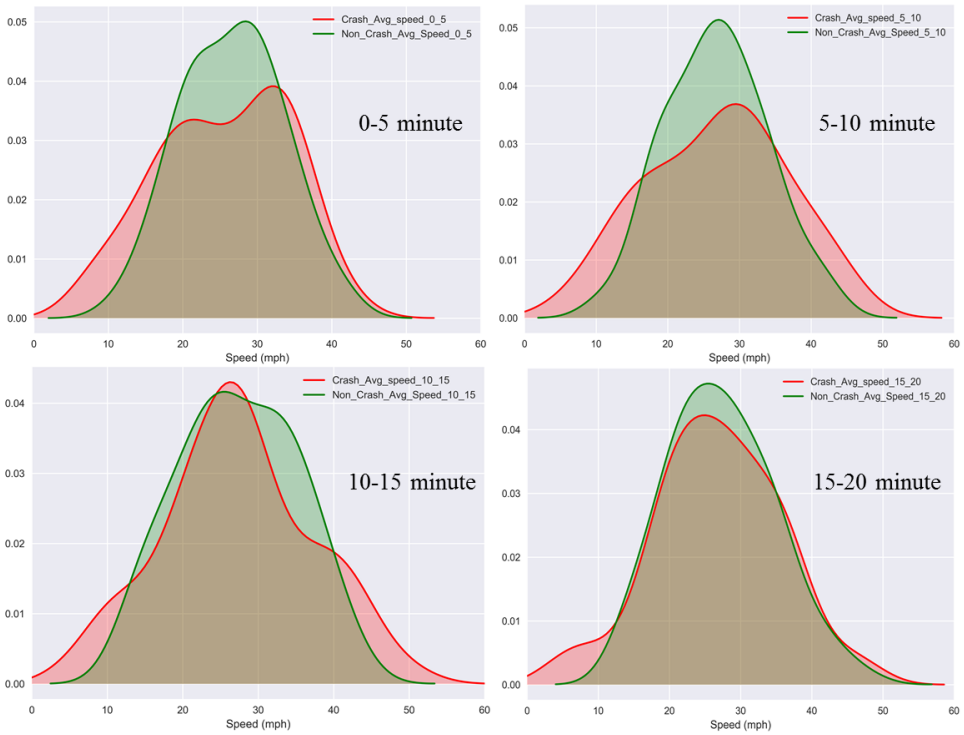


Figure 4-8: Distribution of the Average Speed between Crash and Non-Crash Events among Four Time Slices (Intersection Entrance).

Since the intersection characteristics between different approaches are highly interactive, it is very likely that some of the independent variables are highly correlated. Therefore, two sample correlation matrix for within intersection and intersection entrance datasets, as shown in Figure 4-9 and Figure 4-10, were generated to identify and exclude highly correlated variables.

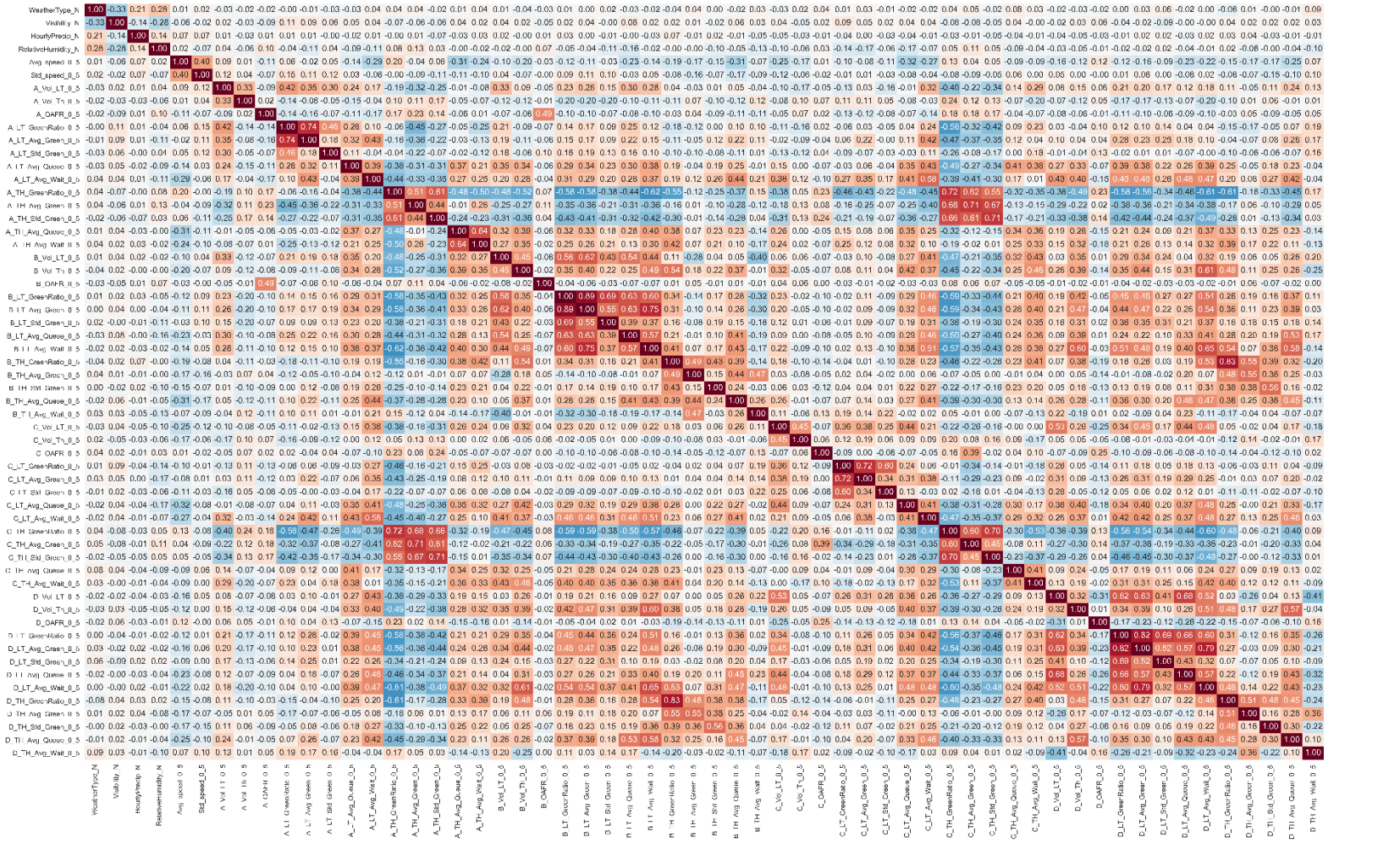


Figure 4-9: Variable Correlation Plot of the Within Intersection Dataset (time-slice 1)

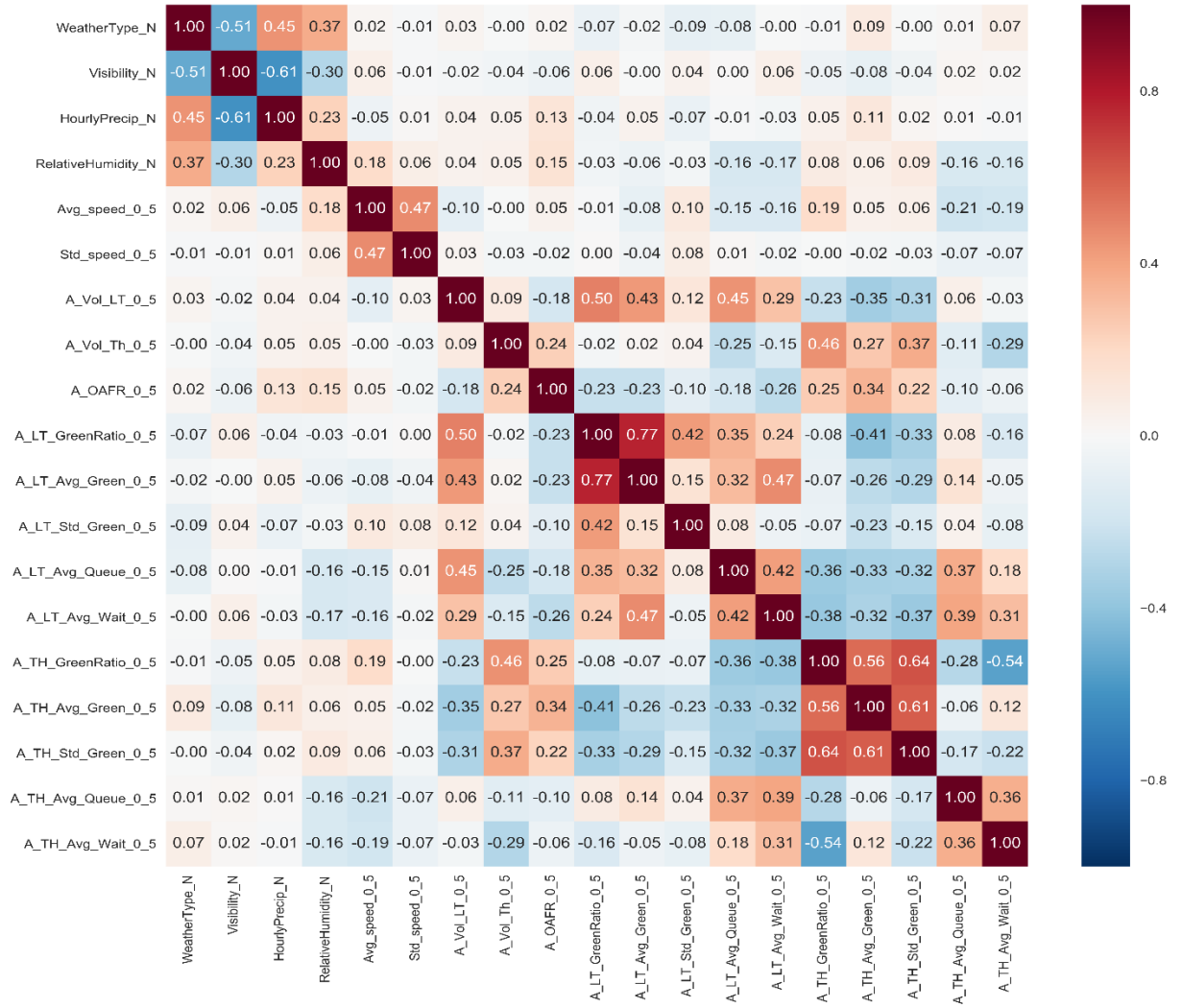


Figure 4-10: Variable Correlation Plot of the Intersection Entrance Dataset (time-slice 1)

The threshold of 0.6 was utilized for the linear Pearson correlation analysis to identify the highly-correlated variables, which is in line with previous research (Kobelo et al., 2008). Moreover, with respect to the nonlinear correlation, one of the mutual information based measures, maximal information coefficient (MIC) was also employed to identify the nonlinear association between two variables (Albanese et al., 2018). As suggested by Albanese et al. (2018), the threshold of MIC was chosen to be 0.7. Above all, the highly correlated pairs of variables were selected based on two criteria: the Pearson correlation coefficient is greater than 0.6 or the MIC is greater than 0.7.

Take the time slice 1 dataset for the within intersection crashes as an example, there are 57 independent variables, which could result in 1596 ($\frac{57!}{2!(57-2)!}$) pairs of variables. The results of correlation analysis indicate that 45 pairs of highly correlated variables were identified and presented in Table 4-3.

Table 4-3: The Highly Correlated Variables for the Within Intersection Dataset (Time Slice 1)

Variables		Pearson Correlation Coefficient	Maximal Information Coefficient (MIC)
A_LT_GreenRatio_0_5	A_LT_Avg_Green_0_5	0.736663	0.480014
A_TH_GreenRatio_0_5	A_TH_Std_Green_0_5	0.610484	0.37036
A_TH_GreenRatio_0_5	B_LT_Avg_Wait_0_5	-0.618134	0.433272
A_TH_GreenRatio_0_5	C_TH_GreenRatio_0_5	0.718776	0.384575
A_TH_GreenRatio_0_5	C_TH_Avg_Green_0_5	0.622481	0.312943
A_TH_GreenRatio_0_5	D_LT_Avg_Wait_0_5	-0.60694	0.399255
A_TH_GreenRatio_0_5	D_TH_GreenRatio_0_5	-0.607984	0.333878
A_TH_Avg_Green_0_5	C_TH_GreenRatio_0_5	0.676753	0.398096
A_TH_Avg_Green_0_5	C_TH_Avg_Green_0_5	0.705255	0.534972
A_TH_Avg_Green_0_5	C_TH_Std_Green_0_5	0.673996	0.533665
A_TH_Std_Green_0_5	C_TH_GreenRatio_0_5	0.65504	0.494755
A_TH_Std_Green_0_5	C_TH_Avg_Green_0_5	0.608811	0.479297
A_TH_Std_Green_0_5	C_TH_Std_Green_0_5	0.7133	0.538673
A_TH_Avg_Queue_0_5	A_TH_Avg_Wait_0_5	0.6426	0.364688
B_Vol_LT_0_5	B_LT_GreenRatio_0_5	0.581439	0.709964
B_Vol_LT_0_5	B_LT_Avg_Green_0_5	0.617656	0.709964
B_Vol_LT_0_5	B_LT_Avg_Queue_0_5	0.537864	0.743575
B_Vol_LT_0_5	B_LT_Avg_Wait_0_5	0.443236	0.727099
B_Vol_Th_0_5	D_LT_Avg_Wait_0_5	0.612108	0.421989
B_LT_GreenRatio_0_5	B_LT_Avg_Green_0_5	0.887087	0.974562
B_LT_GreenRatio_0_5	B_LT_Std_Green_0_5	0.692994	0.746113
B_LT_GreenRatio_0_5	B_LT_Avg_Queue_0_5	0.629947	0.957971
B_LT_GreenRatio_0_5	B_LT_Avg_Wait_0_5	0.596116	0.957971
B_LT_Avg_Green_0_5	B_LT_Std_Green_0_5	0.553333	0.716639
B_LT_Avg_Green_0_5	B_LT_Avg_Queue_0_5	0.632579	0.960151
B_LT_Avg_Green_0_5	B_LT_Avg_Wait_0_5	0.752837	0.960151
B_LT_Avg_Queue_0_5	B_LT_Avg_Wait_0_5	0.574885	0.976834
B_LT_Avg_Queue_0_5	D_LT_Avg_Queue_0_5	0.325929	0.71544
B_LT_Avg_Queue_0_5	D_LT_Avg_Wait_0_5	0.406182	0.71544
B_LT_Avg_Wait_0_5	D_LT_Avg_Queue_0_5	0.400766	0.708458
B_LT_Avg_Wait_0_5	D_LT_Avg_Wait_0_5	0.649914	0.669127
B_TH_GreenRatio_0_5	D_TH_GreenRatio_0_5	0.831784	0.625065
C_LT_GreenRatio_0_5	C_LT_Avg_Green_0_5	0.721476	0.557538
C_TH_GreenRatio_0_5	C_TH_Std_Green_0_5	0.695453	0.432222
D_Vol_LT_0_5	D_LT_GreenRatio_0_5	0.618642	0.837144
D_Vol_LT_0_5	D_LT_Avg_Green_0_5	0.634303	0.838908
D_Vol_LT_0_5	D_LT_Avg_Queue_0_5	0.68267	0.890338
D_Vol_LT_0_5	D_LT_Avg_Wait_0_5	0.521591	0.887042
D_LT_GreenRatio_0_5	D_LT_Avg_Green_0_5	0.821751	0.965754
D_LT_GreenRatio_0_5	D_LT_Std_Green_0_5	0.69115	0.728799
D_LT_GreenRatio_0_5	D_LT_Avg_Queue_0_5	0.657098	0.922204
D_LT_GreenRatio_0_5	D_LT_Avg_Wait_0_5	0.601375	0.922204
D_LT_Avg_Green_0_5	D_LT_Avg_Queue_0_5	0.567144	0.913977
D_LT_Avg_Green_0_5	D_LT_Avg_Wait_0_5	0.788658	0.913977
D_LT_Avg_Queue_0_5	D_LT_Avg_Wait_0_5	0.57161	0.975712

With respect to those pairs of variables which have higher nonlinear correlation coefficients (MIC) but lower linear Pearson correlation coefficients (rows marked in grey in Table 4-3), a scatterplot matrix was generated to further illustrate the nonlinear association between those pairs of variables

(Figure 4-11).

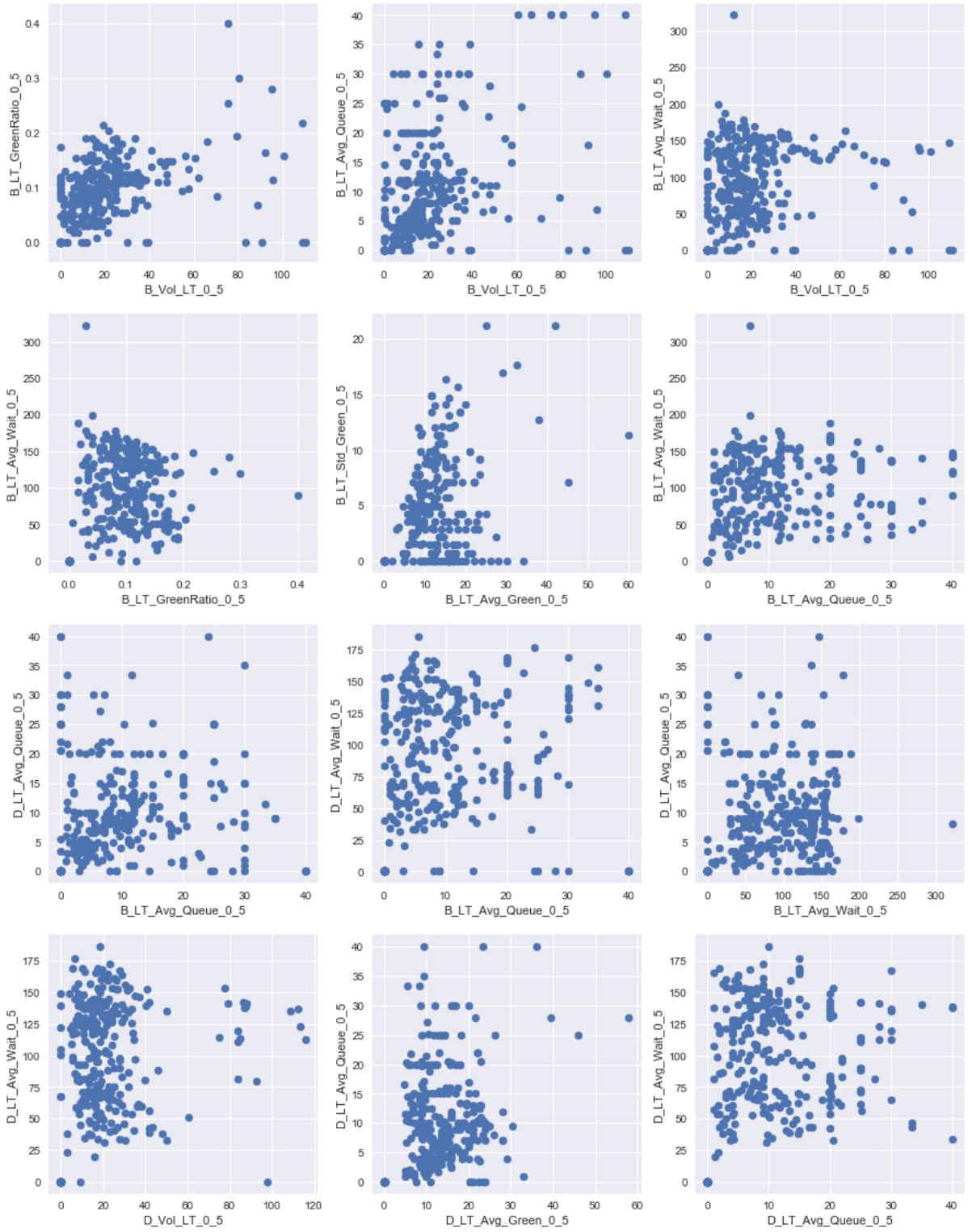


Figure 4-11: Scatterplot Matrix for those Variables which are Nonlinear Associated

4.3 Methodology

Suppose that there are N strata with 1 crash (y_{ij}=1) and m non-crash cases (y_{ij}=0) in stratum i, i=1, 2, …, N. Let p_{ij} be the probability that the *j*th observation in the *i*th stratum is a crash; j=0, 1, 2, …, m. This crash probability could be expressed as:

$$y_{ij} \sim Bernoulli(p_{ij}) \quad (4\text{-}2)$$

$$logit(p_{ij}) = \alpha_i + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij} \quad (4\text{-}3)$$

Where α_i is the intercept term for the *i*th stratum; $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$ is the vector of regression coefficients for *k* independent variables.

In order to take the stratification in the analysis of the observed data, the stratum-specific intercept α_i is considered to be nuisance parameters, and the conditional likelihood for the *i*th stratum would be expressed as (Hosmer Jr et al., 2013):

$$l_i(\boldsymbol{\beta}) = \frac{\exp(\sum_{u=1}^{k} \beta_u X_{ui0})}{\sum_{j=0}^{m} \exp(\sum_{u=1}^{k} \beta_u X_{uij})} \quad (4\text{-}4)$$

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

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} l_i(\boldsymbol{\beta}) \quad (4\text{-}5)$$

Since the full conditional likelihood is independent of stratum-specific intercept α_i, thus Eq. (4-3) cannot be used to estimate the crash probabilities. However, the estimated $\boldsymbol{\beta}$ coefficients are the log-odd ratios of corresponding variables and can be used to approximate the relative risk of an event. Furthermore, the log-odds ratios can also be used to develop a prediction model under this matched case-control analysis. Suppose two observation vectors $\boldsymbol{X_{i1}} = (X_{1i1}, X_{2i1}, \dots, X_{Ki1})$ and

$\boldsymbol{X_{i2}} = (X_{1i2}, X_{2i2}, \ldots, X_{Ki2})$ from the *i*th strata, the odds ratio of crash occurrence caused by observation vector $\boldsymbol{X_{i1}}$ relative to observation vector $\boldsymbol{X_{i2}}$ could be calculated as:

$$\frac{p_{i1}/(1-p_{i1})}{p_{i2}/(1-p_{i2})} = \exp[\sum_{k=1}^{K} \beta_k (X_{ki1} - X_{ki2})] \qquad (4\text{-}6)$$

The right-hand side of Eq. (4-6) is independent of α_i and can be calculated using the estimated $\boldsymbol{\beta}$ coefficients. Thus, the above relative odds ratio could be utilized for predicting crash occurrences by replacing $\boldsymbol{X_{i2}}$ with the vector of the independent variables in the *i*th stratum of non-crash events. One may use simple average of each variable for all non-crash observations within the stratum. Let $\bar{\boldsymbol{X}}_{\boldsymbol{i}} = (\bar{X}_{1i}, \bar{X}_{2i}, \ldots, \bar{X}_{Ki})$ denote the vector of mean values of non-crash events of the *k* variables within the *i*th stratum. Then the odds ratio of a crash relative to the non-crash events in the *i*th stratum could be approximated by:

$$\frac{p_{i1}/(1-p_{i1})}{p_{\bar{i}}/(1-p_{\bar{i}})} = \exp[\sum_{k=1}^{K} \beta_k (X_{ki1} - \bar{X}_{ki})] \qquad (4\text{-}7)$$

Full Bayesian inference was employed in this study. For each model, three chains of 20,000 iterations were set up in WinBUGS (Lunn et al., 2000), the first 5,000 iterations were excluded as burn-in, the latter 15,000 stored iterations were set to estimate the posterior distribution. Convergence was evaluated using the built-in Brooks-Gelman-Rubin (BGR) diagnostic statistic (Brooks and Gelman, 1998).

In terms of model goodness-of-fit, the AUC value which is the area under Receiver Operating Characteristic (ROC) curve was also adopted. The ROC curve illustrates the relationship between the true positive rate (sensitivity) and the false alarm rate (1–specificity) of model classification results based on a given threshold from 0 to 1. It is worth noting that the classification results of

Bayesian random parameters logistic model is based on the predicted crash probabilities, which lie in the range of 0 to 1, while the classification result of Bayesian conditional logistic model and Bayesian random parameters conditional logistic model are based on the predicted odds ratio, which may be larger than 1. In order to be consistent with the other two models, all the odds ratios predicted by Bayesian conditional logistic model were divided by the maximum odds ratio to create adjusted odds ratios. Later, the adjusted odds ratios were used to create the classification result based on different threshold from 0 to 1. In this study, AUC values were calculated using R package pROC (Robin et al., 2011).

4.4 Model Results

4.4.1 Within intersection crashes

This section discusses the modeling results of the Bayesian conditional logistic models for the within intersection crashes based on the full dataset (four time slices) and different time slices datasets, respectively. Table 4-4 shows the results of within intersection model based on full dataset. In total, 14 variables were identified to be significant variables, including speed characteristics, signal timing, queue length, and waiting time related factors collected from different approaches and time slices.

Table 4-4: Results of the Bayesian Conditional Logistic Model based on Full Dataset (Within Intersection).

Variables	Coefficient Estimation		Odds Ratio	
	Mean	95% BCI	Mean	95% BCI
Avg_speed_0_5	**-0.038**	**(-0.07, -0.005)***	**0.963**	**(0.932, 0.995)***
Std_speed_0_5	**0.066**	**(0.001, 0.131)**	**1.068**	**(1.001, 1.14)**
B_TH_Avg_Wait_0_5	**0.013**	**(0.002, 0.024)**	**1.013**	**(1.002, 1.024)**
D_TH_Avg_Wait_0_5	**0.016**	**(0.006, 0.026)**	**1.016**	**(1.006, 1.026)**
B_LT_Std_Green_5_10	**-0.138**	**(-0.248, -0.04)**	**0.871**	**(0.78, 0.961)**
C_TH_Avg_Wait_5_10	**0.017**	**(0.001, 0.032)***	**1.017**	**(1.001, 1.033)***
B_Vol_LT_10_15	**0.029**	**(0.005, 0.054)***	**1.029**	**(1.005, 1.055)***
D_TH_Avg_Green_10_15	**-0.059**	**(-0.103, -0.017)**	**0.943**	**(0.902, 0.983)**
A_LT_Avg_Green_15_20	**-0.055**	**(-0.106, -0.006)**	**0.946**	**(0.899, 0.994)**
A_LT_Std_Green_15_20	**-0.090**	**(-0.161, -0.019)**	**0.914**	**(0.851, 0.981)**
C_LT_Avg_Queue_15_20	**-0.094**	**(-0.18, -0.013)**	**0.910**	**(0.835, 0.987)**
D_TH_GreenRatio_15_20	**-0.088**	**(-0.175, -0.004)**	**0.916**	**(0.839, 0.996)**
D_TH_Std_Green_15_20	**0.060**	**(0.004, 0.114)**	**1.062**	**(1.004, 1.121)**
D_TH_Avg_Queue_15_20	**-0.067**	**(-0.13, -0.005)**	**0.935**	**(0.878, 0.995)**
AUC	0.7596			

*Note: 95% BCI values marked in bold and noted by * indicate that these variables are significant at the 0.1 level, while other variables are significant at the 0.05 level.*

Considering that the traffic and signal characteristics during different time slice may have different relationship with the real-time crash risk. To investigate the differences between different time-slice datasets, four separate time-slice models were developed based on four time slices, respectively. Table 4-5 shows the results of 4 time-slice models for the within intersection dataset. The model comparison results based on AUC values indicate that the slice 2 model performs the

best, followed by the slice 4 and slice 1 models. However, based on slice 1 model, there would be no spare time to implement any proactive traffic management strategy to prevent the possibility of crash occurrence. Moreover, as stated by Golob et al. (2004), there may exist 2.5 min difference between the exact crash time and reported crash time, thus the slice 1 model was treated as a reference. Finally, the slice 2 model was selected to conduct further interpretation.

Table 4-5: Results of Bayesian Conditional Logistic Regression Models based on Different Time Slices (Within Intersection)

Variables	Slice 1		Slice 2		Slice 3		Slice 4	
	Mean (95% BCI)	Odds Ratio (95% BCI)	Mean (95% BCI)	Odds Ratio (95% BCI)	Mean (95% BCI)	Odds Ratio (95% BCI)	Mean (95% BCI)	Odds Ratio (95% BCI)
Avg_speed	**-0.033 (-0.063, -0.004)***	**0.968 (0.939, 0.996)***	-	-	-	-	-	-
Std_speed	**0.056 (0.008, 0.101)***	**1.058 (1.008, 1.106)***	-	-	-	-	-	-
A_Vol_Th	-	-	**0.005 (0.001, 0.011)***	**1.005 (1.001, 1.011)***	-	-	-	-
A_LT_Avg_Green	-	-	-	-	-	-	**-0.041 (-0.08, -0.003)***	**0.96 (0.923, 0.997)***
A_LT_Std_Green	-	-	-	-	-	-	**-0.064 (-0.131, -0.004)**	**0.938 (0.877, 0.996)**
B_Vol_LT	**0.034 (0.009, 0.063)**	**1.035 (1.009, 1.065)**	**0.039 (0.011, 0.07)**	**1.040 (1.011, 1.073)**	**0.031 (0.005, 0.058)**	**1.031 (1.005, 1.06)**	**0.036 (0.006, 0.066)**	**1.037 (1.006, 1.068)**
B_LT_Std_Green	-	-	**-0.106 (-0.206, -0.017)**	**0.899 (0.814, 0.983)**	-	-	-	-
B_TH_Avg_Queue	-	-	**-0.046 (-0.09, -0.005)***	**0.955 (0.914, 0.995)***	-	-	**-0.052 (-0.103, -0.008)**	**0.949 (0.902, 0.992)**
B_TH_Avg_Wait	**0.013 (0.003, 0.022)**	**1.013 (1.003, 1.022)**	-	-	-	-	-	-
C_Vol_Th	**-0.006 (-0.012, 0.000)***	**0.994 (0.988, 1.000)***	-	-	-	-	-	-
C_LT_Avg_Queue	-	-	-	-	-	-	**-0.076 (-0.159, -0.003)**	**0.927 (0.853, 0.997)**
D_Vol_LT	-	-	**-0.036 (-0.067, -0.004)***	**0.965 (0.935, 0.996)***	-	-	**-0.039 (-0.078, -0.004)**	**0.962 (0.925, 0.996)**
D_OAFR	-	-	**0.518 (0.077, 0.978)**	**1.679 (1.08, 2.659)**	-	-	-	-
D_TH_GreenRatio	-	-	-	-	-	-	**-0.074 (-0.145, -0.004)**	**0.929 (0.865, 0.996)**
D_TH_Avg_Green	-	-	-	-	**-0.057 (-0.099, -0.019)**	**0.945 (0.906, 0.981)**	-	-
D_TH_Std_Green	-	-	-	-	-	-	**0.054 (0.006, 0.103)**	**1.055 (1.006, 1.108)**
D_TH_Avg_Wait	**0.009 (0.000, 0.017)**	**1.009 (1.000, 1.017)**	**-0.011 (-0.02, -0.002)**	**0.989 (0.98, 0.998)**	**0.011 (0.001, 0.021)**	**1.011 (1.001, 1.021)**	-	-
AUC	0.6759		0.6927		0.6337		0.6858	

*Note: Mean (95% BCI) values marked in bold are significant at the 0.05 level; Mean (95% BCI) values marked in bold and noted by * are significant at the 0.1 level.*

It is worth noting that the speed related variables were only found to be significant in slice 1 model, which might be explained as that the speed characteristics on the upstream segment only have short-term impacts on the within intersection crash occurrence, and relatively, these within intersection crashes are more likely to be influenced by the signal timing and traffic volume related variables. Based on the estimation results of slice 2 model, seven variables were found to be significantly associated with the crash risk within intersection area: (1) the positive coefficient (0.005) of "A_Vol_Th" indicates that higher through volume from "A" approach tends to increase the crash risk, which is consistent with previous aggregated intersection studies (Abdel-Aty and Wang, 2006; Chin and Quddus, 2003; Guo et al., 2010; Poch and Mannering, 1996) that higher exposure may results in more crashes. The odds ratio of 1.005 means that when other variables held constant, one-unit increase in the through volume from "A" approach would increase the odds of crash occurrence by 0.5%; (2) similarly, the left turn volume from "B" approach (B_Vol_LT) was also found to be positively correlated with the odds of crash occurrence. This could be explained in that higher left turn volume from "B" approach may results in more conflicts between the through vehicles from "A" approach and the left turn vehicle from "B" approach. The odds ratio of 1.04 means that when other variables held constant, one-unit increase in the left turn volume from "B" approach would increase the odds of crash occurrence by 4%; (3) "B_LT_Std_Green" was found to be negatively associated with the odds of crash occurrence within intersection, which means that higher standard deviation of the length of left turn phase on "B" approach could improve the safety performance of intersection. The possible reason is that when the left turn volume from "B" approach, as well as other variables held constant, the higher variation in the length of left turn phase on "B" approach indicates higher adaptability of the left turn phase, which indeed increase the safety performance of intersection; (4) "B_TH_Avg_Queue"

was found to have negative effect on the crash risk within intersection, which could be explained as that higher queue length on the through lanes of "B" approach may represent that more signal priority has been given to the "A" approach, which may reduce the exposed conflicting traffic flow between through vehicles from "A" and "B" approaches; (5) the negative coefficient (-0.036) of "D_Vol_LT" indicates that higher left turn volume from "D" approach tends to reduce the crash risk within intersection. The possible reason might be that more left turn vehicle from "D" approach may raise the awareness of those drivers from the "A" approach, which will therefore reduce the odds of crash occurrence. This is similar to the findings by Guo et al. (2010), which indicates that the left-turn ADT on minor road are significantly negatively associated with the crash frequency at signalized intersections; (6) higher "D_OAFR" tends to increase the odds of crash occurrence, which demonstrates that higher variation in traffic flow across through lanes on "D" approach tends to increase the crash risk within intersection. This could be potentially explained by that higher variation in traffic flow across through lanes on "D" approach may results in many lane change behavior occurring within the intersection, which will increase the complexity of traffic flow within intersection, as well as the odds of crash occurrence within intersection; (7) "D_TH_Avg_Wait" was found to be negatively correlated with the odds of crash occurrence within the intersection. This might be explained by that a longer waiting time on "D" approach indicates higher signal priority was given to the "A" approach, which will indeed reduce the exposed conflicting traffic flows between the through vehicles from "A" and "D" approaches.

4.4.2 Intersection entrance crashes

Similar to the within intersection crashes, a full model was first developed for the intersection entrance crashes based on four time slices. Table 4-6 shows the results of intersection entrance model based on full dataset. In total, 7 variables were identified to be significant variables, including speed characteristics, signal timing, queue length, and waiting time related factors collected from different time slices.

Table 4-6: Results of the Bayesian Conditional Logistic Model based on Full Dataset (Intersection Entrance).

Variables	Coefficient Estimation		Odds Ratio	
	Mean	95% BCI	Mean	95% BCI
A_TH_Avg_Queue_0_5	0.054	(0.018, 0.094)	1.055	(1.018, 1.099)
A_LT_Avg_Green_5_10	-0.056	(-0.107, -0.006)	0.946	(0.899, 0.994)
A_LT_Avg_Queue_5_10	-0.065	(-0.128, -0.007)*	0.937	(0.88, 0.993)*
A_TH_Avg_Wait_5_10	0.014	(0.000, 0.028)	1.014	(1.000, 1.028)
Avg_speed_10_15	-0.046	(-0.078, -0.017)	0.955	(0.925, 0.983)
A_TH_Avg_Green_15_20	-0.037	(-0.069, -0.009)	0.964	(0.933, 0.991)
A_LT_GreenRatio_15_20	-0.084	(-0.167, -0.003)	0.919	(0.846, 0.997)
AUC	0.728			

*Note: 95% BCI values marked in bold and noted by * indicate that these variables are significant at the 0.1 level, while other variables are significant at the 0.05 level.*

In addition to the full model, four separate time-slice models were developed for the intersection entrance crashes based on four time slices, respectively. Table 4-7 shows the results of 4 time-slice models for the intersection entrance dataset. The model comparison results based on AUC values indicate that the slice 2 model performs the best, followed by the slice 4 and slice 1 models, which

is in line with the within intersection models. The possible reason why the slice 4 model also performs very well might be that the traffic environment in the intersection entrance area is simpler than the within intersection area, therefore, the crash risk in the intersection entrance area tends to be more stable over time than the within intersection area. However, there may exist some uncertainty because of the insufficient sample size, which will afterwards influence the performance of different time-slice model. It is worth noting that the sign of the significant variables is consistent in all slices. Therefore, all the 7 significant variables among four time-slice models will be investigated for the intersection entrance dataset.

Table 4-7: Results of Bayesian Conditional Logistic Regression Models based on Different Time Slices (Intersection Entrance)

Variables	Slice 1		Slice 2		Slice 3		Slice 4	
	Mean (95% BCI)	Odds Ratio (95% BCI)	Mean (95% BCI)	Odds Ratio (95% BCI)	Mean (95% BCI)	Odds Ratio (95% BCI)	Mean (95% BCI)	Odds Ratio (95% BCI)
Avg_speed	**-0.050 (-0.077, -0.024)**	**0.951 (0.926, 0.976)**	**-0.041 (-0.072, -0.012)**	**0.96 (0.931, 0.988)**	**-0.038 (-0.066, -0.01)**	**0.963 (0.936, 0.99)**	**-0.037 (-0.068, -0.006)**	**0.964 (0.934, 0.994)**
A_Vol_LT	**-0.048 (-0.086, -0.013)**	**0.953 (0.918, 0.987)**	**-0.037 (-0.07, -0.005)***	**0.964 (0.932, 0.995)***	**-0.046 (-0.086, -0.01)**	**0.955 (0.918, 0.99)**	**-0.047 (-0.091, -0.009)**	**0.954 (0.913, 0.991)**
A_LT_Avg_Green	-	-	-	-	-	-	**-0.050 (-0.096, -0.003)**	**0.951 (0.908, 0.997)**
A_LT_Avg_Wait	-	-	**-0.013 (-0.022, -0.003)**	**0.987 (0.978, 0.997)**	-	-	-	-
A_TH_GreenRatio	-	-	**-0.040 (-0.081, -0.002)**	**0.961 (0.922, 0.998)**	-	-	-	-
A_TH_Std_Green	-	-	-	-	**-0.035 (-0.075, 0)**	**0.966 (0.928, 1)**	**-0.041 (-0.077, -0.007)**	**0.960 (0.926, 0.993)**
A_TH_Avg_Queue	**0.030 (0.001, 0.061)***	**1.030 (1.001, 1.063)***	-	-	-	-	-	-
AUC	0.6679		0.6770		0.6466		0.6767	

*Note: Mean (95% BCI) values marked in bold are significant at the 0.05 level; Mean (95% BCI) values marked in bold and noted by * are significant at the 0.1 level.*

In total, seven variables from the "A" approach were found to be significantly correlated with the crash occurrence in the intersection entrance area: (1) the coefficients of average speed are consistent to be negative among four time-slice models, which means that lower average speed tends to increase the odds of crash occurrence in the intersection entrance area, which is consistent with previous studies (Abdel-Aty et al., 2012; Ahmed et al., 2012a, b; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2016; Yuan et al., 2018a). This could be explained by that the lower average speed, i.e., congested condition, are more likely to have higher crash risk than uncongested condition; (2) the left turn volume was found to have significant negative effect on the odds of crash occurrence, which means that higher left turn volume may results in lower crash risk. The possible reason might be that driver intending to turn left approach the intersection more carefully and with lower speeds. Thus higher left turn volume may increase the driver awareness when approaching the entering approach, which may improve the safety performance; (3) the average length of left turn green phase was found to be negatively correlated with the odds of crash occurrence, which means that when the left turn volume, as well as other variables held constant, longer left turn green time could decrease the odds of crash occurrence; (4) the negative coefficient of the left turn average waiting time demonstrates that the longer waiting time for the left turn vehicles may results in better safety performance. The possible reason might be that the longer waiting time for the left turn vehicles, the less exposure may exist between left turn and through vehicles, which may reduce the crash risk; (5) similarly, the green ratio, as well as the standard deviation of the green time of the through phase were found to have negative effect on the odds of crash occurrence, which indicate that longer and more adaptive green phase for the through vehicles could significantly improve the safety performance of the intersection entrance area. It may be reasoned that longer and more adaptive green phase for the through

vehicles could significantly decrease the frequency of stop-and-go traffic, which will therefore decrease the potential conflicts. Similarly, Lee et al. (2013) found that the implementation of cooperative vehicle intersection control algorithm, which optimize the vehicle trajectory to reduce the stop-and-go frequency, can reduce the number of rear-end crash events by 30-87% for different volume condition; (6) the positive coefficient of average queue on the through lanes indicates that longer queue on the through lanes may increase the odds of crash occurrence.

4.5 Discussion and Conclusion

This research examined the real-time crash risk at signalized intersections based on the disaggregated data from multiple sources, including travel speed collected by Bluetooth detectors, lane-specific traffic volume and signal timing data from adaptive signal controllers, and weather data collected by airport weather station. The intersection and intersection-related crashes were collected and then divided into three types, i.e., within intersection crashes, intersection entrance crashes, and intersection exit crashes. In terms of the sample size, only the within intersection crashes and intersection entrance crashes were considered and then modeled separately. Matched case-control design with a control-to-case ratio of 4:1 was employed to select the corresponding non-crash events for each crash event, and three confounding factors, i.e., location, time of day, and day of the week, were selected as matching factors. Afterwards, all the traffic, signal timing, and weather characteristics during 20-minute window prior to the crash or non-crash events were collected and divided into four 5-minute slices, i.e., 0-5 minute, 5-10 minute, 10-15 minute, and 15-20 minute. Later, Bayesian conditional logistic models were developed for within intersection crashes and intersection entrance crashes, respectively.

For the within intersection crashes, the results of the full model (based on four time-slice datasets) indicate that 14 variables are significantly associated with the real-time crash risk, including speed characteristics, signal timing, queue length, and waiting time related factors collected from different approaches and time slices. The AUC value of the full model is 0.7596, which is much higher than the time-slice models. This comparison result reveals that incorporating all time slices variables could significantly improve the model performance. With respect to the four time-slice models, the model results show that the slice 2 model performs much better than the other modes in terms of the AUC value, which means that the characteristics during 5-10 minutes prior to the crash event have more power in the real-time crash risk prediction than the other time intervals. Among the slice 2 model, three volume related variables, i.e., the through volume from "A" approach (at-fault vehicle traveling approach), the left turn volume from "B" approach (near-side crossing approach), and the OAFR from "D" approach (far-side crossing approach), were found to have significant positive effects on the odds of crash occurrence, which is consistent with previous aggregated studies(Abdel-Aty and Wang, 2006; Chin and Quddus, 2003; Guo et al., 2010; Wang et al., 2016b; Xie et al., 2013). However, the left turn volume from "D" approach was found to have negative effect on the crash risk, this may be reasoned that more left turn vehicle from "D" approach may raise the awareness of those drivers from "A" approach, which will therefore reduce the crash risk.

Moreover, the standard deviation of the length of left turn green phase of "B" approach, the average queue length of the through vehicles on "B" approach, and the average waiting time of the through vehicles on "D" approach were found to be negatively associated with the odds of crash occurrence. These findings imply that the increased adaptability for the left turn signal timing of "B" approach (higher "B_LT_Std_Green") and increased priority for "A" approach (higher "B_TH_Avg_Queue"

and “D_TH_Avg_Wait”) could significantly decrease the odds of crash occurrence caused by the vehicles from “A” approach. It is worth noting that the speed-related variables were only found to be significant in the slice 1 model. This might be because the potential conflicting movements within intersection area are quite dynamic, and the speed characteristics on the upstream segment may only have short-term impacts on the within intersection crash occurrence.

With respect to the intersection entrance crashes, since all the involving vehicles in the intersection entrance crash are traveling on the same approach with the at-fault vehicle, only the characteristics of “A” approach were included in the models. The full model performs much better than the four time-slice models in terms of the AUC value, which is in line with the within intersection models. Among the four time-slice models, the slice 2 model performs the best, which is slightly better than the slice 4 and slice 1 models. The possible reason why the slice 4 model also performs very well might be that the traffic environment in the intersection entrance area is more simple than the within intersection area, therefore, the crash risk in the intersection entrance area tends to be more stable over time than the within intersection area, and the insufficient sample size may also results in some uncertainty among the four time-slice models. Therefore, the significant variables in four time-slice models were investigated. Average speed was found to have significant negative effect on the odds of crash occurrence, which is consistent with previous studies (Abdel-Aty et al., 2012; Ahmed et al., 2012a, b; Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2016; Yuan et al., 2018a). The left turn volume was surprisingly found to be negatively correlated with the odds of crash occurrence, which might be explained as the higher left turn volume may increase the driver awareness when approaching the entering approach, which may improve the safety performance. Moreover, three signal timing variables, i.e., A_LT_Avg_Green, A_TH_GreenRatio, and A_TH_Std_Green, were found to have significant negative effects on the

odds of crash occurrence. These findings imply that longer average green time for the left turn phase, higher green ratio for the through phase, and higher adaptability for the through green phase can significantly improve the safety performance in the intersection entrance area. Besides, the average queue length on the through lanes was found to have positive effect on the odds of crash occurrence, which indicates that longer queue on the through lanes may significantly increase the crash risk.

It is worth noting that all the weather-related variables are insignificant in both within intersection models and intersection entrance models. This might be explained by that the weather-related variables are more likely to have effects on high-speed segment or free-flow facilities, while the signalized intersections are usually operated at low speed and they are highly interrupted by the traffic signals, therefore, the weather-related variables may not have significant effects on the crash occurrence at signalized intersections. Above all, the model results provide a lot of insights on the relationship between the crash risk at signalized intersection and the real-time traffic and signal timing characteristics. For example, the results related to signal timing variables imply that higher adaptability for both left turn and through phases, longer average green time for the left turn phase, and higher green ratio for the through phase could significantly improve the safety performance of signalized intersections. These findings might be incorporated into the adaptive signal control algorithm to better accommodate the real-time safety and efficiency requirements (Gong et al., 2019a).

Overall, this study succeeds in verifying the feasibility of real-time safety analysis for signalized intersections. However, there are still some limitations for the current study. For example, only 23 signalized intersections on three corridors were considered, which may result in some bias in the data collection even though the matched case-control design was utilized. Also, different

geometric characteristics may also have significant effects on real-time crash risk, which has already been demonstrated by Ahmed et al. (2012a). However, the geometric effects were controlled in this study by using matched case-control design. Above all, further investigation would be beneficial to improve the generalization of the model results, which may start from the following aspects: increase the sample size by collecting data from large-scale signalized intersections which may also have various geometric characteristics and try to use unbalanced dataset which is more realistic than the artificially balanced data. It is also worth noting that the vulnerable users (pedestrians, motorcyclists) related crashes were not considered in the current stage, although signalized intersections are typical dangerous hotspots for the vulnerable road users. Therefore, it would be meaningful to investigate the relationship between vulnerable-user-related-crash occurrence and real-time traffic and signal characteristics.

CHAPTER 5: REAL-TIME CRASH RISK PREDICTION USING LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORK[3]

5.1 Introduction

Intersections are well-known high crash risk locations because of the variety of road user's behaviors and interaction. According to Fatality Analysis Reporting System (FARS) database, nearly 25% fatal crashes that occurred in United State in 2016 are intersection-related crashes. This serious traffic safety issue at intersections has been a critical research topic during past decades. However, previous safety studies for intersections mainly focused on static and aggregated analysis, which was limited by the data availability. These analyses were only able to identify some general influence factors, e.g., AADT, speed limit, geometric design, etc. At the same time, many researchers and organizations had calibrated and developed safety performance functions for different states and intersection types, which could be applied to predict annual crash frequency to better support safety evaluation and long-term management.

More recently, with the help of widely deployed traffic detectors along arterials and intersections, real-time traffic data are collected and updated in very short time period (e.g., 1 minute, 20 seconds, or even per individual vehicle). In this context, some researchers started to investigate the crash likelihood on urban arterials by using real-time traffic data (Theofilatos, 2017; Theofilatos et al., 2017; Yuan et al., 2018b). However, seldom research has been

[3] This chapter has been published in Transportation Research Record: Journal of the Transportation Research Board (https://doi.org/10.1177/0361198119840611)

conducted at signalized intersections (Yuan and Abdel-Aty, 2018). It is worth noting that the real-time crash risk prediction at signalized intersections is much more complicated than arterial segments, which could be explained by the conflicting movements, turning movements, and interrupted traffic flow temporally separated by signal control. These differences could result in that the prediction algorithm should consider huge number of influence factors, including the signal timing, volume, and speed characteristics for different movements.

These pioneering research studies mainly focused on the analyses between real-time crash risk and possible influence factors. As we are approaching connected and automated vehicles soon, which will enable more advanced and pro-active management strategies to be deployed at intersections to prevent crash occurrence in real time. Prior to the implementation of safety management strategies, more robust and reliable real-time crash risk prediction algorithms are needed to accurately predict the real-time crash risk at intersections.

Crash risk prediction is a typical binary classification problem, i.e., crash or non-crash. In the real world, non-crash events are much more common than crash events, and the crash events should be considered as very rare events. Therefore, this kind of imbalanced crash and non-crash event dataset can hardly be directly utilized to develop models. In general, there are two kinds of sampling methods could be applied to address this imbalanced issue: (1) under-sampling method aims to reduce the sample size of non-crash events to generate a relatively

balanced dataset; (2) while over-sampling method tends to increase the sample size of crash events by using various resampling methods to create a balanced dataset.

However, previous research mostly applied the under-sampling methods to balance the dataset, which may lose some important information for non-crash events. Among those research studies, matched case-control design was extensively deployed in previous studies: (1) within stratum-matched non-crash data (Abdel-Aty et al., 2004; Abdel-Aty et al., 2012). While over-sampling methods have seldom been utilized to predict real-time crash risk, Basso et al. (Basso et al., 2018) attempted to use synthetic minority over-sampling technique (SMOTE) on the training dataset to calibrate the prediction algorithms on one freeway segment, and then evaluated them based on a real-world imbalanced dataset. Their comparison results showed that the algorithms with SMOTE balanced dataset have better prediction performance than the other algorithms, which is consistent with the previous imbalanced classification studies in other fields (Chawla et al., 2002). It is worth noting that the previous real-time crash risk prediction models were evaluated based on artificially balanced test data, while these evaluation results can hardly represent the prediction performance in real-world application. To the best of the authors' knowledge, only Basso et al. (Basso et al., 2018) evaluated their models based on the original unbalanced dataset (where crashes are quite rare events). They claimed that their model could predict 67.89% of the crashes with a false positive rate of 20.94 for one freeway segment, which is among the best in the literature.

In terms of the methodology, there are generally two categories of modelling methods that are employed in real-time crash risk prediction studies: statistical analyses and machine learning approaches. Statistical methods include matched case-control logistic models (Abdel-Aty and Pande, 2005; Abdel-Aty et al., 2004; Ahmed and Abdel-Aty, 2012; Xu et al., 2012; Zheng et al., 2010), Bayesian logistical models (Ahmed et al., 2012a; Shi and Abdel-Aty, 2015; Wang et al., 2017a; Wang et al., 2015a; Yu et al., 2014), Bayesian random effect logistic models (Shi and Abdel-Aty, 2015; Yu et al., 2016), Bayesian random parameter logistic models (Shi and Abdel-Aty, 2015; Xu et al., 2014; Yu and Abdel-Aty, 2014; Yu et al., 2017). Machine-learning based methods include neural networks (Abdel-Aty and Pande, 2005; Abdel-Aty et al., 2008), support vector machines (Yu and Abdel-Aty, 2013; Yu and Abdel-Aty, 2014), and Bayesian networks (Hossain and Muromachi, 2012; Sun and Sun, 2015). With the rapid development of artificial intelligence and deep learning technologies, there are more and more advanced algorithms, for example, Recurrent Neural Networks (RNNs) have been proved to be very powerful in sequence learning (Chung et al., 2014), which might be more appropriate to predict the real-time crash risk by considering time series characteristics.

Recurrent Neural Networks (RNNs) are a class of artificial neural networks, which were developed in the 1980s. RNNs are distinguished from Feed-Forward Neural Networks by incorporating feedback to previous layers. Because of their internal memory, RNNs can remember important information about the input they received. Therefore, RNNs have been

widely employed in many fields to conduct sequential data analysis and prediction, including language model (Mikolov et al., 2010), speech recognition (Graves et al., 2013), machine translation (Kalchbrenner and Blunsom, 2013), etc. However, the traditional short-term RNNs usually take too much time or do not work well at all, especially when the time lag is long, which may result in exploding or vanishing gradients. Therefore, Hochreiter and Schmidhuber (1997) proposed a novel recurrent network architecture in conjunction with an appropriate gradient-based learning algorithm, which is long short-term memory (LSTM) RNNs. LSTMs have gating mechanism to store the relevant information for future predictions, which are explicitly designed to avoid long-term dependency problem, therefore they were proved to have very good performance on a large variety of sequence learning problems, for example, handwriting sequence generation, sequential trajectory learning, language modeling, speech recognition, visual recognition, etc.

With respect to transportation field, many studies have been conducted by using RNN or LSTM, which mainly focus on driving behavior identification (Wijnands et al., 2018), travel demand prediction (Xu et al., 2017), and roadway traffic speed or travel time prediction (Ma et al., 2015). Ma et al. (2015) applied LSTM to predict travel speed based on the data collected by traffic microwave detectors in Beijing, they found that LSTM achieved the best prediction performance in terms of both accuracy and stability among several prevailing parametric and nonparametric algorithms. Xu et al. (2017) proposed an LSTM-based sequence learning model

to predict future taxi requests in each area of New York City based on the recent demand and other relevant information, they also found that the LSTM algorithm outperforms other prediction methods, such as feed-forward neural network. Wijnands et al. (2018) tried to identify changes in individual driving behavior by using LSTM based on the individual's acceleration and deceleration pattern.

To the best of the authors' knowledge, no research has been conducted for real-time crash risk prediction by using LSTM. However, real-time crash risk prediction is a typical time series related sequential prediction process, and the impacts of long-term and short-term traffic data might be quite different, which could be captured by LSTM efficiently. Therefore, LSTM would be a better solution for real-time crash risk prediction. In summary, this study aims to bridge the following two research gaps for real-time crash risk prediction: (1) first, develop a real-time crash risk prediction algorithm for signalized intersections by using LSTM RNN; (2) second, collect real-world full sample data from 44 intersections, and then compare the prediction performance between proposed LSTM RNN algorithm based on SMOTE over-sampled dataset and conditional logistic models based on within-stratum matched case-control dataset. The whole framework of this chapter is shown in Figure 5-1.

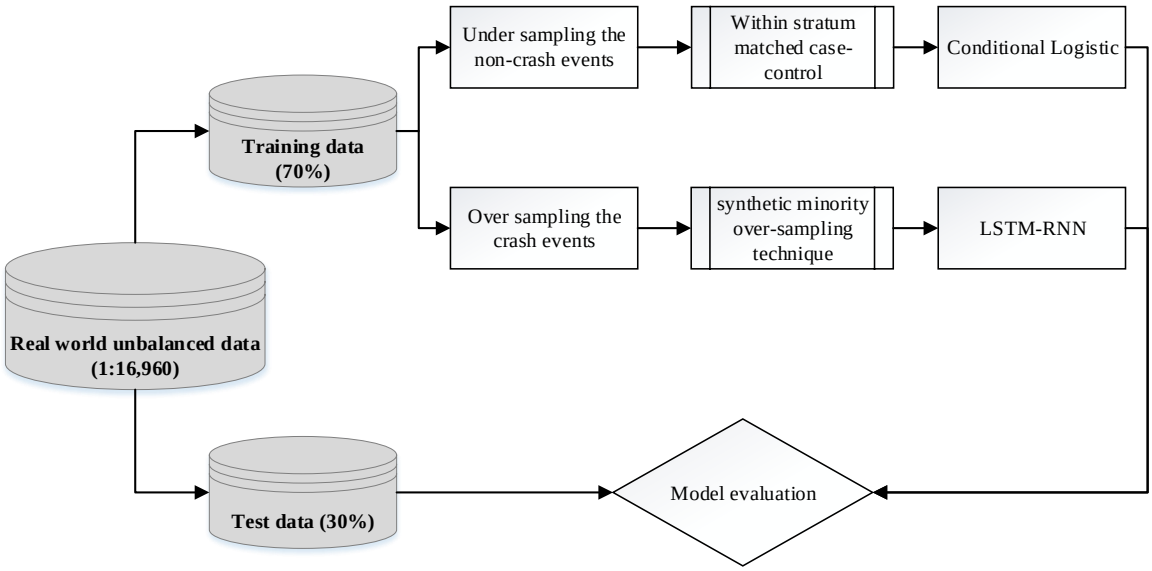


Figure 5-1: Framework of the Study

5.2 Data Preparation

In total, 44 signalized intersections were chosen from Oviedo, Florida, as shown in Figure 5-2. A total of five datasets were utilized: (1) crash data from January 2017 to April 2018 provided by Signal Four Analytics (S4A); (2) travel speed data collected by 44 BlueTOAD detectors installed at 44 intersections; (3) signal timing data provided by Automated Traffic Signal Performance Measures (ATSPM) database; (4) loop detector data were also provided by ATSPM database; (5) weather characteristics collected by the nearest airport weather station.

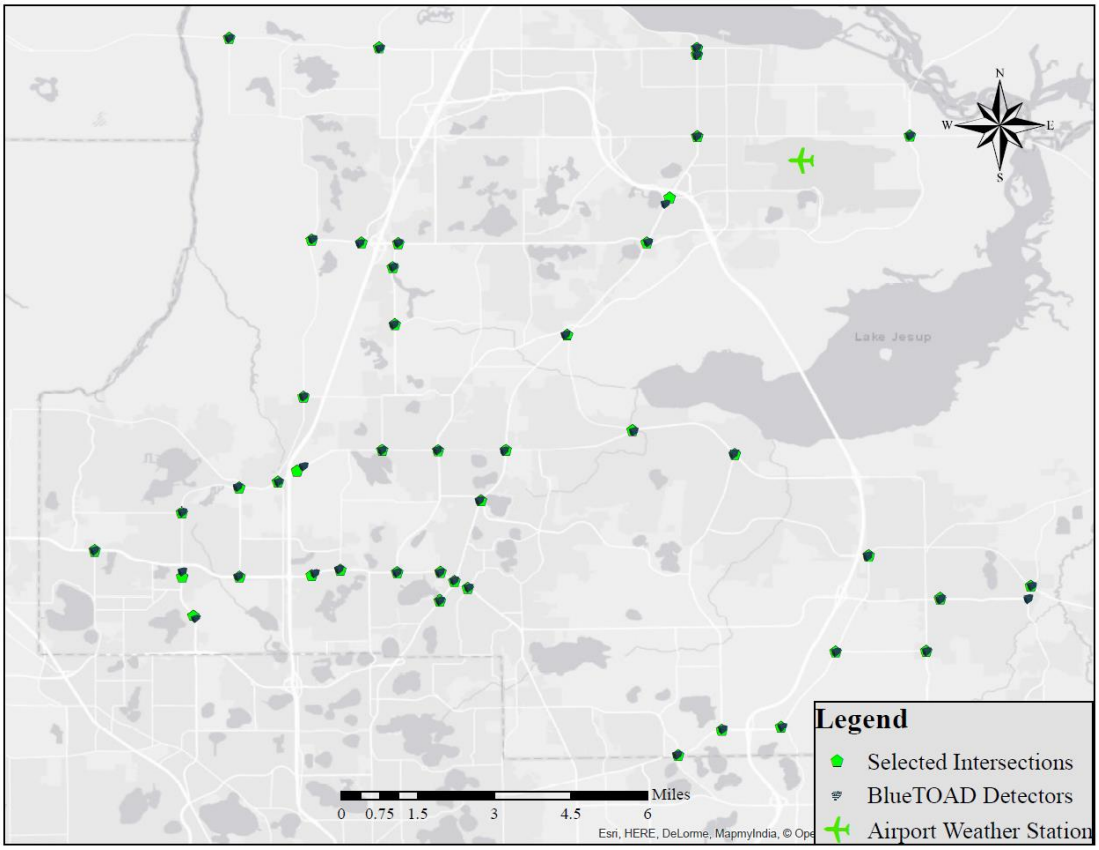


Figure 5-2: Selected Intersections

Signal four analytics (S4A) system provides detailed crash information, including crash time, location, severity, type, weather condition, etc. First, all crashes occurred at intersections or influenced by intersections (within 250 feet of intersections) from January 2017 to April 2018 were collected. Second, 16 (2.35%) crashes under the influence of alcohol and drugs were excluded, since these kinds of crashes are usually not attributed to real-time traffic and signal characteristics which are the focus of this study. Since the percentage of alcohol and drug related crashes is very low, therefore, it is assumed that this process would not result in bias estimation. After data preprocessing, 665 crashes were collected and divided into three types based on their location, namely, within intersection areas, intersection entrance areas, and

intersection exit areas, which were defined in our previous research (Yuan and Abdel-Aty, 2018). 335 (50.37%) crashes occurred within intersection areas, 230 (34.59%) crashes occurred in intersection entrance areas, and 90 (13.53%) crashes occurred in intersection exit areas. As shown in the previous study (Yuan and Abdel-Aty, 2018), crash occurrences within intersection areas are more likely to be predicted, therefore, only within-intersection crashes were selected as an example to prove the concept of this study.

In this study, all the within intersection crashes were associated to the travelling approach of the at-fault vehicle, which is consistent with Yuan and Abdel-Aty (2018). It is worth noting that all the intersection approaches were renamed as "A", "B", "C", and "D" for each crash and non-crash event based on the relative-direction nomenclature proposed by Yuan and Abdel-Aty (Yuan and Abdel-Aty, 2018), which aims to keep all the characteristics from different approaches comparable. For example, the impacts of southbound real-time traffic volume on southbound crash occurrences should be different from that of eastbound crash occurrences due to the different conflict patterns. Therefore, all the data collection from different approaches were based on relative direction. More specifically, the "A" approach indicates the approach where the crash or non-crash event occurred at, and the "B" approach indicates the left-side approach of the "A" approach. Similarly, the "C" and "D" approaches follow a clockwise sequence (please refer to Yuan and Abdel-Aty (Yuan and Abdel-Aty, 2018) for details).

Speed data are provided by 44 BlueTOAD® detectors which measure the individual vehicular speed on a specific segment, as shown in Figure 5-3. Bluetooth detectors can only detect the vehicles equipped with Bluetooth devices which are working on discoverable mode. The individual vehicular speed on a specific segment is calculated as the segment length divided by the travel time of each detected vehicle on the segment based on the detection data of two Bluetooth detectors located at two intersections. In this study, the Bluetooth penetration rate is 3.69%, which is higher than the threshold suggested by the previous studies (Chen and Chien, 2000; Long Cheu et al., 2002a). Also, the validity of Bluetooth detectors for measuring individual vehicular speed on urban arterials has been proved by our previous research (Yuan and Abdel-Aty, 2018; Yuan et al., 2018b). In this study, speed data (including average speed and speed standard deviation) were only collected for the segment of "A" approach, which represents the approach where crashes occurred, or at-fault vehicles traveled.

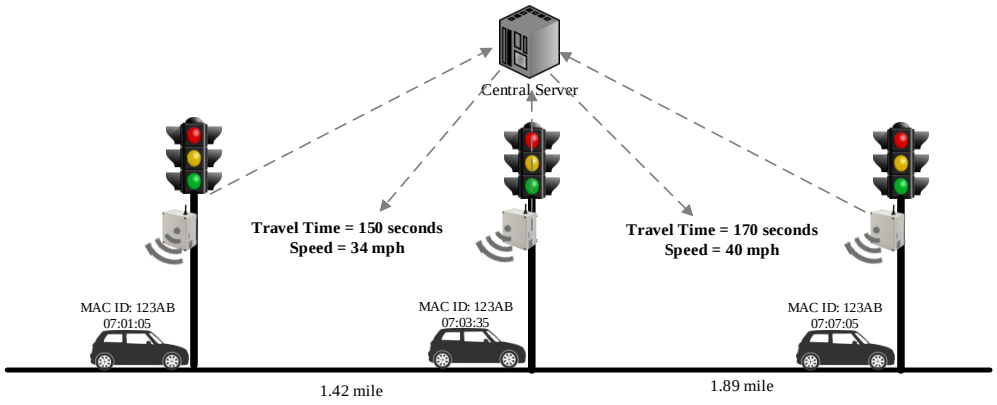


Figure 5-3: Illustration of Bluetooth Data Collection (Yuan et al., 2018b)

Signal timing and lane-specific vehicle count data are archived by the Automated Traffic Signal Performance Measure (ATSPM) database, which is recorded in the highest time resolution of controllers (0.1 seconds). All events generated by signal controllers are recorded in sets of four bytes per event: one byte for event code type, one byte for event parameter (for signifying detector numbers and phases), and two bytes for timestamp of when the event occurred. The event code is important for determining the type of reported activity, which could be phase initiation or termination, detection on/off, etc. In this study, there are three lane-specific volume-related variables and three signal timing related variables collected for every phase and then aggregated in 5 minutes with 1-minute updating increments. All the required information for the six measures is shown in Table 5-1.

Table 5-1: Required Data Elements for Selected ATSPM Measures.

Variable	Description	Required Event Code
Total Volume	Number of vehicles detected on a specific lane during given time period.	82. Detector On
Arrive on Green (AOG)	Number of vehicles detected on a specific lane while the intersection is green.	1. Phase Green 82. Detector On
Arrive on Yellow (AOY)	Number of vehicles detected on a specific lane while the intersection is yellow.	8. Phase Yellow 82. Detector On
Green Ratio	Ratio between phase green duration and given time period.	1. Phase Green
Average Green Time	The average value of all the green phase duration during given time period.	
Standard Deviation of Green Time	The standard deviation of all the green phase duration during given time period.	

Three weather related variables (weather type, visibility, and hourly precipitation) were collected from the nearest airport weather station (as shown in Figure 5-2). Since weather data are not recorded continuously, once weather condition changes and reaches a preset threshold, a new record will be added to the archived data. For every crash and non-crash event, the closest weather record prior to the crash time was extracted. It is worth noting that all the study locations are within 20 miles of the selected airport weather location, which is valid according to the previous research (Chung et al., 2018).

Since this study aims to provide 5-minute crash risk prediction (i.e., crash risk during next 5-10 minutes), while updating every minute. Therefore, it will generate an observation every minute for every intersection approach. Originally, there are 39,441,600 (83 approaches*11 months *30 days *24 hours*60 minutes) observations. With respect to crash events, all the observations whose prediction time periods include historical crash occurrences would be labeled as crash events. Meanwhile, all the observations within 3 hours after crash occurrences were excluded to eliminate the influence of crash events. For every observation, all the real-time traffic, signal, and weather characteristics were extracted for the period from 0 to 30 minutes (divided into six 5-minute time slices) prior to the observation time. For example, if an observation i at 18:26, the corresponding traffic and signal timing data from 17:56 to 18:26 were extracted and named as time slice 6, 5, 4, 3, 2, and 1, respectively.

After data matching and cleaning, there are 8,463,751 observations (499 crash events and 8,463,252 non-crash events) for within intersection area. The crash to non-crash ratio is 1:16,960, where the crash events are rare events. The full dataset was divided into training dataset (70%) and test dataset (30%). It is worth noting that two sampling methods were applied on the training dataset, while the test dataset was still the original imbalanced data with 150 crash events and 2,539,130 non-crash events. Given the real-world full sample data, two kinds of sampling methods (within stratum matched case-control and SMOTE) were employed on the training dataset and then evaluated based on the original test dataset. Therefore, a comprehensive comparison on the prediction performance could be conducted between traditional real-time crash risk prediction based on under-sampling methods and the proposed real-time crash risk prediction based on over-sampling method.

In terms of the within stratum matched case-control sampling method, four confounding factors, i.e., intersection ID, approach ID, time of day, and day of week, were controlled as matching factors. Therefore, all the corresponding non-crash events could be identified by using these matching factors and then a specific number of non-crash events would be randomly selected from the group of non-crash events for every crash event. Abdel-Aty et al. (Abdel-Aty et al., 2004) found that there is no significant difference when the control-to-case ratio changing from one to five. Among the previous research, 4:1 is the most commonly used control-to-case ratio (Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Xu et al., 2012; Yu et al., 2018; Yu

et al., 2016; Zheng et al., 2010). Other than that, 10:1 (Wang et al., 2015a) and 20:1 (Xu et al., 2013a) had also been applied by few researchers. Since the full sample data have already been collected in this study, it is more appropriate to use high control-to-case ratio, which is closer to the real condition and they can capture more information. However, highly imbalanced dataset may decrease the performance of traditional logistic model. Therefore, 10:1 was chosen in this study to compare with the other sampling methods. Consequently, 10 non-crash events from the same intersection, approach, time of day, and day of week were randomly selected for each crash event. For some crash events, there is no any matched non-crash events and some crash events may have less than 10 non-crash events dues to data missing issue. Finally, 3215 non-crash events and 349 crash events were collected as the matched case-control dataset.

With respect to the over-sampling method, the SMOTE was employed to create synthetic examples of the minority class (i.e., crash events) to achieve an equal number of samples with the majority class. These synthetic examples were randomly introduced among the minority class and along the line segments joining any of the k minority class nearest neighbors. In this study, k was set to be 5, which is consistent with Chawla et al. (2002).

In summary, three datasets were generated, i.e., two training datasets and one test dataset. Since all the variables are too many to be shown, therefore, Table 5-2 only shows the summary statistics in the full sample dataset for the characteristics during time slice 1 on the “A” approach.

Table 5-2: Summary of Variables Descriptive Statistics (Crash and Non-Crash Cases)

Type	Variable	Description	Crash Event		Non-Crash Event	
			Mean (Std)	(Min, Max)	Mean (Std)	(Min, Max)
Traffic data	Avg_speed	Average speed on the upstream segment	34.44 (10.28)	(4.00, 76.9)	32.13 (10.28)	(5.83, 60.00)
	Std_speed	Speed standard deviation on the upstream segment	5.22 (5.22)	(0.00, 59.75)	5.45 (4.67)	(0.00, 29.66)
	TH_Volume	Through volume of "A" approach	86.71 (51.86)	(0.00, 754.00)	104.46 (53.73)	(0.00, 315.00)
	LT_Volume	Left turn volume of "A" approach	7.85 (8.95)	(0.00, 101.00)	9.77 (9.76)	(0.00, 70.00)
	TH_AOG	Number of through vehicles arrived on green	64.38 (46.3)	(0.00, 435.00)	78.73 (49.91)	(0.00, 297.00)
	LT_AOG	Number of left turn vehicles arrived at intersection on green	4.13 (6.71)	(0.00, 89.00)	5.49 (7.91)	(0.00, 65.00)
	TH_AOY	Number of through vehicles arrived on yellow	2.44 (2.81)	(0.00, 61.00)	3.00 (3.44)	(0.00, 30.00)
	LT_AOY	Number of left turn vehicles arrived on yellow	0.65 (1.29)	(0.00, 27.00)	1.03 (1.72)	(0.00, 11.00)
	TH_Volume_OAFR	Overall average flow ratio among all the through lanes	1.13 (0.57)	(0.94, 51.00)	1.21 (1.46)	(0.94, 17.02)
	TH_AOG_OAFR	Overall average flow ratio among all the through lanes	1.10 (0.42)	(0.94, 41.01)	1.15 (1.11)	(0.94, 13.22)
	TH_AOY_OAFR	Overall average flow ratio among all the through lanes	1.15 (0.25)	(0.94, 6.15)	1.16 (0.27)	(0.94, 2.60)
Signal Timing	TH_Green_Ratio	Ratio of through green time within 5-minute interval	0.47 (0.19)	(0.00, 1.00)	0.46 (0.17)	(0.04, 0.91)
	LT_Green_Ratio	Ratio of left turn green time within 5-minute interval	0.08 (0.07)	(0.00, 1.00)	0.09 (0.07)	(0.00, 0.52)
	TH_Avg_green	Average length of through green phase within 5-minute interval	299.74 186.29)	(0.00, 2754.00)	79.27 (40.58)	(0.00, 360.00)
	LT_Avg_green	Average length of left turn green phase within 5-minute interval	159.49 346.39)	(0.00, 624.00)	14.78 (9.75)	(0.00, 57.00)
	TH_Std_green	Standard deviation of the length of through green phase within 5-minute interval	136.02 (12.15)	(0.00, 194.62)	6.16 (11.77)	(0.00, 139.30)
	LT_Std_green	Standard deviation of the length of left turn green phase within 5-minute interval	148.63 (24.09)	(0.00, 441.99)	1.43 (2.60)	(0.00, 21.92)
Weather data	Weather_type	Weather type: 0 for normal and 1 for adverse weather.	0.06 (0.25)	(0.00, 1.00)	0.06 (0.23)	(0.00, 1.00)
	Visibility	Visibility (mile).	9.80 (1.16)	(0.00, 10.00)	9.83 (0.96)	(3.00, 10.00)
	Precipitation	Hourly precipitation (inch).	0.00 (0.03)	(0.00, 1.48)	0.00 (0.03)	(0.00, 0.32)
	Humidity	Percentage (%)	59.63 (19.77)	(14.00, 100.00)	58.55 (20.07)	(14.80, 100)
Geometry	App_throu	Number of through lanes	2.48 (0.69)	(1.00, 4.00)	2.62 (0.66)	(1.00, 4.00)
	App_right	Number of exclusive right-turn lanes	0.57 (0.53)	(0.00, 2.00)	0.56 (0.51)	(0.00, 2.00)
	App_left	Number of left lanes	1.36 (0.52)	(0.00, 2.00)	1.43 (0.52)	(0.00, 2.00)
	App_lim	Speed limit	42.26 (5.01)	(30.00, 55.00)	42.03 (4.16)	(30.00, 55.00)
	App_mn_mj	Minor or Major (0 or 1)	0.90 (0.30)	(0.00, 1.00)	0.91 (0.28)	(0.00, 1.00)
	Leg_3_or_4	3-legged vs 4-legged (0 or 1)	0.94 (0.23)	(0.00, 1.00)	0.97 (0.18)	(0.00, 1.00)

Type	Variable	Description	Crash Event		Non-Crash Event	
			Mean (Std)	**(Min, Max)**	**Mean (Std)**	**(Min, Max)**
	Left_protected	Left turn protected	0.51 (0.5)	(0.00, 1.00)	0.51 (0.50)	(0.00, 1.00)
	Inter_size	Intersection size (Total lane number)	15.56 (4.31)	(7.00, 24.00)	16.64 (4.55)	(7.00, 24.00)
	N_left_maj	Number of left-turn lanes on major road	2.70 (0.9)	(1.00, 4.00)	2.87 (0.93)	(1.00, 4.00)
	N_left_min	Number of left-turn lanes on minor road	2.57 (1.22)	(0.00, 4.00)	2.77 (1.20)	(0.00, 4.00)
	N_righ_maj	Number of exclusive right-turn lanes on major road	0.98 (0.74)	(0.00, 2.00)	0.98 (0.79)	(0.00, 2.00)
	N_righ_min	Number of exclusive right-turn lanes on minor road	1.37 (0.62)	(0.00, 2.00)	1.49 (0.59)	(0.00, 2.00)
	Sp_lim_maj	Speed limit on major road	42.69 (4.83)	(30.00, 55.00)	42.47 (3.95)	(35.00, 55.00)
	Sp_lim_min	Speed limit on minor road	35.22 (5.96)	(25.00, 45.00)	35.99 (6.37)	(25.00, 45.00)

5.3 Methodology

To calibrate real-time crash prediction models based on the two kinds of training datasets, two different methodologies were employed, respectively. More specifically, conditional logistic model was developed based on the match case-control dataset and LSTM was calibrated based on the SMOTE oversampled dataset. At the end, these two kinds of real-time crash prediction algorithms were compared based on the same unbalanced test dataset. Detailed explanation of these two models are as shown in the following sections.

5.3.1 Long Short-Term Memory Recurrent Neural Network

The LSTM addresses the long-term dependency problem by introducing a memory cell which is able to preserve state over long periods of time (Hochreiter and Schmidhuber, 1997). A multilayer LSTM was developed to predict the crash risk during next 5-10 minutes based on sequence inputs. As shown in Figure 5-4, six input vectors for six time slices are mapped to a probability vector at the output layer for identification. The hidden state of the LSTM unit in the first LSTM layer is

used as input to the LSTM unit in the second LSTM layer in the same time step (Graves et al., 2013).

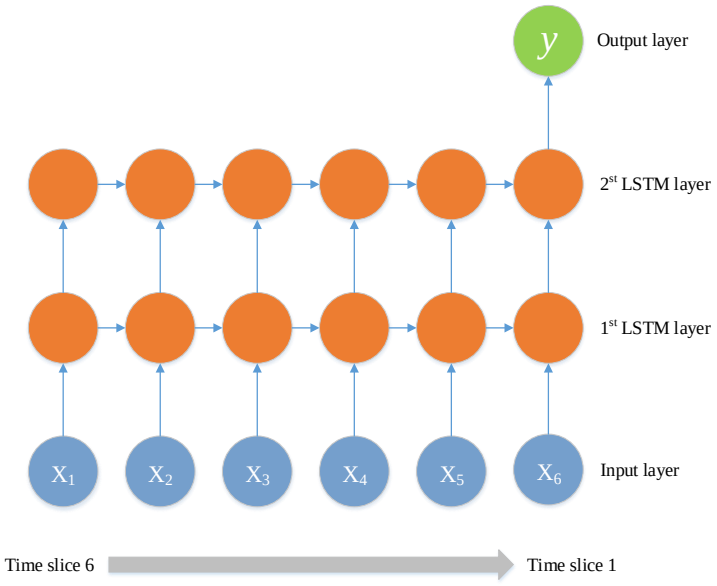


Figure 5-4: Illustration of the LSTM Architecture

A standard LSTM unit contains an input gate i_t, a forget gate f_t, an output gate O_t, a memory cell C_t, and a hidden state h_t. The values of gating vectors i_t, f_t, and o_t are in [0, 1]. The LSTM unit at each time step is illustrated in Figure 5-5.

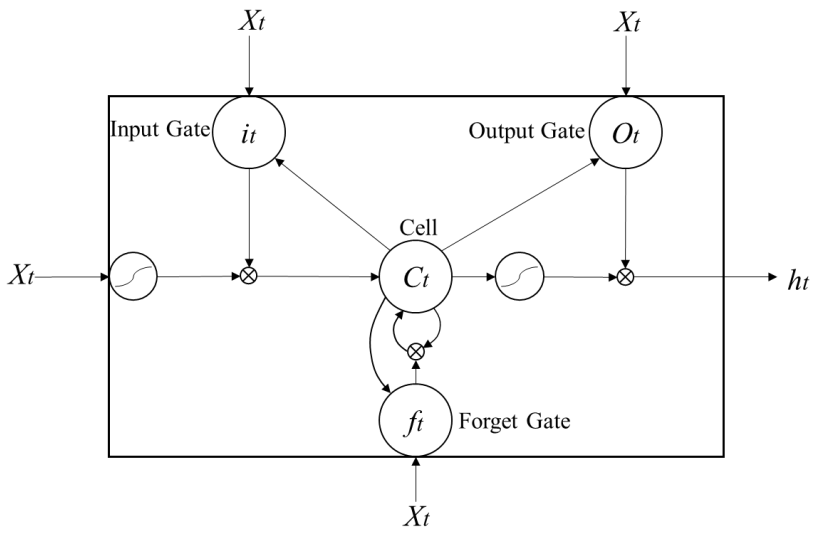


Figure 5-5: Illustration of LSTM Unit (Graves et al., 2013)

The LSTM generates a mapping from an input sequence vectors $X = (X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6)$ to an output probability vector by calculating the network unit activations using the following equations, iterated from $t = 1$ to 6:

$$\boldsymbol{i_t = \sigma(W_{ix}X_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i)} \quad (5\text{-}1)$$

$$f_t = \sigma(W_{fx}X_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (5\text{-}2)$$

$$o_t = \sigma(W_{ox}X_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o) \quad (5\text{-}3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{cx}X_t + W_{ch}h_{t-1} + b_c) \quad (5\text{-}4)$$

$$h_t = o_t \odot tanh(c_t) \quad (5\text{-}5)$$

$$y_t = W_{yh}h_{t-1} + b_y \quad (5\text{-}6)$$

where W represent weight matrices, for example, W_{ix} denotes the weight matrix from the input gate to the input, σ is the logistic sigmoid function, and ⊙ indicates elementwise product of the vectors. The forget gate f_t controls the extent to which the previous step memory cell is forgotten, the input gate i_t determines how much to update for each unit, and the output gate o_t controls the

exposure of the internal memory state. Since the value of all the gating variables vary for each time step, therefore, the model could learn how to represent information over multiple time steps.

5.3.2 Conditional Logistic Model

Suppose that there are *N* strata, where one crash (y_{ij}=1) and *m* non-crash cases (y_{ij}=0) in stratum *i*, *i*=1, 2, …, *N*. Let p_{ij} be the probability that the *j*th observation in the *i*th stratum is a crash; *j*=0, 1, 2, …, *m*. This crash probability could be expressed as:

$$\boldsymbol{y_{ij} \sim Bernoulli(p_{ij})} \quad (5\text{-}7)$$

$$logit(p_{ij}) = \alpha_i + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij} \quad (5\text{-}8)$$

Where α_i is the intercept term for the *i*th stratum; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ is the vector of regression coefficients for *k* independent variables; $\boldsymbol{X_{ij}} = (X_{1ij}, X_{2ij}, \ldots, X_{kij})$ is the vector of *k* independent variables.

In order to consider stratification in the analysis of the observed data, the stratum-specific intercept α_i is considered to be nuisance parameter, and the conditional likelihood for the *i*th stratum would be expressed as (Hosmer Jr et al., 2013):

$$\boldsymbol{l_i(\beta) = \frac{\exp(\sum_{u=1}^{k} \beta_u X_{ui0})}{\sum_{j=0}^{m} \exp(\sum_{u=1}^{k} \beta_u X_{uij})}} \quad (5\text{-}9)$$

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

$$\boldsymbol{L(\beta) = \prod_{i=1}^{N} l_i(\beta)} \qquad (5\text{-}10)$$

Since the full conditional likelihood is independent of stratum-specific intercept α_i, thus Equation 8 cannot be used to estimate the crash probabilities. However, the $\boldsymbol{\beta}$ coefficients can be estimated by Eq. (5-10). These estimates are the log-odds ratios of corresponding variables and can be used to approximate the relative risk of a crash. Furthermore, the log-odds ratios can also be used to develop a prediction model under this matched case-control analysis. Suppose two observation vectors $\boldsymbol{X_{i1}} = (X_{1i1}, X_{2i1}, \dots, X_{ki1})$ and $\boldsymbol{X_{i2}} = (X_{1i2}, X_{2i2}, \dots, X_{ki2})$ from the *i*th strata, the odds ratio of crash occurrence caused by observation vector $\boldsymbol{X_{i1}}$ relative to observation vector $\boldsymbol{X_{i2}}$ could be calculated as:

$$\boldsymbol{\frac{p_{i1}/(1-p_{i1})}{p_{i2}/(1-p_{i2})} = \exp[\sum_{u=1}^{k} \beta_u (X_{ui1} - X_{ui2})]} \qquad (5\text{-}11)$$

The right side of Equation 11 is independent of α_i and can be calculated using the estimated $\boldsymbol{\beta}$ coefficients. Thus, the above relative odds ratio may be utilized for predicting crash occurrences by replacing $\boldsymbol{X_{i2}}$ with the vector of the independent variables in the *i*th stratum of non-crash cases. One may use simple average of all non-crash observations within the stratum for each variable. Let $\boldsymbol{\bar{X}_i} = (\bar{X}_{1i}, \bar{X}_{2i}, \dots, \bar{X}_{ki})$ denotes the vector of mean values of non-crash cases of the *k* variables within the *i*th stratum. Then the odds ratio of a crash relative to the non-crash cases in the *i*th stratum could be approximated by:

$$\boldsymbol{\frac{p_{i1}/(1-p_{i1})}{p_{\bar{i}}/(1-p_{\bar{i}})} = \exp[\sum_{u=1}^{k} \beta_u (X_{ui1} - \bar{X}_{ui})]} \qquad (5\text{-}12)$$

5.3.3 Performance Metrics

In terms of model performance, AUC, which is the area under Receiver Operating Characteristic (ROC) curve was adopted. The ROC curve illustrates the relationship between true positive rate (sensitivity) and false alarm rate (1–specificity) for a given threshold from 0 to 1. It is worth noting that the classification results of binary logistic model are based on the predicted crash probabilities, which lie in the range of 0 to 1, while the classification result of conditional logistic model are based on the predicted odds ratio over the average condition of the matched non-crash events at the same location, which may be larger than 1. To be consistent with the other model, all the odds ratios predicted by conditional logistic model were scaled by using min-max normalization. Later, the normalized odds ratios were used to generate the classification result based on different threshold from 0 to 1.

To calculate specific values for sensitivity and false alarm rate, the threshold needs to be determined. In this study, the threshold value was chosen as the point where sensitivity equals to specificity. Based on this determined threshold, both sensitivity and false alarm rate were calculated for every model.

5.4 Result Analysis and Comparison

For every time-slice dataset, there are 84 variables from four intersection approaches. However, these variables from different intersection approaches and time slices might be highly correlated. Therefore, both Pearson linear correlation analysis and maximal information coefficient (MIC) nonlinear correlation analysis were conducted to identify the highly correlated variables. In terms of the threshold, 0.6 was utilized for the linear Pearson correlation analysis, while 0.65 was applied for the MIC value, which is suggested by Albanese et al. (2018). Finally, the highly correlated

pairs of variables were selected based on two criteria: the Pearson correlation coefficient is greater than 0.6 or the MIC is greater than 0.65.

Based on the correlation analysis results, variable selection procedure were conducted by incorporating the results of highly correlated variables and the variable importance (decrease in Gini impurity index) which was generated by using random forest (RF) algorithm (Ahmed and Abdel-Aty, 2012). For example, if two variables were identified to be highly correlated, then the less important variable would be excluded from the next step. Based on the selected variables, conditional logistic model was developed based on the filtered variables. Table 5-3 shows the final model results, since the major objective of this study is prediction rather than association analysis, the significance threshold of p-value was relaxed to 0.2 which indicates that all the variables with p-value smaller than 0.2 were included in the final model to ensure all the possible contributing factors were included.

Table 5-3: Model Results of Conditional Logistic Regression

Variable	Coefficient	S.E.	P-value	Odds Ratio (S.D.)
Avg_speed_slice2	-0.013	0.010	0.198	0.988 (1.01)
A_TH_AOY_slice2	0.052	0.025	0.037*	1.053 (1.025)
A_TH_AOY_slice4	0.087	0.027	0.001*	1.091 (1.027)
A_TH_AOG_OAFR_slice6	0.280	0.087	0.001*	1.324 (1.091)
A_LT_AOY_slice1	0.180	0.053	0.001*	1.197 (1.054)
A_LT_AOY_slice2	-0.317	0.067	0.000*	0.728 (1.069)
A_LT_AOY_slice3	-0.307	0.070	0.000*	0.736 (1.072)
A_TH_Avg_green_slice2	0.004	0.002	0.045*	1.004 (1.002)
A_LT_Std_green_slice3	0.040	0.021	0.064**	1.04 (1.022)
B_TH_AOY_slice4	-0.082	0.052	0.116	0.921 (1.053)
B_TH_AOY_slice6	0.074	0.044	0.091**	1.077 (1.045)
B_TH_AOG_OAFR_slice1	0.165	0.102	0.108	1.179 (1.108)
B_TH_AOG_OAFR_slice3	-0.368	0.154	0.017*	0.692 (1.166)
B_TH_AOG_OAFR_slice5	0.170	0.110	0.121	1.185 (1.116)
B_TH_Green_Ratio_slice3	-3.097	1.331	0.020*	0.045 (3.787)
B_LT_Green_Ratio_slice1	4.735	1.387	0.001*	113.833 (4.001)
B_LT_Std_green_slice4	-0.057	0.029	0.047*	0.945 (1.029)
D_TH_AOY_slice3	-0.121	0.050	0.015*	0.886 (1.051)
D_TH_AOY_slice4	-0.079	0.049	0.108	0.924 (1.05)
D_TH_AOY_slice6	-0.086	0.047	0.067**	0.917 (1.048)
D_TH_AOG_OAFR_slice2	-0.334	0.209	0.110	0.716 (1.233)
D_TH_AOG_OAFR_slice4	0.394	0.109	0.000*	1.483 (1.116)
D_TH_AOG_OAFR_slice5	0.505	0.131	0.000*	1.657 (1.14)
D_TH_AOG_OAFR_slice6	0.195	0.136	0.152	1.215 (1.146)

*Note: the p value noted by * indicate that these variables are significant at the 0.05 level, while the value noted by ** indicate that these variables are significant at the 0.1 level.*

It is worth noting that several AOY and AOG related variables were found to be significantly associated with real-time crash risk, especially for the “A” approach. These findings indicate that more through vehicles arrive on yellow may significantly increase the crash risk, which could be explained by the impacts of intersection dilemma zone. The above model estimation results were then applied on the unbalanced test dataset (150 crash events and 2,539,130 non-crash events).

Figure 5-6 shows the ROC curve of the model prediction performance, the area under ROC curve is 0.61, which is lower than the previous research (Yuan and Abdel-Aty, 2018). This could be mainly attributed to the test dataset, where the real-world unbalanced dataset might be much difficult to achieve high sensitivity and keep low false alarm rate. However, the evaluation based on unbalanced dataset is very meaningful, which could represent the prediction performance in real world. The threshold for prediction were determined based on the condition where sensitivity (i.e., true positive rate) equals to specificity (i.e., true negative rate), which is consistent with previous research (Ahmed and Abdel-Aty, 2013; Xu et al., 2013b).

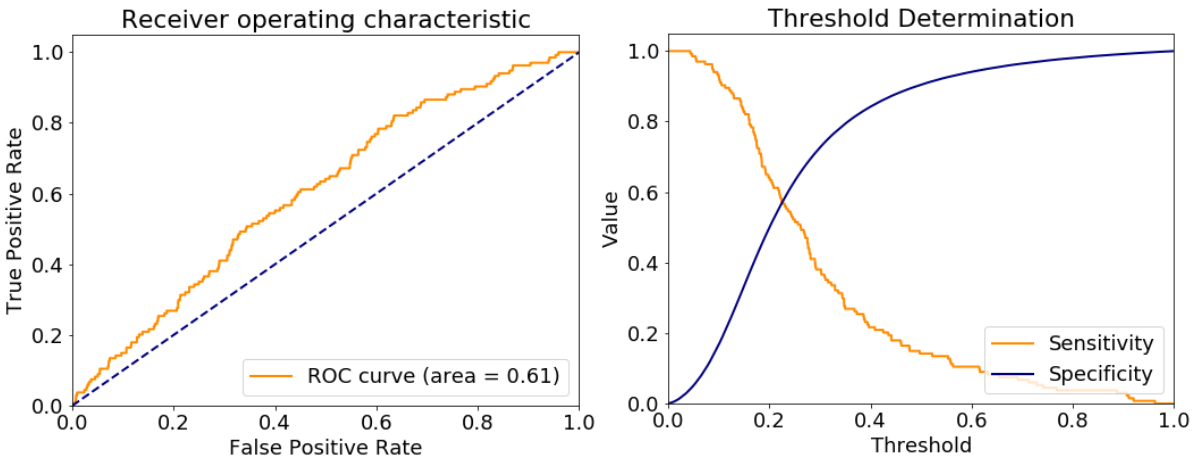


Figure 5-6: The ROC Curve and Threshold Determination of Conditional Logistic model

With respect to the multilayer LSTM, since the input of LSTM requires that the shape of the input vector for each time step to be the same, the variables in any time slice should be included into every time slice. For example, A_TH_AOG_OAFR is only significant from time slice 6 dataset, while this variable should be included into every time slice. Therefore, for every time slice, 13 variables from the above model results, together with all the 14 geometric variables were collected as the input of LSTM since we do not control the geometric location.

The LSTM algorithm was implemented based on TensorFlow™ 1.11 using the NVIDIA GTX 1080 Ti 11G GPU. Adam optimizer (Kingma and Ba, 2014) was utilized as the optimization algorithm. To prevent overfitting, the dropout (Srivastava et al., 2014) strategy with the probability of 0.5 was applied in our experiment. Three hyper-parameters, i.e., learning rate, training epoch, and mini-batch size, were tuned to achieve the best prediction performance. The training time for each run is around 2 hours. Table 5-4 presents all the parameters in the training phase of the LSTM.

Table 5-4: Parameters for LSTM

Parameter	Range	Result
Learning rate	(0.0001, 0.001, 0.005, 0.01)	0.01
Training epoch	(50, 100, 150, 200)	100
Mini-batch size	(100,000, 150,000, 200,000, 250,000)	150,000

Figure 5-7 shows the sensitivity and false alarm rate during the final training procedure, which indicate that there is no significant overfitting issue appeared in our final model.

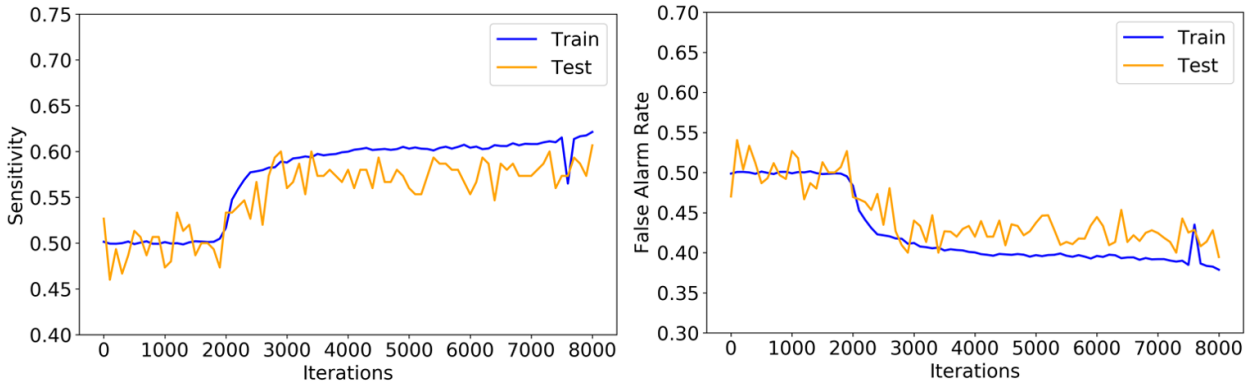


Figure 5-7: Training and Validation Metrics of the Final Model

As shown in Figure 5-8, the AUC value of LSTM is close to the conditional logistic model, while the ROC curve of LSTM is a little bit different from the conditional logistic model. The ROC

curve of LSTM tends to be close to the upper left corner, which indicates that the LSTM algorithm are slightly more powerful to predict the crash occurrence. Figure 5-8 also shows the determination of the classification threshold, where the intersecting point of sensitivity and specificity curve are very close to 1, which means that the algorithm will predict the crash occurrence only when the predicted crash risk is high.

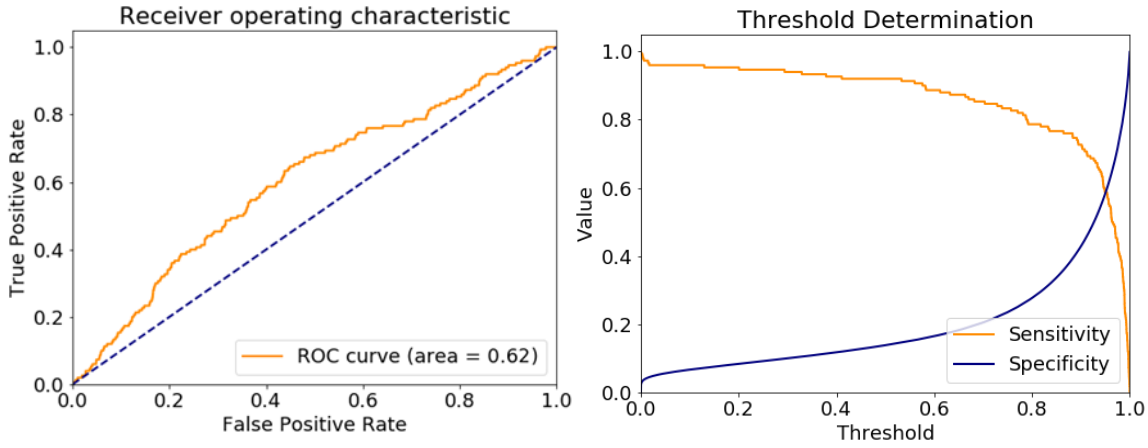


Figure 5-8: The ROC Curve and Threshold Determination of LSTM

Based on the determined threshold values, the sensitivity and false alarm rate of the final prediction were calculated for both conditional logistic model and LSTM. As can be seen from Figure 5-9, the prediction sensitivity is around 7% higher than the conditional logistic model, accompany with around 7% lower false alarm rate.

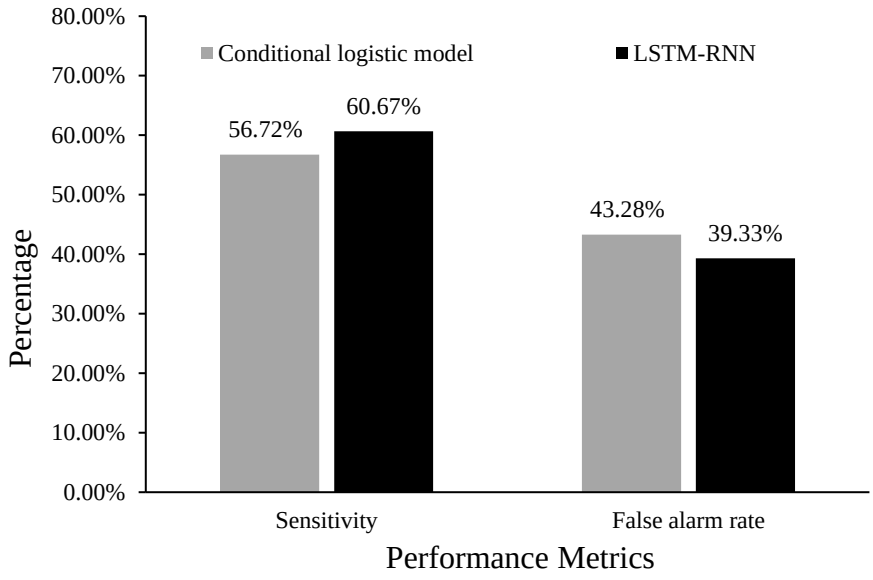


Figure 5-9: Model Comparison Results

From the application point of view, the best prediction sensitivity of 60.67% with 39.33% false alarm rate is still not good enough for practical deployment. However, these prediction algorithms were evaluated based on the unbalanced dataset, which could represent the practical performance, while the previous research evaluated based on artificially balanced dataset cannot guarantee their performance in real world situations. Moreover, the model comparison results showed the promising potential of deep learning algorithms over the conditional logistic model. More specifically, the conditional logistic model based on matched case-control dataset has been widely used by previous researchers on real-time crash risk analysis, which is quite maturing and robust. Nevertheless, the LSTM algorithms have only been used during recent years, especially for the sequence learning. There are still many potential future improvements and modifications could be done to improve the prediction performance. For example, only two LSTM layers were utilized in this study, while more LSTM layers could be included to build a deeper LSTM.

5.5 Conclusion and Discussion

This study tried to predict the real-time crash risk at signalized intersections by using multilayer LSTM recurrent neural network, which is designed for sequence modeling, and they can consider the time series characteristics automatically. First, a real-world unbalanced dataset was collected for every minute by incorporating real-time traffic, signal, and weather data. Also, both the approach-level and intersection-level geometric characteristics were included into the algorithm. To train the algorithm without losing any non-crash information, the synthetic minority over-sampling technique (SMOTE) was employed in this study to generate a balanced training dataset. In comparison, a traditional conditional logistic model was developed based on the matched case-control dataset with the control-to-case ratio of 10:1.

The prediction results showed that the LSTM with SMOTE could predicts 60.67% of the intersection crashes with a false alarm rate of 39.33%, which is better than the conditional logistic model (i.e., sensitivity: 56.72% and false alarm rate: 43.28%). This comparison results succeed in verifying the feasibility of applying LSTM in real-time crash risk prediction. Since this study is the first attempt in predicting real-time crash risk by using LSTM, therefore, the feasibility proof of the of LSTM with SMOTE is the major objective of this study.

With respect to the prediction performance, there are three possible reasons which may results in this relative low sensitivity. First, this study was tested on actual imbalanced data rather than the artificially balanced data. Second, the signalized intersections are much more complicated than freeway segments, therefore the crash occurrence at signalized intersections might be attributed to many other factors which were not captured by our algorithm. For example, there are many driving

behavior related factors, e.g., drowsy driving and distracted driving. Third, this study aims to predict the real-time crash risk during next 5-10 minutes based on the current data, which might be not long enough to be accurately predicted, if we increase it up to 5-15 minutes, or 5-20 minutes, they may have better prediction results, and the long prediction period might be more appropriate for proactive traffic management.

In summary, this study succeeds in verifying the feasibility of real-time crash risk prediction at signalized intersections by using LSTM recurrent neural network together with SMOTE over-sampling method. The results of this study could be utilized to predict real-time crash risk at signalized intersections in advance, which could assist operators to implement various pro-active traffic management strategies to reduce the risk in real-time. However, there are still some limitations for the current study. For example, there are several modified RNN structures which might be used in the future to improve the prediction performance. Even for the LSTM itself, there are several ways to improve the model performance, e.g., more LSTM layers, parameter regularization could reduce the over-fitting problem. For the resampling methods, there are many other ensemble sampling methods which can be used to generate balanced dataset, e.g., adaptive boosting and gradient tree boosting. In addition, crash occurrences have been widely proved to be highly influenced by drivers' characteristics and their driving behavior before crash occurrence, while these driver factors were not considered in this study. With the help of real-time driving behavior data, which could be enabled by connected vehicle technologies (Ekram and Rahman, 2018; Rahman and Abdel-Aty, 2018; Rahman et al., 2019; Rahman et al., 2018; Wu et al., 2019), more microscopic driver-level crash risk could be predicted in real-time.

CHAPTER 6: MODELING REAL-TIME CYCLE-LEVEL CRASH RISK AT SIGNALIZED INTERSECTIONS BASED ON HIGH-RESOLUTION EVENT-BASED DATA

6.1 Introduction

Signalized intersections serve a variety of road users to sequence right-of-way between intersecting streams of users. Due to the complex conflicting movements and frequently changing signals, signalized intersections are identified as typical high-risk locations. In the United States, nearly 27% (9047 fatalities) of all traffic fatalities are caused by intersection and intersection-related crashes in 2017 according to the data extracted from the Fatality Analysis and Reporting System (FARS). Given the serious traffic safety situation, investigating crash precursors for signalized intersections has been a critical research topic during past decades. Previous intersection safety studies mainly focused on modeling the relationships between annual crash frequency and static contributing factors, such as annual average daily traffic (AADT), traffic control, geometric design, etc. However, those static and yearly aggregated studies cannot capture the impacts of the real-time variation in traffic, weather, signal control characteristics, which might lead to misunderstanding of potential crash precursors. Also, with the advance of sensing technologies and smart city initiative, more and more real-time traffic data are available on arterials, which could be utilized to assist real-time pro-active traffic management.

As a prerequisite component for pro-active traffic safety management, real-time crash risk evaluation has gained a lot of attention from all over the world. However, previous research mainly focused on freeways, seldom on signalized intersections. Yuan and Abdel-Aty (2018) investigated the relationships between intersection approach-level crash risk and real-time traffic, signal timing, and weather characteristics based on 23 signalized intersections in Central Florida. More recently, Yuan et al. (2019) employed Long-Short Term Memory (LSTM) algorithm to predict real-time crash risk at signalized intersections based on Synthetic Minority Over-Sampling Technique (SMOTE), where they achieved better performance than traditional models. However, the previous two studies were conducted based on 5-min time intervals, which is inconsistent with the cyclical characteristics of the traffic flow at signalized intersections. Specifically, if the cycle length of a signalized intersection is 2 minutes, thus the 5-min time interval includes two complete cycles and one half-cycle. The data for the incomplete half-cycle might be collected during green phase, red phase or even both green and red phases. This uncertainty in data preparation may lead to biased model estimation results.

On the other hand, cycle-level traffic characteristics were proved to have significant impacts on intersection safety. For example, Essa and Sayed (2018b) developed cycle-level safety performance functions for signalized intersections based on automated traffic conflict analysis and they found that cycle-level traffic variables (e.g., maximum queue length, shockwave characteristics, and platoon ratio) are significantly correlated with the frequency of conflicts.

Similar findings were also been reached in their another study (Essa and Sayed, 2018a). However, these studies were conducted based on video-based conflict analyses, which has several limitations. For example, it's hard to collect and process long-period video data to get enough traffic conflict data. Also, the intrinsic relationship between traffic conflicts and crash occurrences is still quite obscure. Above all, cycle-level real-time crash risk analysis should be conducted while considering the cyclical characteristics of the traffic flow at signalized intersections.

In this context, the first step is to identify the exact signal cycle for every crash, which plays an important role in the identification of crash precursors. As the cycle lengths are usually 2-3 minutes, which may require that the precision of the reported crash time should be less than 1 minute. However, after carefully check the distribution of minutes of the reported crash times in four different crash databases, Imprialou and Quddus (2017) found that a disproportionate number of crashes have been reported at times when the minute indication ended with zero or five. Also, many of previous real-time safety studies utilized the traffic data several minutes (typically 5 minutes) prior to the reported crash time (Shi and Abdel-Aty, 2015; Wang et al., 2019a; Xu et al., 2013a; Yu et al., 2018). In order to determine the actual time of crashes, Lee et al. (2003) employed the shockwave theory which assumes that the time that the shockwave arrived at the crash location is assumed to be the actual crash time. While this method can only be used for uninterrupted roadway facilities (e.g., freeway) where the shockwave propagation

only appears during incidents. For signalized intersections, shockwave propagation could appear on both crash and normal conditions, which indicates that the identification of the actual times of intersection crashes might not be appropriate to use shockwave theory. In this study, with the help of high-resolution event-based Automated Traffic Signal Performance Measures (ATSPM) data on intersection approaches, the exact cycle of every crash occurrence could be verified based on the identification of abnormal detections.

As we all know, crash occurrences are usually considered as rare events due to the extreme high imbalance ratio between non-crash and crash cases (Basso et al., 2018; Theofilatos et al., 2018b; Yuan et al., 2019). However, traditional statistical models, such as logistic regression, can sharply underestimate the probability of rare events (King and Zeng, 2001). Therefore, undersampling strategies, which aim to balance the class distribution by eliminating samples from the majority class, have been widely employed in previous studies while modeling the probability of crash occurrences. Among them, the matched case-control design is the most popular undersampling strategy used in the field of real-time crash risk analysis (Abdel-Aty et al., 2004; Wang et al., 2019a; Yu et al., 2018). As stated by Theofilatos et al. (2018b), the choice of statistical method depends heavily on the sampling strategy. In order to evaluate the impact of undersampling strategies, two kinds of undersampling strategies (matched case-control and random undersampling) were employed in this study, while conditional logistic regression and

regular binary logistic regression were developed respectively for two kinds of balanced datasets.

Above all, this study aims to bridge the following research gaps: (1) determine the exact signal cycle where every crash occurred based on the high-resolution event-based ATSPM dataset; (2) model real-time crash risk at cycle-level for signalized intersections with the consideration of shockwave characteristics; (3) determine the best undersampling strategy while calibrating real-time crash risk prediction models for signalized intersections.

6.2 Data Preparation

Since all the ATSPM loop detectors are installed in the intersection approach areas, thus the ATSPM data are only capable to verify the exact crash time for those crashes occurred in the intersection approach areas. In total, 42 intersection approaches from 28 intersections were selected from Seminole County, Florida, as shown in Figure 6-1. A total of three datasets were collected in this study: (1) crash data from January 2017 to December 2018 provided by Signal Four Analytics (S4A); (2) high-resolution event-based signal timing and vehicle detection data during the same time period provided by Automated Traffic Signal Performance Measures (ATSPM) database; (3) weather characteristics collected by the nearest Local Climatological Data (LCD) station, which were archived by NOAA.

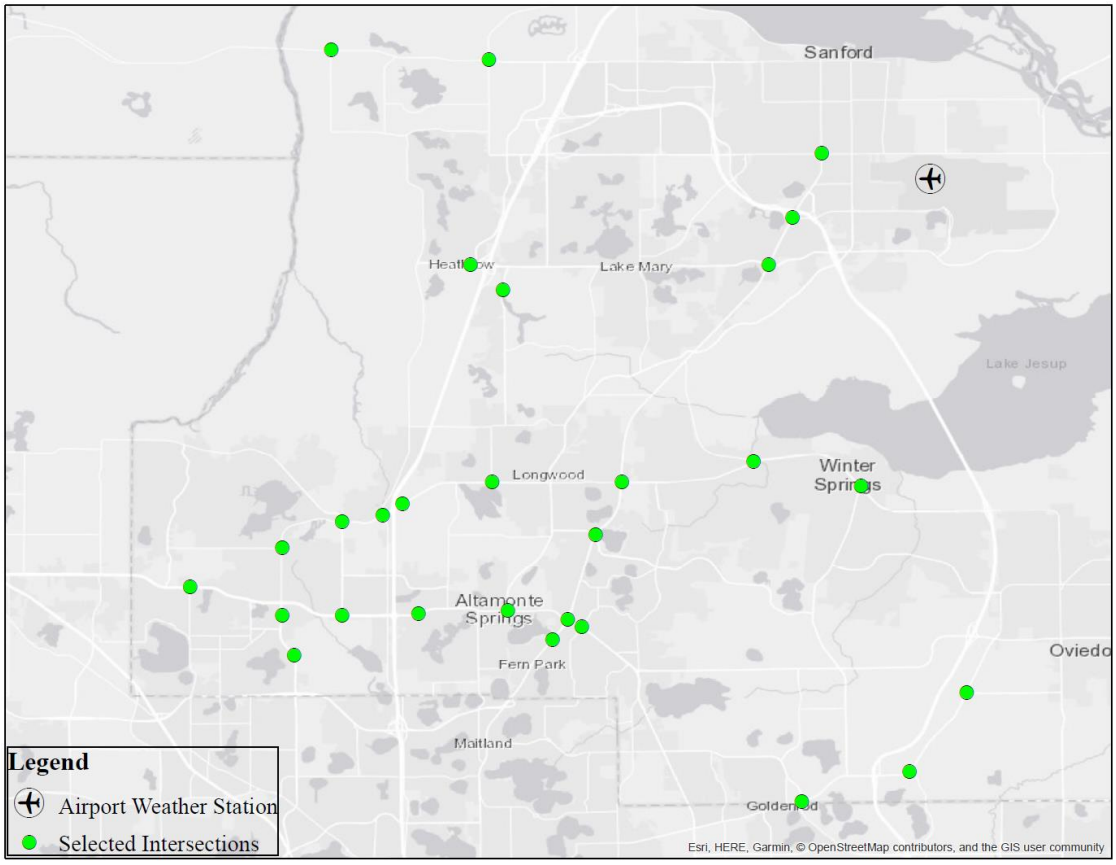


Figure 6-1: Selected Intersections.

6.2.1 Signal Timing and Vehicle Detection

For the selected 42 intersection approaches, a total of 210 loop detectors are installed on the through lanes. In order to analyze the traffic variation within intersection approach area, two sets of detector locations (henceforth referred to as front detectors and back detectors, respectively) were considered for all the selected intersection approaches. As the detector availability are different among intersection approaches, there exist three kinds of detector configurations for the selected approaches (configuration 1: 3 intersection approaches;

configuration 2: 6 intersection approaches; configuration 3: 33 intersection approaches), as shown in Figure 6-2.

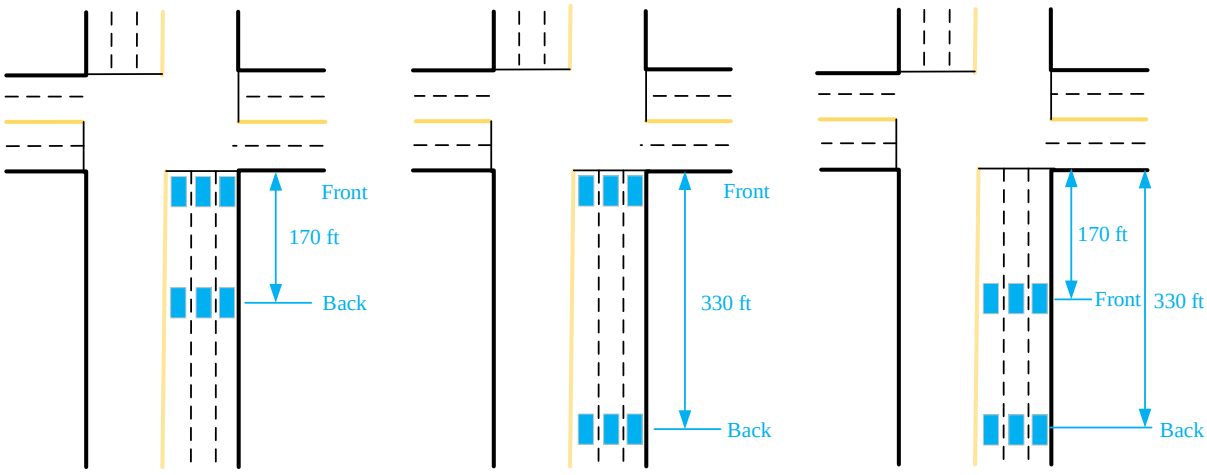


Figure 6-2: Detector Configurations on Intersection Approach.

Signal timing and lane-specific vehicle count data are calculated from the Automated Traffic Signal Performance Measure (ATSPM) database, which is recorded in the highest time resolution of controllers (0.1 seconds). Every event generated by signal controllers or loop detectors is recorded in sets of four bytes per event: two bytes for the timestamp of when the event occurred, one byte for event code type, and one byte for event parameter (for signifying detector numbers and phases). The event code is important for determining the type of reported activity, which could be phase initiation or termination, detection on/off, etc. Table 6-1 shows the sample set of events generated by a signal controller and 16 loop detectors at the intersection of US17-92 & 25th St.

Table 6-1: Sample Data Collected at the Intersection (US17-92 & 25th St).

Sample Raw Data			Description
Timestamp	Event Code	Event Parameter	
2018/12/1 00:00:09.1	1	6	Phase 6 Begin Green
2018/12/1 00:00:09.1	1	2	Phase 2 Begin Green
2018/12/1 00:00:37.4	82	9	Detector 9 On
2018/12/1 00:00:39.0	81	9	Detector 9 Off
2018/12/1 00:00:41.5	82	14	Detector 14 On
2018/12/1 00:00:42.2	81	14	Detector 14 Off
2018/12/1 00:00:50.2	82	10	Detector 10 On
2018/12/1 00:00:50.5	81	10	Detector 10 Off
2018/12/1 00:00:50.5	82	9	Detector 9 On
2018/12/1 00:00:51.3	82	13	Detector 13 On
2018/12/1 00:00:51.4	81	13	Detector 13 Off
2018/12/1 00:00:59.5	81	9	Detector 9 Off
2018/12/1 00:01:08.2	82	13	Detector 13 On
2018/12/1 00:01:08.3	81	13	Detector 13 Off
2018/12/1 00:01:09.2	82	14	Detector 14 On
2018/12/1 00:01:10.4	82	10	Detector 10 On
2018/12/1 00:01:10.4	81	14	Detector 14 Off
2018/12/1 00:01:10.7	81	10	Detector 10 Off
2018/12/1 00:01:32.9	82	13	Detector 13 On
2018/12/1 00:01:33.0	81	13	Detector 13 Off
2018/12/1 00:01:36.0	82	10	Detector 10 On
2018/12/1 00:01:36.4	81	10	Detector 10 Off
2018/12/1 00:01:43.5	82	13	Detector 13 On
2018/12/1 00:01:43.7	81	13	Detector 13 Off
2018/12/1 00:01:50.8	7	6	Phase 6 Green Termination
2018/12/1 00:01:50.8	8	2	Phase 2 Begin Yellow Clearance
2018/12/1 00:01:50.8	8	6	Phase 6 Begin Yellow Clearance
2018/12/1 00:01:50.8	7	2	Phase 2 Green Termination
2018/12/1 00:01:53.9	82	13	Detector 13 On
2018/12/1 00:01:54.0	81	13	Detector 13 Off
2018/12/1 00:01:55.6	9	6	Phase 6 End Yellow Clearance
2018/12/1 00:01:55.6	9	2	Phase 2 End Yellow Clearance
2018/12/1 00:01:55.6	10	2	Phase 2 Begin Red Clearance
2018/12/1 00:01:55.6	10	6	Phase 6 Begin Red Clearance

Based on the high-resolution event based ATSPM data, several signal timing and vehicle detection related metrics could be inferred. For example, the time difference between the start and the end of a signal event represent the phase duration, the time interval between “detector on” and “detector off” indicates the detector occupancy time, and the time interval between “detector off” and “detector on” denotes the vehicle gap. In this study, all the variables were

aggregated at cycle-level, which means that every variable will be generated for every signal cycle (i.e., the time interval between the start of red phase and the end of yellow phase).

For traffic volume characteristics, various types of overall average flow ratio (OAFR) were collected in addition to the basic cycle volume to consider the variation in traffic flow across lanes. For further details about the calculation of OAFR, please refer to Yuan and Abdel-Aty (2018). In addition, real-time traffic progression measures were also collected, including percent of green (POG), percent on yellow (POY), arrival on green ratio (AOGR), arrival on yellow ratio (AOYR), and platoon ratio (PR).

$$\boldsymbol{GR_i = \frac{t_{g,i}}{C_i}} \quad (6\text{-}1)$$

$$POG_i = \frac{V_{g,i}}{V_{c,i}} \quad (6\text{-}2)$$

$$POY_i = \frac{V_{y,i}}{V_{c,i}} \quad (6\text{-}3)$$

$$AOGR_i = \frac{POG_i}{t_{g,i}} = \frac{V_{g,i}}{V_{c,i}} / t_{g,i} \quad (6\text{-}4)$$

$$AOYR_i = \frac{POY_i}{t_{y,i}} = \frac{V_{y,i}}{V_{c,i}} / t_{y,i} \quad (6\text{-}5)$$

$$PR_i = \frac{POG_i}{GR_i} = (\frac{V_{g,i}}{V_{c,i}}) / (\frac{t_{g,i}}{C_i}) \quad (6\text{-}6)$$

Where $\boldsymbol{t_{g,i}}$ is the duration of the green phase during the *i*th cycle; $t_{y,i}$ is the duration of the yellow phase during the *i*th cycle; $\boldsymbol{C_i}$ is the cycle length of the *i*th cycle; $\boldsymbol{V_{g,i}}$ represents the

number of vehicles arriving on green during the *i*th cycle; $\boldsymbol{V_{c,i}}$ is the total volume during the *i*th cycle; $V_{y,i}$ represents the number of vehicles arriving on yellow during the *i*th cycle.

Moreover, different kinds of headway and occupancy related variables were collected based on the vehicle detector actuation events. Meanwhile, the traffic variation between front detectors and back detectors were also collected to represent the traffic variation within the intersection approach areas. Table 6-2 summarizes the required information for every cycle-level variable.

Table 6-2: Required Data Elements for Selected ATSPM Measures.

Type	Variables	Description	Required Event Code
Traffic Volume	Cycle_Volume	Through volume per intersection approach per cycle	1. Phase Begin Green 8. Phase Begin Yellow Clearance 10. Phase Begin Red Clearance 82. Detector On
	OAFR_Back_Cycle	Overall average flow ratio among back detectors per cycle	
	OAFR_Back_Green	Overall average flow ratio among back detectors during green phase per cycle	
	OAFR_Back_Red	Overall average flow ratio among back detectors during red phase per cycle	
	OAFR_Front_Green	Overall average flow ratio among front detectors during green phase per cycle	
Signal Timing	Cycle_Len	Cycle length (s)	1. Phase Begin Green 8. Phase Begin Yellow Clearance 10. Phase Begin Red Clearance
	Green_Ratio	Percentage of the length of green time per cycle	
Traffic Progression	POG_Back	Percentage of arrival on green of back detectors	1. Phase Begin Green 8. Phase Begin Yellow Clearance 10. Phase Begin Red Clearance 82. Detector On
	POY_Back	Percentage of arrival on yellow of back detectors	
	POR_Back	Percentage of arrival on red of back detectors	
	AOGR_Back	Arrival on green ratio of back detectors	
	AOYR_Back	Arrival on yellow ratio of back detectors	
	AORR_Back	Arrival on red ratio of back detectors	
	Platoon_Ratio	Platoon ratio	
Headway and Occupancy	AVG_Occupancy_Back_Green	Average occupancy of back detectors during green phase (s)	1. Phase Begin Green 8. Phase Begin Yellow Clearance 10. Phase Begin Red Clearance 81. Detector Off 82. Detector On
	STD_Occupancy_Back_Green	Standard deviation of occupancy of back detectors during green phase (s)	
	AVG_Headway_Back_Green	Average headway of back detectors during green phase (s)	
	STD_Headway_Back_Green	Standard deviation of headway of back detectors during green phase (s)	
	AVG_Occupancy_Back_Red	Average occupancy of back detectors during red phase (s)	
	STD_Occupancy_Back_Red	Standard deviation of occupancy of back detectors during red phase (s)	
	AVG_Headway_Back_Red	Average headway of back detectors during red phase (s)	
	STD_Headway_Back_Red	Standard deviation of headway of back detectors during red phase (s)	
	AVG_Occupancy_Front_Green	Average occupancy of front detectors during green phase (s)	
	STD_Occupancy_Front_Green	Standard deviation of occupancy of front detectors during green phase (s)	
	AVG_Headway_Front_Green	Average headway of front detectors during green phase (s)	
	STD_Headway_Front_Green	Standard deviation of headway of front detectors during green phase (s)	
Traffic Variation	Diff_OAFR_Green	Difference in the OAFR during green phase between front detectors and back detectors	
	Diff_AVG_Occupancy_Green	Difference in the average occupancy during green phase between front detectors and back detectors (s)	
	Diff_STD_Occupancy_Green	Difference in the standard deviation of occupancy during green phase between front detectors and back detectors (s)	
	Diff_AVG_Headway_Green	Difference in the average headway during green phase between front detectors and back detectors (s)	
	Diff_STD_Headway_Green	Difference in the standard deviation of headway during green phase between front detectors and back detectors (s)	

6.2.2 Shockwave Characteristics

Given the high-resolution event-based ATSPM data, all the shockwave characteristics could be estimated in real-time by applying shockwave theory (Liu et al., 2009; Wu and Liu, 2014). In this study, the maximum queue length, queuing shockwave speed, and shockwave area were calculated based on the high-resolution data collected by back detectors. Figure 6-3 shows a typical traffic shockwave at an intersection, where L_{max} indicates the maximum queue length, V_1 represents the queuing shockwave speed, S is the shockwave area.

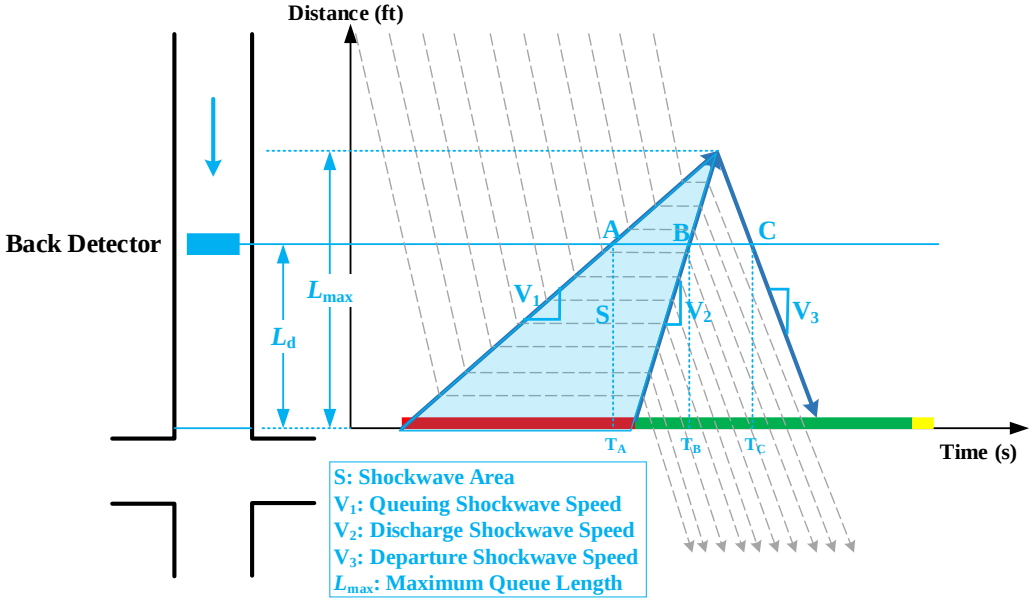


Figure 6-3: Illustration of Traffic Shockwave at an Intersection.

As demonstrated by Liu et al. (2009), if point A exists (i.e., queuing shockwave (V_1) propagates beyond the location of the back detectors), the back detectors would be occupied by a long time until TB when B point appears (i.e., discharge shockwave (V_2) propagates to the location of the

detector). Also, the C point where the end of the queue passes the detector could be identified as the time when the traffic flow at the detector changes from saturated discharging flow to normal arrival flow. Therefore, the three shockwave characteristics (i.e., queuing shockwave speed (V_1^i), maximum queue length (L_{max}^i), and shockwave area (S^i)) could be calculated:

$$\boldsymbol{V_1^i = \frac{0 - Q_a^i}{k_j - k_a^i}} \quad (6\text{-}7)$$

$$V_2^i = \frac{L_d}{T_B^i - T_g^i} \quad (6\text{-}8)$$

$$V_3^i = \frac{Q_m - Q_a^i}{k_m - k_a^i} \quad (6\text{-}9)$$

$$L_{max}^i = L_d + \frac{(T_C^i - T_B^i)}{(\frac{1}{V_2^i} + \frac{1}{V_3^i})} \quad (6\text{-}10)$$

$$S^i = \frac{(t_{r,i} \times L_{max}^i)}{2} \quad (6\text{-}11)$$

Where Q_a^i and k_a^i are the average arrival flow rate and density during the *i*th cycle; k_j indicates the jammed density; V_2^i and V_3^i represent the speed of discharge shockwave and departure shockwave. L_d represents the distance between stop bar and back detectors; T_B^i is the time when the discharge shockwave propagates to the location of the detector during the *i*th cycle; T_g^i is the time when the green phase starts during the *i*th cycle; Q_m and k_m represent the saturated flow rate and density; T_C^i represents the time when the end of the queue passes the detector; T_B^i is the time when the discharge shockwave propagates to the location of the detector; $t_{r,i}$ is the duration of the red phase during the *i*th cycle.

In addition, if point A does not exist, i.e., the maximum queue length is less than L_d, the maximum queue length could be estimated based on the simple input-output method (Liu et al., 2009). For more details about the identification procedures of points A, B, C, please refer to Liu et al. (2009). Table 6-3 summarizes the required information for the three shockwave related variables.

Table 6-3: Required Data Elements for Shockwave Characteristics.

Variables	Description	Required Event Code
Max_Queue_Length	Maximum queue length (mile)	1. Phase Begin Green 10. Phase Begin Red Clearance 81. Detector Off 82. Detector On
Shock_Wave_Area	Shockwave area (mile.s)	
Queuing_Shockwave_Spd	Queuing shockwave speed (ft/s)	

Figure 6-4 shows a one-day sample of shockwave characteristics for the intersection of US17-92 & 25th St on 05/03/2017. Among the figure, the light blue line indicates the maximum queue length, the light green line represents the absolute queuing shockwave speed, and the red bar indicates the actual crash. This figure clearly shows that the crash occurred during the time period with longer queue length and higher absolute queuing shockwave speed.

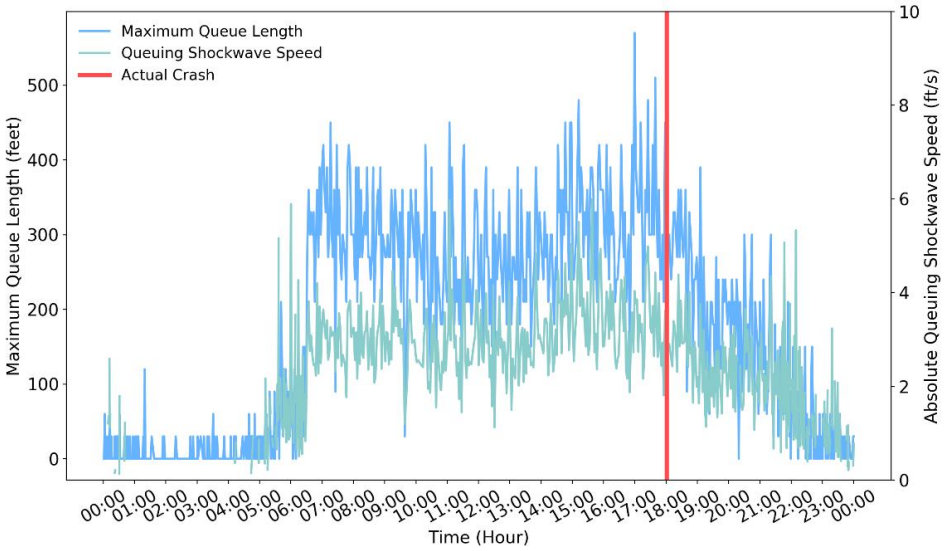


Figure 6-4: Shockwave Characteristics Data for an Intersection (US17-92 & 25th St) on 05/03/2017.

6.2.3 Weather

Three weather-related variables (weather type, visibility, and hourly precipitation) were collected from the nearest LCD airport weather station (as shown in Figure 6-1). As weather data are not recorded continuously, once weather condition changes and reaches a preset threshold, a new record will be added to the archived data. For every cycle, the closest weather record prior to the begin of every cycle was extracted. Table 6-4 shows the detailed description of weather data.

Table 6-4: Description of Weather Data.

Variables	Description
Visibility	Visibility (mile)
Weather_Types	Weather type: 0 for normal and 1 for adverse weather.
Precipitation	Hourly precipitation (inch).

6.2.4 Crash Data and Corresponding Signal Cycle

Signal four analytics (S4A) system provides detailed crash information, including crash time, location, severity, type, etc. First, 362 crashes occurred within the selected intersection approaches (from stop bar to 250 feet upstream) from January 2017 to December 2018 were collected. It is worth noting that the left-turn phases of the selected intersections are served with the combination of lead-lead sequence, lag-lag sequence, and lead-lag sequence, which results in huge complexity in the interaction between the left-turn and through movements. Consequently, only the through movement related variables and the corresponding crashes were considered in this study as an instance to verify the feasibility of cycle-level real-time safety analysis. After excluding all the crashes occurred on the left turning lanes, as well as the crashes without corresponding traffic data, there are 252 crashes remaining in the final dataset. Among them, 190 (75.40%) crashes are rear-end crashes, 41 (16.27%) crashes are sideswipe crashes, and 21 (8.33%) crashes are other types of crashes.

In order to determine the actual signal cycle of crash occurrence, the corresponding high-resolution vehicle detection and signal timing data during the time interval starts from 15 minutes before the recorded crash time to 15 minutes after the reported crash time were extracted for every crash and then plotted to identify the potential abnormal detections, which is consistent with the previous research (Wang et al., 2019b). As shown in Figure 6-5, two kinds of abnormal detections were considered to verify the reported crash time.

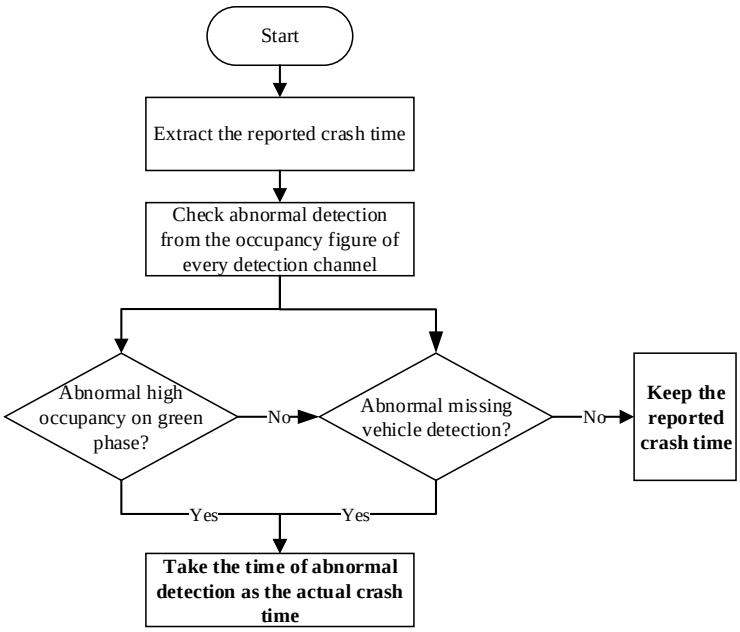
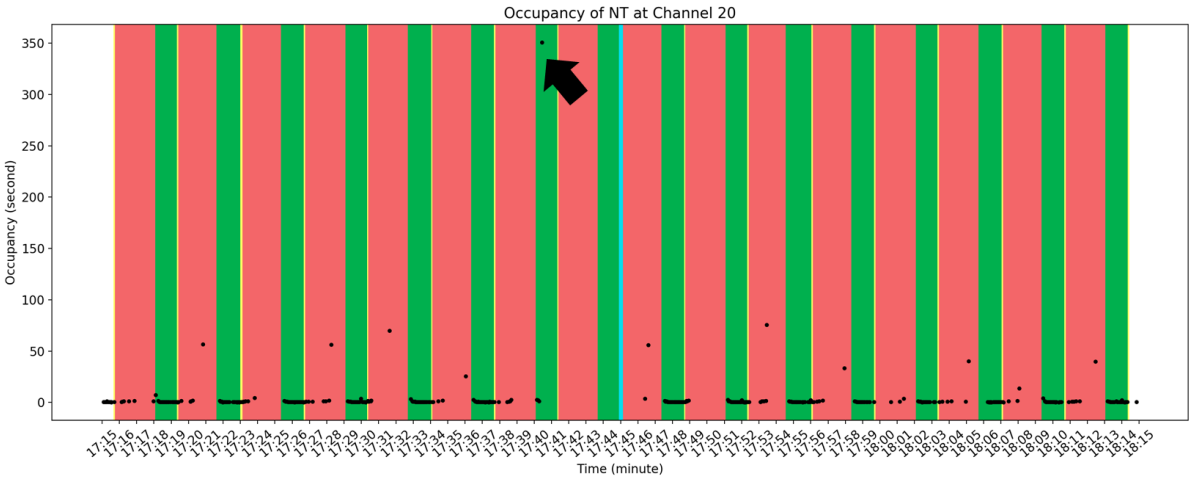
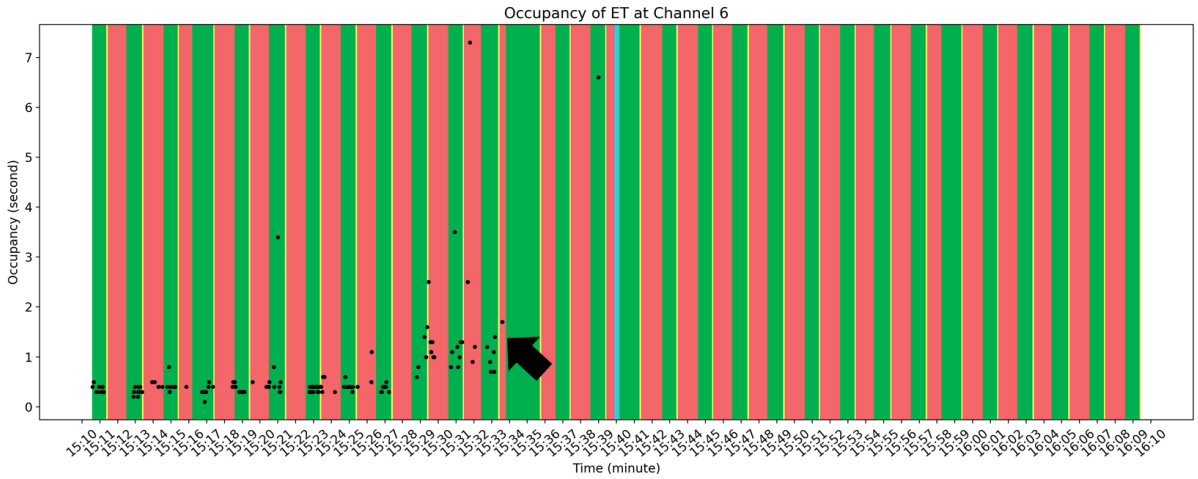


Figure 6-5: Determination of the Actual Time of Crash.

Figure 6-6 shows the examples of two kinds of abnormal detections, i.e., abnormal high occupancy on green phase, and abnormal missing vehicle detection. The light blue bars in the middle of the x-axis indicate the reported crash time, the y-axis represents the occupancy of every vehicle detection, every black dot indicates every detected vehicle, and the black arrows point at the abnormal detections. In the first crash example, the reported crash time is 17:45:00, while the time of the abnormal detection with extremely high occupancy (350 seconds) on green phase is 17:40:27. Therefore, the reported time of this crash was modified to be 17:40:27. In the second crash example, the reported crash time is 15:40:00, while the time of the abnormal detection with unusual missing detection is 15:33:36. Therefore, the reported time of this crash was modified to be 15:33:36.



(1) Crash example 1: abnormal high occupancy on green phase



(2) Crash example 2: abnormal missing vehicle detection

Figure 6-6: Examples of Abnormal Events.

Above all, 80 (32%) crashes were identified with abnormal detection and the reported crash times were modified to be the corresponding time of abnormal detections. Figure 6-7 shows the distribution of the time difference between reported crash time and modified crash time for the 80 identified crashes. The average time difference for those crashes which were identified with abnormal detections is 133 seconds, which is in line with previous research (Imprialou and Quddus, 2017).

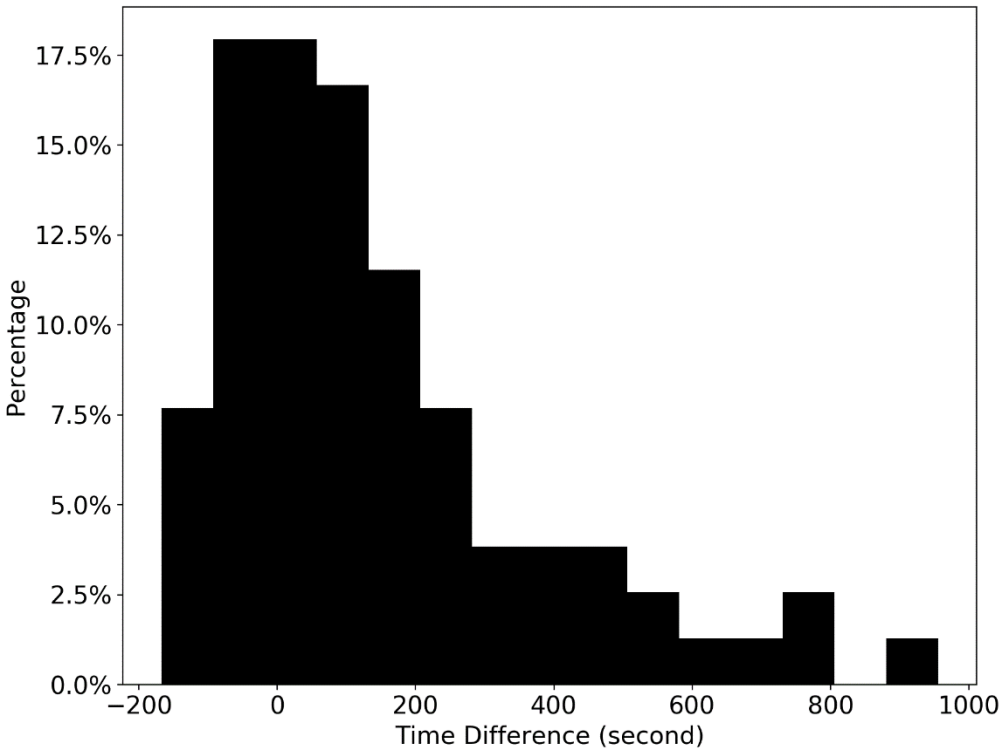


Figure 6-7: Distribution of the Time Difference between Reported Crash Time and Modified Crash Time.

All the abovementioned traffic and weather-related variables were prepared at the cycle-level. Based on the modified crash time, the corresponding cycle for every crash could be identified. In order to consider the effect of time dependency and then model the impact of the traffic status during preceding cycles on the risk of the crash cycle, five cycles prior to the crash cycle were considered to develop five models, respectively. Different labelling strategies were employed for different cycle models, as shown in Figure 6-8. For example, for the cycle-1 model, only the first cycle prior to the crash cycle was labeled as "1" (crash event), and the crash cycles and all the cycles within two hours after the crash cycles were excluded to eliminate the influence of crash occurrence on traffic condition.

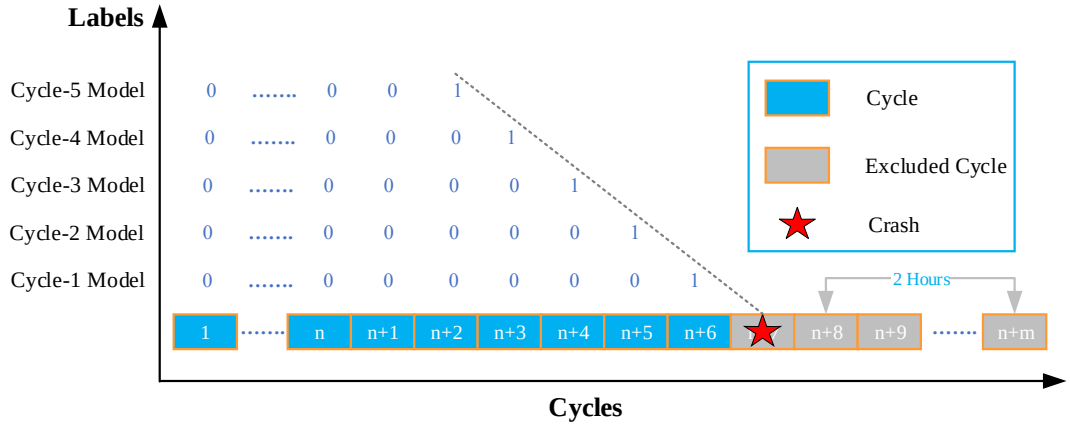


Figure 6-8: Illustration of Data Labelling for Every Consecutive Time Series Data

In summary, the final dataset includes 12,291,308 cycles, where 252 of them are crash events and 12,291,056 cycles are non-crash events. Table 6-5 shows the descriptive statistics of collected variables for both crash and non-crash events.

Table 6-5: Descriptive Statistics of Collected Variables (Crash and Non-Crash Events).

Type	Variables	Crash Event		Non-Crash Event	
		Mean (Std)	(Min, Max)	Mean (Std)	(Min, Max)
Traffic Volume	Cycle_Volume	62.421 (38.294)	(0, 173)	25.85 (29.323)	(0, 332)
	OAFR_Cycle_Back	1.07 (0.176)	(0.943, 2.184)	1.123 (0.26)	(0.943, 4.465)
	OAFR_Green_Back	1.073 (0.173)	(0.944, 2.19)	1.123 (0.259)	(0.943, 4.283)
	OAFR_Red_Back	1.18 (0.412)	(0.944, 4.803)	1.124 (0.246)	(0.943, 4.803)
	OAFR_Green_Front	1.38 (1.608)	(0.944, 13.683)	1.193 (0.579)	(0.943, 13.683)
Signal Timing	Cycle_Len	165.338 (47.035)	(36.5, 399.8)	123.44 (71.188)	(15.1, 994.9)
	Green_Ratio	0.473 (0.139)	(0, 0.896)	0.532 (0.184)	(0, 1)
	POG_Back	0.624 (0.213)	(0, 1)	0.609 (0.305)	(0, 1)
	POY_Back	0.038 (0.051)	(0, 0.333)	0.05 (0.136)	(0, 1)
	POR_Back	0.338 (0.212)	(0, 1)	0.341 (0.295)	(0, 1)
	AOGR_Back	0.851 (0.323)	(0, 2.353)	1.121 (1.042)	(0, 14.925)
	AOYR_Back	0.757 (0.931)	(0, 6.803)	0.925 (2.574)	(0, 37.037)
	AORR_Back	0.501 (0.562)	(0, 6.135)	0.922 (1.203)	(0, 14.354)
	Platoon_Ratio	134.773 (40.833)	(0, 274.91)	114.128 (57.886)	(0, 671.519)
Headway and Occupancy	Avg_Occupancy_Green_Back	0.751 (5.305)	(0.1, 84.013)	0.3 (0.521)	(0.1, 84.013)
	Std_Occupancy_Green_Back	1.202 (13.806)	(0, 219.103)	0.135 (0.395)	(0, 219.103)
	Avg_Headway_Green_Back	10.637 (17.938)	(1.64, 155.7)	42.29 (75.625)	(0.2, 1066.1)
	Std_Headway_Green_Back	11.708 (23.684)	(0.212, 266.296)	29.296 (48.461)	(0, 687.873)
	Avg_Occupancy_Red_Back	0.616 (2.473)	(0.1, 36.85)	0.428 (1.729)	(0.1, 47.3)
	Std_Occupancy_Red_Back	0.246 (0.602)	(0, 7.038)	0.123 (0.36)	(0, 9.509)
	Avg_Headway_Red_Back	16.304 (19.176)	(2.173, 186.867)	45.003 (73.769)	(1.1, 1065.1)
	Std_Headway_Red_Back	15.305 (29.911)	(0.424, 439.739)	28.24 (42.585)	(0, 586.121)
	Avg_Occupancy_Green_Front	0.873 (0.869)	(0.29, 7.1)	0.587 (0.571)	(0.1, 18.25)
	Std_Occupancy_Green_Front	1.318 (10.427)	(0, 164.968)	0.297 (0.505)	(0, 164.968)
	Avg_Headway_Green_Front	10.285 (21.251)	(1.643, 205.2)	45.159 (82.189)	(0.4, 1330.2)
	Std_Headway_Green_Front	10.858 (23.184)	(0.523, 264.882)	30.147 (50.481)	(0, 728.886)

Type	Variables	Crash Event		Non-Crash Event	
		Mean (Std)	(Min, Max)	Mean (Std)	(Min, Max)
Traffic Variation	Diff_OAFR_Green	0.392 (1.584)	(0, 12.536)	0.161 (0.547)	(0, 12.536)
	Diff_Avg_Occupancy_Green	0.509 (0.578)	(0, 6)	0.321 (0.354)	(0, 6)
	Diff_Std_Occupancy_Green	0.458 (0.866)	(0, 9.591)	0.24 (0.457)	(0, 9.591)
	Diff_ Avg _Headway_Green	3.601 (13.379)	(0, 196.935)	11.082 (37.458)	(0, 646.425)
	Diff_Std_Headway_Green	4.555 (8.809)	(0, 121.21)	7.153 (19.221)	(0, 290.762)
Shockwave Characteristics	Max_Queue_Length	0.065 (0.05)	(0, 0.297)	0.026 (0.033)	(0, 0.384)
	Shock_Wave_Area	2.959 (2.824)	(0, 20.384)	0.942 (1.55)	(0, 20.384)
	Queuing_Shockwave_Spd	-4.2 (4.938)	(-59.041, -0.188)	-2.271 (2.381)	(-59.041, -0.016)
Weather	Visibility	9.822 (0.822)	(2.75, 10)	9.758 (1.031)	(0.125, 10)
	Weather_Type	0.115 (0.32)	(0, 1)	0.075 (0.263)	(0, 1)
	Precipitation	0.004 (0.016)	(0, 0.14)	0.002 (0.011)	(0, 0.185)

6.3 Methodology

Figure 6-9 shows the framework of model development. First, the original imbalanced dataset (imbalance ratio: 1: 48,954) was split into training dataset and test dataset based on time sequence, where the data before 9/16/2018 were selected as training dataset (200 crash events and 9,829,994 non-crash events) and the remaining data were selected as test dataset (52 crash events and 2,460,803 non-crash events). Second, two kinds of undersampling strategies (i.e., matched case-control and random undersampling) were employed on the training dataset to generate balanced datasets to calibrate the statistical models. For the matched case-control strategy, four factors, i.e., intersection ID, approach ID, hour of day, and day of week, were controlled as matching factors. Therefore, all the corresponding non-crash events for every crash event could be identified by using these matching factors and then a specific number of non-crash events would be randomly selected from the group of non-crash events. According to previous studies, 4:1 is the most

commonly used control-to-case ratio (Ahmed and Abdel-Aty, 2012; Shi and Abdel-Aty, 2015; Yu et al., 2018; Yu et al., 2016; Zheng et al., 2010). Therefore, 4 non-crash events were selected for every crash event, and the final matched case-control dataset includes 252 crash events and 1008 non-crash events. For the random undersampling strategy, the same crash to non-crash ratio was utilized, and 1008 non-crash events were randomly selected from 1,324,453 non-crash events in the training dataset. At last, all the models were evaluated based on the same imbalanced raw dataset.

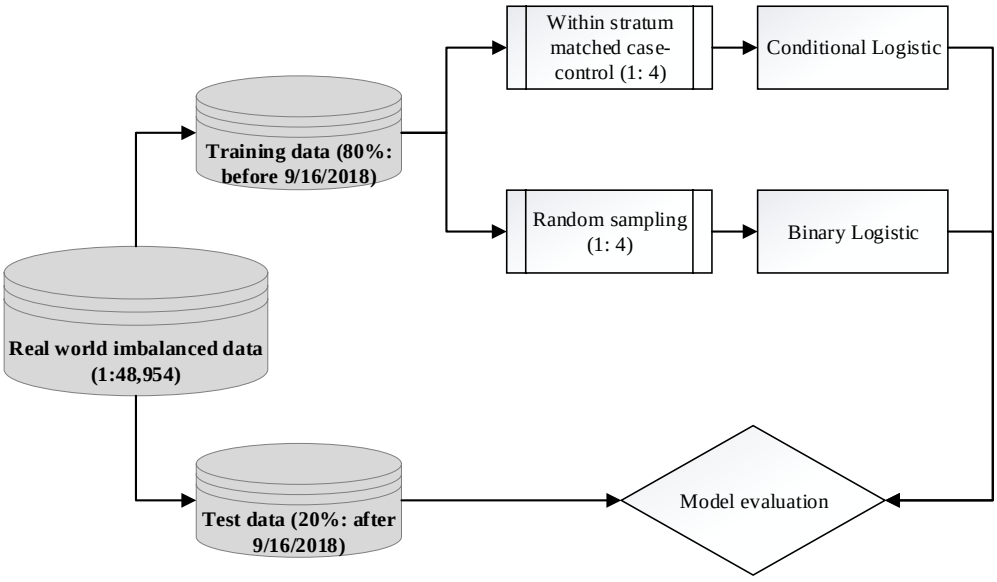


Figure 6-9: Framework of Model Development.

6.3.1 Conditional Logistic Model

Suppose there are one crash case (y_{i0}=1) and m non-crash cases (y_{ij}=0) in stratum i, i=1, 2, …, N and j=1, 2, …, m. p_{ij} is the probability that the jth observation in the ith stratum is a crash. This crash probability could be expressed as:

$$\boldsymbol{y_{ij} \sim Bernoulli(p_{ij})} \quad (6\text{-}12)$$

$$logit(p_{ij}) = \alpha_i + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij} \quad (6\text{-}13)$$

Where α_i is the intercept term for the *i*th stratum; $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$ is the vector of regression coefficients for *k* independent variables; $\boldsymbol{X_{ij}} = (X_{1ij}, X_{2ij}, \dots, X_{kij})$ is the vector of *k* independent variables.

To consider the stratification in the observed data, the stratum-specific intercept α_i is considered to be nuisance parameter, and the conditional likelihood for the *i*th stratum would be expressed as (Hosmer Jr et al., 2013):

$$\boldsymbol{l_i(\beta) = \frac{\exp(\sum_{u=1}^{k} \beta_u X_{ui0})}{\sum_{j=0}^{m} \exp(\sum_{u=1}^{k} \beta_u X_{uij})}} \quad (6\text{-}14)$$

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

$$\boldsymbol{L(\beta) = \prod_{i=1}^{N} l_i(\beta)} \quad (6\text{-}15)$$

Since the full conditional likelihood is independent of stratum-specific intercept α_i, thus Eq. 13 cannot be used to estimate the crash probabilities. However, the $\boldsymbol{\beta}$ coefficients can be estimated by Eq. (6-15). These estimates represent the logarithm of the odds ratios of corresponding variables and can be used to approximate the relative risk of a crash. Furthermore, the log-odds-ratios can also be used to develop a prediction model under this matched case-control analysis. Suppose two observation vectors $\boldsymbol{X_{i1}} = (X_{1i1}, X_{2i1}, \dots, X_{ki1})$ and $\boldsymbol{X_{i2}} = (X_{1i2}, X_{2i2}, \dots, X_{ki2})$ from the *i*th strata, the odds ratio of crash occurrence caused by observation vector $\boldsymbol{X_{i1}}$ relative to observation vector $\boldsymbol{X_{i2}}$ could be calculated as:

$$\frac{\boldsymbol{p_{i1}/(1-p_{i1})}}{\boldsymbol{p_{i2}/(1-p_{i2})}} = \mathbf{exp}[\sum_{\boldsymbol{u=1}}^{\boldsymbol{k}} \boldsymbol{\beta_u}(\boldsymbol{X_{ui1}} - \boldsymbol{X_{ui2}})] \quad (6\text{-}16)$$

The right side of Eq. (6-16) is independent of α_i and can be calculated using the estimated $\boldsymbol{\beta}$ coefficients. Thus, the above odds ratio could be utilized for predicting crash occurrences by replacing $\boldsymbol{X_{i2}}$ with the vector of the independent variables in the *i*th stratum of non-crash cases. One may use the simple average of all non-crash observations within the stratum for each variable. Let $\overline{\boldsymbol{X}}_{\boldsymbol{i}} = (\bar{X}_{1i}, \bar{X}_{2i}, \ldots, \bar{X}_{ki})$ denotes the vector of mean values of non-crash cases of the *k* variables within the *i*th stratum. Then the odds ratio of a crash relative to the non-crash cases in the *i*th stratum could be approximated by:

$$\frac{\boldsymbol{p_{i1}/(1-p_{i1})}}{\boldsymbol{p_{\bar{i}}/(1-p_{\bar{i}})}} = \mathbf{exp}[\sum_{\boldsymbol{u=1}}^{\boldsymbol{k}} \boldsymbol{\beta_u}(\boldsymbol{X_{ui1}} - \overline{\boldsymbol{X}}_{\boldsymbol{ui}})] \quad (6\text{-}17)$$

6.3.2 Binary Logistic Model

Suppose the crash occurrence has the outcomes y_i=1 (crash event) and y_i=0 (non-crash event) with the respective probabilities of p_i and 1-p_i, *i*=1, 2, …, *M*. *M* represents the total number of samples, which equals to *N(m+1)* in this study. The binary logistic regression can be expressed as follows:

$$y_i \sim Bernoulli(p_i) \quad (6\text{-}18)$$

$$logit(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} \quad (6\text{-}19)$$

Where β_0 is the intercept; $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_K)$ is the vector of coefficients for K independent variables; $\boldsymbol{X_i} = (X_{1i}, X_{2i}, \dots, X_{Ki})$ is the vector of K independent variables for the ith observation.

6.4 Result analysis

6.4.1 Effect Analysis

In order to consider the effect of time dependency, five cycles prior to the crash cycle were considered to develop five models, respectively. Table 6-6 shows the estimation results of the conditional logistic model, which was developed based on the data prepared by using matched case-control design. The model comparison results based on the test AUC values indicate that the cycle-1 model performs much better than the other four models, which means that the closest signal cycle plays the most important role in the real-time crash risk of the current signal cycle.

Table 6-6: Estimation Results of Conditional Logistic Model.

Variables	Cycle-1	Cycle-2	Cycle-3	Cycle-4	Cycle-5
	Mean (P-value)	**Mean (P-value)**	**Mean (P-value)**	**Mean (P-value)**	**Mean (P-value)**
Cycle_Volume	0.01 (0.095) *	-	-	-	-
OAFR_Green_Front	-	-	-	-	0.418 (0.045) **
OAFR_Red_Back	0.719 (0.009) **	-	-	-	-
OAFR_Cycle_Back	-	1.303 (0.021) **	-	-	-
AOYR_back	-	-	-	0.126 (0.097) *	-
Std_Headway_Red_Back	-	-	0.008 (0.059) *	0.008 (0.096) *	-
Avg_Occupancy_Green_Front	-	0.824 (<0.001) **	-	0.323 (0.029) **	0.329 (0.052) *
Std_Occupancy_Green_Front	-	-	0.276 (0.009) **	-	-
Diff_OAFR_Green	0.531 (0.011) **	-	0.713 (0.026) **	-	-
Diff_Avg_Occupancy_Green	0.827 (<0.001) **	-	-	-	-
Diff_Avg_Headway_Green	-	-0.019 (0.098) *	-	-	-
Max_Queue_Length	-	-	3.655 (0.073) *	5.209 (0.013) **	4.672 (0.048) **
Shock_Wave_Area	0.142 (0.004) **	-	-	-	-
AUC	0.8046	0.6005	0.6597	0.6902	0.7239

*Note: The cells noted by ** are significant at the 0.05 level; The cells noted by * are significant at the 0.1 level.*

Overall, 13 variables are found to be significant across all the cycle models, which could be classified as five types, i.e., traffic volume, signal timing, headway and occupancy, traffic variation, and shockwave characteristics. (1) Four traffic-volume-related variables (Cycle_Volume, OAFR_Green_Front, OAFR_Red_Back, and OAFR_Cycle_Back) are found to be positively associated with the real-time crash risk, which means that higher cycle volume and overall average flow ratio across lanes could significantly increase the crash likelihood. (2) The signal-timing-related variable, i.e., AOYR_back is proved to have significant positive effect on crash risk, which means that given the same yellow time, more vehicles arrive on yellow could significantly increase the crash likelihood. (3) Three headway-and-occupancy-related variables (Avg_Occupancy_Green_Front, Std_Occupancy_Green_Front, and Std_Headway_Red_Back) are also found to be positively correlated with crash occurrences, which reveals that more congested and fluctuating traffic condition could result in high crash risk. (4) Three traffic-variation-related variables (Diff_OAFR_Green, Diff_Avg_Occupancy_Green, and Diff_Avg_Headway_Green) are found to be significantly associated with crash likelihood. The higher differences in the OAFR and average occupancy during green time between the front and back set of detectors tend to result in higher crash occurrence. However, the higher difference between the average green headway of the front and back set of detectors is proved to be associated with lower crash risk. The possible reason might be that the difference in OAFR and average occupancy are associated with traffic congestion and turbulence between the front and back set of detectors, therefore, higher traffic variation tends to result in higher crash likelihood. However, the difference in average headway represents the vehicle arrival pattern where the higher difference in average headway means that sparser vehicle arrival, which could lead to lower crash risk. (5) Two

shockwave-related variables are recognized to be positively associated with real-time crash risk, which is consistent with previous conflict-based research (Essa and Sayed, 2018a, b).

As suggested by Yuan and Abdel-Aty (2018), the model based on combined time series data may have better performance. Table 6-7 shows the estimation results of the model with combined cycles. Since the variables from different signal cycles are highly correlated, the final model only includes 7 variables after excluding highly correlated and insignificant variables. The performance of the combined model is slightly better than the best cycle model (Cycle-1), however, this improvement is almost negligible which means that the addition of two variables from preceding cycles cannot improve the model performance.

Table 6-7: Estimation Results of Conditional Logistic Model (Combined Cycles).

Variables	Coefficient Estimation		
	Mean	**Std. Error**	**P-value**
Cycle_Volume_1	0.012	0.006	0.054
OAFR_Red_Back_1	0.755	0.289	0.009
Diff_OAFR_Green_1	0.524	0.217	0.016
Diff_Avg_Occupancy_Green_1	0.712	0.231	0.002
Shock_Wave_Area_1	0.144	0.051	0.004
Avg_Occupancy_Green_Front_2	0.632	0.213	0.003
OAFR_Green_Back_3	0.661	0.334	0.048
AUC	0.8094		

To analyze the importance ranking among the variables in the cycle-combined conditional logistic model, an appropriate feature importance measure needs to be selected. Generally, there are two standard random forest feature importance measures, i.e., Gini feature importance and permutation feature importance. As demonstrated in Strobl et al. (2007), Gini feature importance measure is not reliable in situations where potential predictor variables vary in their scale of measurement or their number of categories while permutation feature importance measures are almost unbiased

and more reliable than the Gini feature importance measure. Moreover, Janitza et al. (2013) found that the AUC-based permutation feature importance measure outperforms the standard permutation feature importance measure for imbalanced dataset where the standard permutation feature importance measure loses its ability to discriminate between associated predictors and predictors not associated with the response for increasing class imbalance. Above all, the AUC-based permutation feature importance measure was employed in this study.

Figure 6-10 illustrates the AUC-based permutation feature importance ranking for the cycle-combined conditional logistic model. Among the seven significant factors, the shockwave area during cycle 1 is the most important factor which indicates that the total vehicle delay of all the vehicles during cycle 1 plays the most important role in resulting crash occurrence during the next cycle. Moreover, it's worth noting that the total volume during cycle 1 is the least important factor which is almost expected because the hour of day and day of week were controlled in the matched case-control design. Therefore, the actual effect of total volume can hardly be captured.

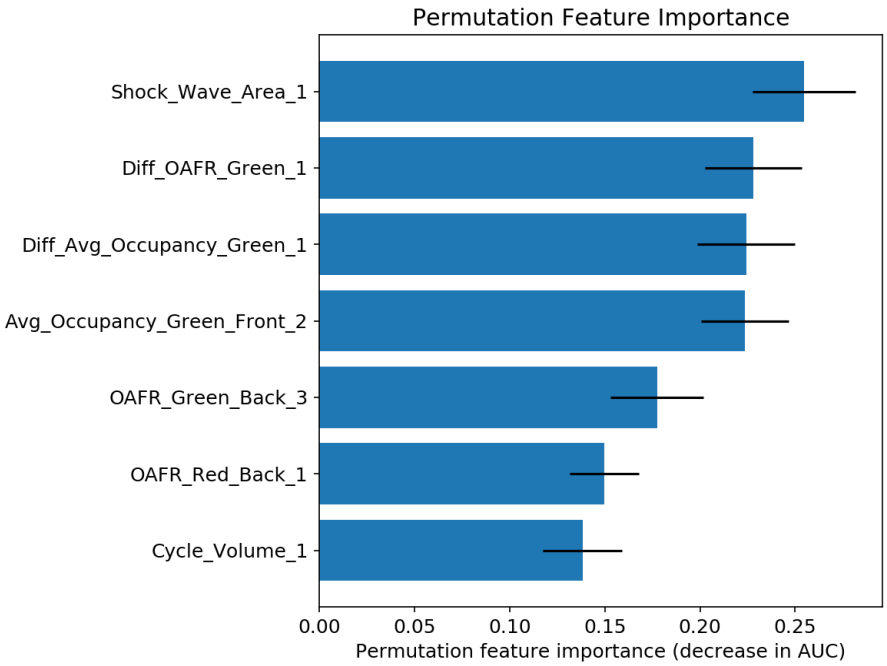


Figure 6-10: Permutation Feature Importance Plot for the Conditional Logistic Model (Matched Case-Control).

To compare the difference between two kinds of undersampling strategies (i.e., matched case-control design and random undersampling) given the same raw imbalanced dataset, two logistic models were developed, respectively. Table 6-8 shows the estimation results of the binary logistic model, which was developed based on the random undersampled dataset. The model comparison results based on the test AUC values reveal that the cycle-3 model performs the best while the cycle-5 model performs the worst. Moreover, every binary logistic model in Table 6-8 outperforms the best conditional logistic model. This could be potentially explained by that the random undersampling method is able to capture the effects of many factors which are controlled in the matched case-control design. For example, the cycle volume, cycle length, green ratio, and queuing shockwave speed might be controlled in the matched case-control design, while they are very significant and important variables in the design of random undersampling.

Table 6-8: Estimation Results of Binary Logistic Model.

Variables	Cycle-1	Cycle-2	Cycle-3	Cycle-4	Cycle-5
	Mean (P-value)	Mean (P-value)	Mean (P-value)	Mean (P-value)	Mean (P-value)
Intercept	-3.072 (<0.001) **	-1.147 (<0.001) **	-2.068 (<0.001) **	-2.03 (<0.001) **	-1.673 (<0.001) **
Cycle_Volume	0.021 (<0.001) **	0.023 (<0.001) **	0.02 (<0.001) **	0.022 (<0.001) **	0.02 (<0.001) **
OAFR_Red_Back	0.752 (0.003) **	-	-	-	-
Cycle_Len	0.005 (0.021) **	-	0.005 (0.024) **	0.005 (0.014) **	0.005 (0.006) **
Green_Ratio	-2.54 (<0.001) **	-2.958 (<0.001) **	-2.777 (<0.001) **	-3.339 (<0.001) **	-3.325 (<0.001) **
Avg_Headway_Green_Back	-	-0.011 (0.038) **	-0.009 (0.059) *	-	-
Avg_Headway_Green_Front	-0.01 (0.062) *	-	-	-	-0.013 (0.013) **
Std_Headway_Red_Back	-	-	0.005 (0.088) *	-	0.006 (0.037) **
Avg_Occupancy_Green_Front	0.419 (0.005) **	-	-	-	-
Std_Occupancy_Green_Front	-	0.348 (0.005) **	0.623 (<0.001) **	-	0.435 (0.002) **
Std_Occupancy_Green_Back	-	-	0.446 (0.024) **	0.401 (0.026) **	-
Diff_Avg_Occupancy_Green	0.422 (0.039) **	-	-	0.468 (0.076) *	-
Queuing_Shockwave_Spd	-0.091 (0.008) **	-0.115 (0.002) **	-0.084 (0.013) **	-0.117 (<0.001) **	-0.127 (<0.001) **
AUC	0.8421	0.862	0.8853	0.8811	0.8348

*Note: The cells noted by ** are significant at the 0.05 level; The cells noted by * are significant at the 0.1 level.*

In total, there are 12 significant variables among the five binary logistic models. (1) Two volume-related variables (Cycle_Volume and OAFR_Red_Back) are found to have positive effects on crash occurrence. Higher cycle volume and OAFR could result are proved to be crash-prone conditions. (2) Two signal-timing-related variables (Cycle_Len and Green_Ratio) are found to be significantly associated with cycle-level crash risk, which implies that longer cycle length and lower green ratio could significantly increase the crash likelihood. The possible reason might be that the longer cycle length and lower green ratio may result in longer waiting time for those vehicles who arrive on red, which could significantly increase the crash risk (Yuan and Abdel-Aty, 2018). (3) Six headway-and-occupancy-related variables (Avg_Headway_Green_Back, Avg_Headway_Green_Front, Std_Headway_Red_Back, Avg_Occupancy_Green_Front, Std_Occupancy_Green_Front, Std_Occupancy_Green_Back) are found to be significant. The

average headways during green time from both front and back sets of detectors are uncovered to be negatively correlated with crash occurrence while the standard deviation of headway is proved to have positive effect on crash likelihood. These findings reveal that sparser and more uniform vehicle arrivals could significantly decrease the crash risk. In terms of occupancy, both the average and standard deviation of occupancy are found to be positively associated with crash occurrences, which in turn verified the findings from headway-related variables that more congested and volatile traffic flow may result in higher crash risk. (4) The traffic-variation-related variable (Diff_Avg_Occupancy_Green) is recognized to be positively correlated with crash occurrences, which could also be explained by that the higher traffic volatility could significantly increase the crash likelihood. This finding is in line with previous research on the safety effect of traffic volatility (Wali et al., 2018). (5) The Queuing_Shockwave_Spd is found to be negatively correlated with crash occurrences, which means that higher queuing shockwave speed may lead to lower crash risk. It is worth noting that the queuing shockwave speed is always negative, and the higher value of queuing shockwave speed stands for the lower absolute value of queuing shockwave speed. Therefore, slow absolute queuing shockwave speed appears to be associated with lower crash likelihood.

Meanwhile, the combined model was also developed for the random undersampled dataset. Table 6-9 presents the estimation results of the binary logistic model with combined cycles. The model performance is almost at the same level as the best cycle model in Table 6-8. In total, 10 variables from four cycles are found to be significant. Generally, the logical signs of all the variables are consistent with the aforementioned models.

Table 6-9: Estimation Results of Binary Logistic Model (Combined Cycles).

Variables	Coefficient Estimation		
	Mean	**Std. Error**	**P-value**
(Intercept)	-3.373	0.555	<0.001**
Cycle_Volume_1	0.021	0.003	<0.001**
Diff_Avg_Occupancy_Green_1	0.651	0.224	0.004**
OAFR_Red_Back_1	0.841	0.263	0.001**
Green_Ratio_2	-2.29	0.677	0.001**
Std_Occupancy_Green_Front_3	0.568	0.177	0.001**
Std_Occupancy_Green_Back_3	0.551	0.214	0.01**
Queuing_Shockwave_Spd_3	-0.062	0.035	0.076*
Cycle_Len_3	0.003	0.002	0.095*
Avg_Headway_Red_Back_4	-0.009	0.005	0.086*
Shock_Wave_Area_4	0.114	0.055	0.039**
AUC	0.886		

*Note: The cells noted by ** are significant at the 0.05 level; The cells noted by * are significant at the 0.1 level.*

Figure 6-11 shows the AUC-based permutation feature importance ranking for the cycle-combined binary logistic model. It can be clearly observed that the shockwave area is the most important feature, which is consistent with the conditional logistic model. In the binary logistic model, however, the factor of shockwave area was collected from cycle 4 while the factor of shockwave area in the conditional logistic model was collected from cycle 1. The possible reason for the difference in cycles might be that the random undersampling method does not control the effects of time of day and day of week, and effects of factors over time is much more dispersed than the matched case-control method which has been verified by the difference between the AUCs of different cycle models. Another notable finding is that the cycle volume is the second most important factor in the binary logistic model while it is the least important factor in the conditional logistic model, which might be the most important reason why the binary logistic models outperform the conditional logistic models.

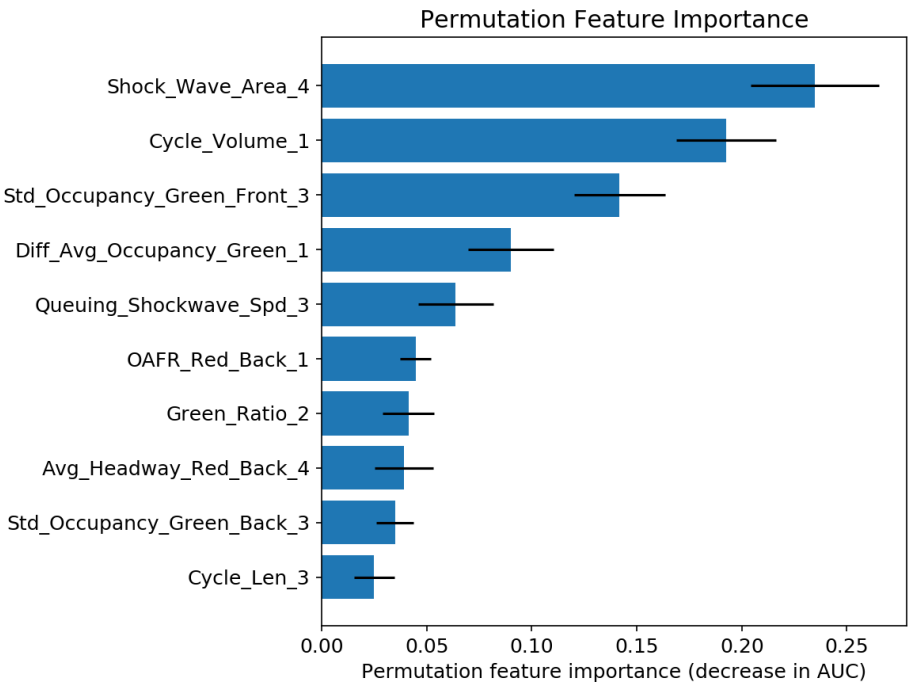


Figure 6-11: Permutation Feature Importance Plot for the Binary Logistic Model (Random Undersampling).

6.4.2 Classification Evaluation

Receiver operating characteristic (ROC) curve is widely used to illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Figure 6-12 shows the ROC curves for the final conditional logistic model and binary logistic model. As indicated in the figure, the area under ROC curve of the binary logistic model is 0.886, which is much higher than that of the conditional logistic model.

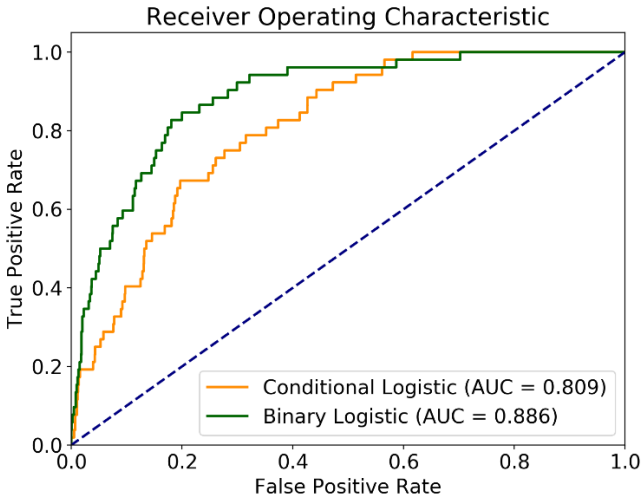


Figure 6-12: Receiver Operating Characteristics Curve.

To evaluate the model performance in terms of the specific sensitivity and false positive rate, the cut-off threshold needs to be determined. In this study, the cut-off threshold was determined as the optimal point where sensitivity and specificity curves cross (Shi and Abdel-Aty, 2015; Yuan et al., 2019). Table 6-10 shows the prediction outcomes, sensitivities, and false positive rates for the two models on the test dataset. The sensitivity of the binary logistic model is much higher than the conditional logistic model where the binary logistic model successfully predicted 43 crashes while the conditional logistic model only successfully predicted 38 crashes over the total 52 crashes.

Table 6-10: Model Classification Results on Test Dataset

Observed	**Predicted**			
	Conditional Logistic		**Binary Logistic**	
	Non-Crash Event	Crash Event	Non-Crash Event	Crash Event
Non-Crash Event	1,798,294	662,509	2,016,496	444,307
Crash Event	14	38	9	43
Sensitivity	0.731		0.827	
False Positive Rate	0.269		0.181	

In terms of the impacts of crash characteristics on model prediction performance, Table 6-11 shows the model prediction performance among different crash types, crash severities, light conditions, and time periods. More specifically, the prediction sensitivity of sideswipe crashes is much higher than the rear-end crashes in both models, and the strength of binary logistic model mainly relies on rear-end crashes where the binary logistic model can predict four more rear-end crashes than the conditional logistic model. In terms of crash severity, property damage only (PDO) crashes are more likely to be predicted than injury crashes according to both models. Moreover, the model prediction performance for daylight crashes is much higher than dusk and dark-lighted conditions, the binary logistic model can even predict more than 90% daylight crashes. With respect to time periods, both models perform better during peak condition than that during non-peak condition. It is worth noting that the superiority of the binary logistic model over the conditional logistic model during non-peak condition is more significant than that during peak condition. This could be attributed to the fact that the binary logistic model based on random undersampled dataset could capture the effects of traffic exposure-related factors in addition to the intrinsic traffic fluctuation, while those exposure factors can hardly be considered by the conditional logistic model as they are mainly controlled as confounding factors. In addition, traffic exposure is supposed to have a more important role in the crash risk during non-peak condition than that during peak condition .

Table 6-11: Comparison between the Model Performance for Different Types of Crashes

Observed Test Crash Events		Conditional Logistic			Binary Logistic		
		Predicted		Sensitivity	Predicted		Sensitivity
		0	1		0	1	
Crash Type	Rear-End	12	29	0.707	8	33	0.805
	Sideswipe	1	8	**0.889**	1	8	**0.889**
	Right Turn	1	0	0.000	0	1	1.000
	Other	0	1	1.000	0	1	1.000
Crash Severity	Injury	3	7	0.700	2	8	0.800
	Property Damage Only (PDO)	11	31	0.738	7	35	0.833
Light Condition	Daylight	10	31	0.756	4	37	**0.902**
	Dusk	1	1	0.500	1	1	0.500
	Dark - Lighted	3	5	0.625	4	4	0.500
	Dark - Not Lighted	0	1	1.000	0	1	1.000
Peak/Non-Peak	Peak	4	15	0.789	3	16	0.842
	Non-Peak	10	23	**0.697**	6	27	**0.818**

6.5 Conclusion and Discussion

This study aims to reveal the relationships between real-time crash occurrences and cycle-level characteristics at signalized intersection approaches. 42 intersection approaches in Seminole County were selected and the high-resolution ATSPM database was utilized to collect real-time cycle-level signal timing, lane-specific volume, headway, and occupancy related variables. Moreover, cycle-level shockwave characteristics, including maximum queue length, shock wave area, and queuing shockwave speed, were also collected from ATSPM database. Prior to the modeling process, the actual times of crash events were determined based on abnormal vehicle detections from ATSPM data. To consider the effect of time dependency, five preceding cycles

were considered to examine their relationships with the crash occurrences during the current cycle. In terms of the model framework, the imbalanced raw data (ratio of crash cycle to non-crash cycle is 1: 48,954) were collected for nearly two years and then split into approximately 80% training data (before 9/16/2018) and 20% test data (after 9/16/2018). For the training dataset, both matched case-control and random undersampling techniques were employed, and conditional logistic and binary logistic models were calibrated respectively to investigate the difference between various undersampling techniques as well as the corresponding statistical models.

Model results reveal that the binary logistic model based on the random undersampled dataset performs much better than the conditional logistic model based on the matched case-control dataset. This could be attributed to that the binary logistic model based on the random undersampled dataset is able to capture the effects of traffic exposure-related factors in addition to the intrinsic traffic fluctuation, while those exposure factors can hardly be considered by the conditional logistic model as they are mainly controlled as confounding factors. This has also been verified through the permutation feature importance figures (Figure 6-10 and Figure 6-11) where the cycle volume is the second most important variable in the binary logistic model while it is the least important variable in the conditional logistic model.

In terms of the time dependency, among the five conditional logistic cycle models, there is a significant trend that the closest preceding cycle (cycle 1) outperforms the other cycle models. However, for the five binary logistic cycle models, there are no significant differences between different cycles, which could be explained by the difference between two undersampling strategies. More specifically, the matched case-control design mainly aims to capture the effects of intrinsic traffic fluctuation rather than the controlled factors which are also very important factors. Therefore, the matched case-control design could better capture the differences between the five

preceding cycles. On the other hand, the random undersampling method aims to model all the potential contributing factors, including both traffic exposure and intrinsic fluctuation characteristics. However, those exposure factors (e.g., cycle volume) are very likely to be similar among five consecutive cycles, which might lead to similar model performance among five cycle models.

Overall, there are five groups of variables (i.e., traffic volume, signal timing, headway and occupancy, traffic variation, and shockwave characteristics) found be significantly associated with the cycle-level crash risk at signalized intersections. (1) Higher cycle volume and overall average flow ratio across lanes could significantly increase the crash likelihood at signalized intersections, which is in line with previous studies (Essa and Sayed, 2018a, b). (2) Longer cycle length, higher arrivals on yellow ratio, and lower green ratio tend to increase the crash risk, which is also consistent with our previous research (Yuan and Abdel-Aty, 2018; Yuan et al., 2019). (3) More congested (higher average occupancy and lower average headway) and fluctuating (higher standard deviation of vehicle occupancy and headway) traffic flow is more likely to be crash-prone conditions. (4) Higher traffic volatility across approach sections could significantly increase the crash likelihood, which is similar to the aggregated intersection safety research (Kamrani et al., 2018). (5) Three shockwave-related variables are found to have significant effects on the cycle-level crash risk. Longer maximum queue length, larger shockwave area, and higher absolute queuing shockwave speed are proved to be crash-prone conditions, which consistent with previous conflict-based research (Essa and Sayed, 2018a, b).

With respect to the model classification performance on the test dataset, the model results indicate that the prediction sensitivity of sideswipe crashes is much higher than the rear-end crashes. Also, PDO crashes, as well as those crashes occurred during daylight and peak conditions are more likely

to be predicted. In terms of model comparison, it is worth noting that the binary logistic model is found to have superior performance on rear-end crashes, as well as those crashes happened during non-peak and daylight conditions, while the conditional logistic model performs better on those crashes occurred during the dark-lighted condition. These findings inspire us that ensembled classifiers could be considered in the future to achieve better prediction performance.

Above all, this is the first attempt to investigate the cycle-level crash risk at signalized intersections based on high-resolution event-based data. Even though the model performance is very promising, there are still some limitations and possible improvements could be made in the future. (1) Only five preceding signal cycles and the crash to non-crash ratio of 1:4 are considered in this study. More cycles and various crash to non-crash ratios could be considered, or event sensitivity analyses could be conducted in the future. (2) While estimating the shockwave characteristics, there is an assumption that the breakpoint C could be identified. However, there might exist oversaturation conditions, e.g., extreme congestion where the intersection is blocked by the downstream queue, therefore, the breakpoint C cannot be identified. (3) It is also worth pointing out that the spatial relationships between upstream and downstream intersections have not been considered in this study, which might be very important in improving the model prediction performance. In this context, more advanced spatial-temporal modeling techniques, e.g., Conv_LSTM (convolutional long short-term memory), could be employed in future research.

CHAPTER 7: CONCLUSIONS

7.1 Summary

This dissertation aims to investigate the relationship between the real-time crash risk on arterials and all the possible contributing factors, and then improve the model prediction performance by employing deep learning algorithms, sampling techniques, and high-resolution data. More specifically, the main objectives of this dissertation are to (1) reveal the relationship between real-time crash occurrence and real-time traffic and signal characteristics on arterials, (2) investigate the effects of real-time traffic, signal timing, and weather characteristics on intersection approach-level crash likelihood, (3) develop a real-time crash risk prediction algorithm for signalized intersections by integrating LSTM and oversampling techniques, (4) predict real-time crash risk at the cycle-level for signalized intersections with the consideration of shockwave characteristics based on high-resolution data.

In Chapter 3, this study investigated the crash risk on urban arterials based on real-time data from multiple sources, including travel speed provided by Bluetooth detectors, traffic volume and signal timing extracted from adaptive signal controllers, and weather data collected by the airport weather station. Matched case-control design with a control-to-case ratio of 4:1 was applied to collect data for crash and non-crash events. Four BCL models were developed separately for four 5-minute interval datasets (20-minute window prior to the reported crash time). In terms of AUC values, the model estimation results indicated that slice 2 (5-10 minute) model performs the best, followed by the slice 1 (0-5 minute) model. It revealed that the average speed, upstream left-turn volume, downstream green ratio, and rainy indicator were found to have significant effects on crash occurrence. Furthermore, Bayesian random parameters conditional logistic model (BRPCL) outperformed Bayesian random parameters logistic (BRPL) and Bayesian conditional logistic

models (BCL) in terms of the area under the receiver operating characteristics curve (AUC) and Deviance Information Criterion (DIC) values.

In Chapter 4, this research examined the real-time crash risk at signalized intersections based on the disaggregated data from multiple sources, including travel speed collected by Bluetooth detectors, lane-specific traffic volume and signal timing data from adaptive signal controllers, and weather data collected by airport weather station. The intersection and intersection-related crashes were collected and then divided into three types, i.e., within intersection crashes, intersection entrance crashes, and intersection exit crashes. In terms of the sample size, only the within intersection crashes and intersection entrance crashes were considered and then modeled separately. Matched case-control design with a control-to-case ratio of 4:1 was employed to select the corresponding non-crash events for each crash event. Afterwards, all the traffic, signal timing, and weather characteristics during 20-minute window prior to the crash or non-crash events were collected and divided into four 5-minute slices. Later, Bayesian conditional logistic models were developed for within intersection crashes and intersection entrance crashes, respectively.

In Chapter 5, this study predicted the real-time crash risk at signalized intersections by using multilayer LSTM recurrent neural network, which is designed for sequence modeling, and they can consider the time series characteristics automatically. First, a real-world unbalanced dataset was collected for every minute by incorporating real-time traffic, signal, and weather data. Also, both the approach-level and intersection-level geometric characteristics were included into the algorithm. To train the algorithm without losing any non-crash information, the synthetic minority over-sampling technique (SMOTE) was employed in this study to generate a balanced training dataset. In comparison, a traditional conditional logistic model was developed based on the matched case-control dataset with the control-to-case ratio of 10:1. The prediction results showed

that the LSTM with SMOTE could predicts 60.67% of the intersection crashes with a false alarm rate of 39.33%, which is better than the conditional logistic model (i.e., sensitivity: 56.72% and false alarm rate: 43.28%). This comparison results succeed in verifying the feasibility of applying LSTM in real-time crash risk prediction. Since this study is the first attempt in predicting real-time crash risk by using LSTM, therefore, the feasibility proof of the of LSTM with SMOTE is the major objective of this study.

In Chapter 6, this study aims to reveal the relationship between real-time crash occurrences and cycle-level characteristics at signalized intersection approaches. Forty-two intersection approaches in Seminole County were selected and the high-resolution ATSPM database was utilized to calculate real-time cycle-level signal timing, lane-specific volume, headway, and occupancy related variables. Moreover, cycle-level shockwave characteristics, including maximum queue length, shock wave area, and queuing shockwave speed, were also estimated from ATSPM database. Prior to the modeling process, the actual times of crash events were determined based on abnormal vehicle detections from ATSPM data. To consider the effect of time dependency, five preceding cycles were considered to examine their relationships with the crash occurrences during the current cycle. In terms of the model framework, the imbalanced raw data (ratio of crash cycle to non-crash cycle is 1: 48,954) were collected for nearly two years and then split into approximately 80% training data (before 9/16/2018) and 20% test data (after 9/16/2018). For the training dataset, both matched case-control and random undersampling techniques were employed, and conditional logistic and binary logistic models were calibrated respectively to investigate the difference between various undersampling techniques as well as the corresponding statistical models. Model results reveal that the binary logistic model based on the random undersampled dataset performs much better than the conditional logistic model based on the matched case-control

dataset. This could be attributed to that the binary logistic model based on the random undersampled dataset is able to capture the effects of traffic exposure-related factors in addition to the intrinsic traffic fluctuation, while those exposure factors can hardly be considered by the conditional logistic model as they are more likely to be controlled when we are controlling the confounding factors (i.e., intersection approach, hour of day, and day of week).

7.2 Implications

Chapter 3 presents multiple algorithms on predicting the real-time crash risk on arterial segments. The outcome of this study might be implemented on urban arterials from several aspects. The most straightforward application is to apply this algorithm to develop an arterial real-time crash risk prediction system. The real-time prediction results could be fed into the implementation of proactive traffic management strategies (e.g., variable speed limit or queue warning), which can efficiently mitigate the crash risk in advance of the potential crash occurrence. Also, the real-time prediction results could be provided to drivers to assist with the route choice decisions. Furthermore, the real-time crash prediction results could be delivered to the drivers through connected-vehicle technology to provide crash risk warning information (Yue et al., 2018). In addition, the arterial real-time crash risk prediction system could be integrated with the real-time crash prediction on freeways. Therefore, an integrated arterial/freeway active traffic management strategy could be employed to proactively mitigate the safety of the road network.

Chapter 4 developed Bayesian conditional logistic models for within intersection crashes and intersection entrance crashes. For the within intersection models, the model results showed that the through volume from “A” approach (the traveling approach of at-fault vehicle), the left turn

volume from “B” approach (near-side crossing approach), and the overall average flow ratio (OAFR) from “D” approach (far-side crossing approach), were found to have significant positive effects on the odds of crash occurrence. Moreover, the increased adaptability for the left turn signal timing of “B” approach and more priority for “A” approach could significantly decrease the odds of crash occurrence. For the intersection entrance models, average speed was found to have significant negative effect on the odds of crash occurrence. The longer average green time and longer average waiting time for the left turn phase, higher green ratio for the through phase, and higher adaptability for the through phase can significantly improve the safety performance of intersection entrance area. In addition, the average queue length on the through lanes was found to have positive effect on the odds of crash occurrence.

Chapter 5 succeeded in verifying the feasibility of real-time crash risk prediction at signalized intersections by using LSTM recurrent neural network together with SMOTE over-sampling method. The results of this study could be utilized to predict real-time crash risk at signalized intersections in advance, which could assist operators to implement various pro-active traffic management strategies to reduce the risk in real-time (for example using adaptive signal control).

Chapter 6 unveiled that the binary logistic model based on the random undersampled dataset performs much better than the conditional logistic model based on the matched case-control dataset. Among the model results, there are five groups of variables (i.e., traffic volume, signal timing, headway and occupancy, traffic variation, and shockwave characteristics) found be significantly associated with the cycle-level crash risk at signalized intersections. Higher cycle volume and overall average flow ratio across lanes could significantly increase the crash likelihood at signalized intersections. Also, longer cycle length, higher arrivals on yellow ratio, and lower green ratio tend to increase the crash risk. More congested (higher average occupancy and lower average

headway) and fluctuating (higher standard deviation of vehicle occupancy and headway) traffic flow is more likely to be crash-prone conditions. Higher traffic volatility across approach sections could significantly increase the crash likelihood. Longer maximum queue length, larger shockwave area, and higher absolute queuing shockwave speed are proved to be crash-prone conditions. These findings inspire us that proactive traffic management strategies, e.g., adaptive signal control, could be implemented to alleviate the real-time crash risk at signalized intersections.

REFERENCES

Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention* 38(2), 215-224.

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. *Journal of safety Research* 36(1), 97-108.

Abdel-Aty, M., Pande, A., Das, A., Knibbe, W., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. *Transportation Research Record: Journal of the Transportation Research Board*(2083), 153-161.

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., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 98-111.

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. A data fusion framework for real-time risk assessment on freeways. *Transportation Research Part C: Emerging Technologies* 26, 203-213.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012a. Assessment of Interaction of Crash Occurrence, Mountainous Freeway Geometry, Real-Time Weather, and Traffic Data. *Transportation Research Record: Journal of the Transportation Research Board* 2280, 51-59.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012b. 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.A., 2012. The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction. *IEEE Transactions on Intelligent Transportation Systems* 13(2), 459-468.

Albanese, D., Riccadonna, S., Donati, C., Franceschi, P., 2018. A practical tool for maximal information coefficient analysis. *GigaScience* 7(4), giy032.

Bagdadi, O., 2013. Assessing safety critical braking events in naturalistic driving studies. *Transportation Research Part F: Traffic Psychology and Behaviour* 16, 117-126.

Baruya, A., 1998. Speed-accident relationships on European roads, *9th International Conference on Road Safety in Europe*, Bergisch Gladbach.

Basso, F., Basso, L.J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. *Transportation Research Part C: Emerging Technologies* 86, 202-219.

Bonneson, J., McCoy, P., 1997. Effect of median treatment on urban arterial safety an accident prediction model. *Transportation Research Record: Journal of the Transportation Research Board*(1581), 27-36.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* 7(4), 434-455.

Cai, Q., Abdel-Aty, M., Lee, J., Huang, H., 2018a. Integrating macro-and micro-level safety analyses: a Bayesian approach incorporating spatial interaction. *Transportmetrica A: Transport Science*, 1-22.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018b. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection

crash modeling. *Analytic Methods in Accident Research* 19, 1-15.

Cai, Q., Wang, Z., Zheng, L., Wu, B., Wang, Y., 2014. Shock wave approach for estimating queue length at signalized intersections by fusing data from point and mobile sensors. *Transportation Research Record: Journal of the Transportation Research Board*(2422), 79-87.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, 321-357.

Chen, M., Chien, S., 2000. Determining the number of probe vehicles for freeway travel time estimation by microscopic simulation. *Transportation Research Record: Journal of the Transportation Research Board*(1719), 61-68.

Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention* 35(2), 253-259.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Chung, W., Abdel-Aty, M., Lee, J., 2018. Spatial analysis of the effective coverage of land-based weather stations for traffic crashes. *Applied geography* 90, 17-27.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis & Prevention* 70, 320-329.

Eboli, L., Mazzulla, G., Pungillo, G., 2016. Combining speed and acceleration to define car users' safe or unsafe driving behaviour. *Transportation Research Part C: Emerging Technologies* 68, 113-125.

Ekram, A.-A., Rahman, M.S., 2018. Effects of Connected and Autonomous Vehicles on Contraflow Operations for Emergency Evacuation: a Microsimulation Study.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. *Accident Analysis & Prevention* 41(5), 1118-1123.

Elvik, R., 2009. *The Power Model of the relationship between speed and road safety: update and new analyses*. Institute of Transport Economics, Oslo.

Essa, M., Sayed, T., 2018a. Full Bayesian conflict-based models for real time safety evaluation of signalized intersections. *Accident Analysis & Prevention*.

Essa, M., Sayed, T., 2018b. Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. *Transportation Research Part C: Emerging Technologies* 89, 289-302.

Golob, T.F., Recker, W.W., Alvarez, V.M., 2004. Freeway safety as a function of traffic flow. *Accident Analysis & Prevention* 36(6), 933-946.

Gomes, S.V., 2013. The influence of the infrastructure characteristics in urban road accidents occurrence. *Accident Analysis & Prevention* 60, 289-297.

Gong, Y., Abdel-Aty, M., Cai, Q., Rahman, M.S., 2019a. Decentralized network level adaptive signal control by multi-agent deep reinforcement learning. *Transportation Research Interdisciplinary Perspectives* 1, 100020.

Gong, Y., Abdel-Aty, M., Park, J., 2019b. Evaluation and augmentation of traffic data including Bluetooth detection system on arterials. *Journal of Intelligent Transportation Systems*, 1-13.

Graves, A., Mohamed, A.-r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks, *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, pp. 6645-6649.

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

Guo, F., Wang, X., Abdel-Aty, M.A., 2010. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention* 42(1), 84-92.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation* 9(8), 1735-1780.

Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2013. *Applied logistic regression*. John Wiley & Sons, Hoboken, New Jersey.

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.

Imprialou, M.-I.M., Quddus, M., Pitfield, D.E., Lord, D., 2016. Re-visiting crash–speed relationships: A new perspective in crash modelling. *Accident Analysis & Prevention* 86, 173-185.

Imprialou, M., Quddus, M., 2017. Crash data quality for road safety research: current state and future directions. *Accident Analysis & Prevention*.

Janitza, S., Strobl, C., Boulesteix, A.-L., 2013. An AUC-based permutation variable importance measure for random forests. *BMC bioinformatics* 14(1), 119.

Jun, J., Guensler, R., Ogle, J., 2011. Differences in observed speed patterns between crash-involved and crash-not-involved drivers: Application of in-vehicle monitoring technology. *Transportation Research Part C: Emerging Technologies* 19(4), 569-578.

Kalchbrenner, N., Blunsom, P., 2013. Recurrent continuous translation models, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700-1709.

Kamrani, M., Arvin, R., Khattak, A.J., 2018. Extracting useful information from Basic Safety Message Data: an empirical study of driving volatility measures and crash frequency at intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 0361198118773869.

Khattak, Z.H., Fontaine, M.D., Boateng, R.A., 2018a. Evaluating the Impact of Adaptive Signal Control Technology on Driver Stress and Behavior, *Transportation Research Board 97th Annual MeetingTransportation Research Board*, Washington D.C.

Khattak, Z.H., Magalotti, M.J., Fontaine, M.D., 2018b. Estimating safety effects of adaptive signal control technology using the Empirical Bayes method. *Journal of Safety Research* 64, 121-128.

King, G., Zeng, L., 2001. Logistic regression in rare events data. *Political Analysis* 9(2), 137-163.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv*.

Kobelo, D., Patrangenaru, V., Mussa, R., 2008. Safety analysis of Florida urban limited access highways with special focus on the influence of truck lane restriction policy. *Journal of Transportation Engineering* 134(7), 297-306.

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.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. *Accident Analysis & Prevention* 102, 213-226.

Lee, J., Park, B.B., Malakorn, K., So, J.J., 2013. Sustainability assessments of cooperative vehicle intersection control at an urban corridor. *Transportation Research Part C: Emerging Technologies* 32, 193-206.

Liu, H.X., Wu, X., Ma, W., Hu, H., 2009. Real-time queue length estimation for congested signalized intersections. *Transportation research part C: emerging technologies* 17(4), 412-427.

Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G.J.I.t.o.p.a., intelligence, m., 2018. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. 40(12), 3007-3021.

Long Cheu, R., Xie, C., Lee, D.H., 2002a. Probe vehicle population and sample size for arterial speed estimation. *Computer-Aided Civil Infrastructure Engineering* 17(1), 53-60.

Long Cheu, R., Xie, C., Lee, D.H., 2002b. Probe vehicle population and sample size for arterial speed estimation. *Computer-Aided Civil and Infrastructure Engineering* 17(1), 53-60.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing* 10(4), 325-337.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies* 54, 187-197.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S., 2010. Recurrent neural network based language model, *Eleventh Annual Conference of the International Speech Communication Association*.

Mussone, L., Bassani, M., Masci, P., 2017. Analysis of factors affecting the severity of crashes in urban road intersections. *Accident; analysis and prevention* 103, 112-122.

Nilsson, G., 2004. Traffic safety dimensions and the power model to describe the effect of speed on safety. *Bulletin-Lunds Tekniska Högskola, Inst för Teknik och Samhälle, Lunds Universitet* 221.

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

Pei, X., Wong, S., Sze, N.-N., 2012. The roles of exposure and speed in road safety analysis. *Accident Analysis & Prevention* 48, 464-471.

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. *Journal of transportation engineering* 122(2), 105-113.

Rahman, M.S., Abdel-Aty, M., 2018. Longitudinal safety evaluation of connected vehicles' platooning on expressways. *Accident Analysis & Prevention* 117, 381-391.

Rahman, M.S., Abdel-Aty, M., Lee, J., Rahman, M.H., 2019. Safety benefits of arterials' crash risk under connected and automated vehicles. *Transportation Research Part C: Emerging Technologies* 100, 354-371.

Rahman, M.S., Abdel-Aty, M., Wang, L., Lee, J., 2018. Understanding the Highway Safety Benefits of Different Approaches of Connected Vehicles in Reduced Visibility Conditions. *Transportation Research Record*, 0361198118776113.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* 12(1), 77.

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.

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.

Simons-Morton, B.G., Cheon, K., Guo, F., Albert, P., 2013. Trajectories of kinematic risky driving among novice teenagers. *Accident Analysis & Prevention* 51, 27-32.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2003. *WinBUGS user manual*. MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4), 583-639.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1), 1929-1958.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* 8(1), 25.

Stuster, J., 2004. *Aggressive Driving Enforcement: Evaluations of Two Demonstration Programs*. US Department of Transportation, National Highway Traffic Safety Administration.

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.

Taylor, M.C., Baruya, A., Kennedy, J.V., 2002. *The relationship between speed and accidents on rural single-carriageway roads*. TRL.

Theofilatos, A., 2017. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. *Journal of Safety Research* 61, 9-21.

Theofilatos, A., Yannis, G., Kopelias, P., Papadimitriou, F., 2018a. Impact of real-time traffic characteristics on crash occurrence: preliminary results of the case of rare events. *Accident Analysis & Prevention*.

Theofilatos, A., Yannis, G., Kopelias, P., Papadimitriou, F., 2018b. Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident; analysis and prevention*.

Theofilatos, A., Yannis, G., Vlahogianni, E.I., Golias, J.C., 2017. Modeling the effect of traffic regimes on safety of urban arterials: The case study of Athens. *Journal of Traffic and Transportation Engineering (English Edition)* 4(3), 240-251.

Wali, B., Khattak, A.J., Bozdogan, H., Kamrani, M., 2018. How is driving volatility related to intersection safety? A Bayesian heterogeneity-based analysis of instrumented vehicles data. *Transportation Research Part C: Emerging Technologies* 92, 504-524.

Wang, C., Quddus, M.A., Ison, S.G., 2013. The effect of traffic and road characteristics on road safety: A review and future research direction. *Safety science* 57, 264-275.

Wang, L., Abdel-Aty, M., Lee, J., 2017a. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. *Accident Analysis & Prevention* 104, 58-64.

Wang, L., Abdel-Aty, M., Lee, J., Shi, Q., 2019a. Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. *Accident Analysis & Prevention* 122, 378-384.

Wang, L., Abdel-Aty, M., Ma, W., Hu, J., Zhong, H., 2019b. Quasi-vehicle-trajectory-based real-time safety analysis for expressways. *Transportation Research Part C: Emerging*

Technologies 103, 30-38.

Wang, L., Abdel-Aty, M., Shi, Q., Park, J., 2015a. Real-time crash prediction for expressway weaving segments. *Transportation Research Part C: Emerging Technologies* 61, 1-10.

Wang, X., Abdel-Aty, M., Almonte, A., Darwiche, A., 2009. Incorporating traffic operation measures in safety analysis at signalized intersections. *Transportation Research Record: Journal of the Transportation Research Board*(2103), 98-107.

Wang, X., Abdel-Aty, M., Brady, P., 2006. Crash estimation at signalized intersections: significant factors and temporal effect. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 10-20.

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

Wang, X., Khattak, A.J., Liu, J., Masghati-Amoli, G., Son, S., 2015c. What is the level of volatility in instantaneous driving decisions? *Transportation Research Part C: Emerging Technologies* 58, 413-427.

Wang, X., Yuan, J., 2017. Safety Impacts Study of Roadway Network Features on Suburban Highways. *China Journal of Highway and Transport* 30(4), 106-114.

Wang, X., Yuan, J., Schultz, G., Meng, W., 2016a. Investigating Safety Impacts of Roadway Network Features of Suburban Arterials in Shanghai, China, *95th Annual Meeting of the Transportation Research Board*. Transportation Research Board, Washington, D.C.

Wang, X., Yuan, J., Schultz, G.G., Fang, S., 2018. Investigating the safety impact of roadway network features of suburban arterials in Shanghai. *Accident Analysis & Prevention* 113, 137-148.

Wang, X., Yuan, J., Yang, X., 2016b. Modeling of Crash Types at Signalized Intersections Based

on Random Effect Model. *Journal of Tongji University (Natural Science)* 44(1), 81-86.

Wang, Z., Cai, Q., Wu, B., Zheng, L., Wang, Y., 2017b. Shockwave-based queue estimation approach for undersaturated and oversaturated signalized intersections using multi-source detection data. *Journal of Intelligent Transportation Systems* 21(3), 167-178.

Wijnands, J.S., Thompson, J., Aschwanden, G.D., Stevenson, M., 2018. Identifying behavioural change among drivers using Long Short-Term Memory recurrent neural networks. *Transportation Research Part F: Traffic Psychology and Behaviour* 53, 34-49.

Wu, X., Liu, H.X., 2014. Using high-resolution event-based data for traffic modeling and control: An overview. *Transportation research part C: emerging technologies* 42, 28-43.

Wu, Y., Abdel-Aty, M., Wang, L., Rahman, M.S., 2019. Combined connected vehicles and variable speed limit strategies to reduce rear-end crash risk under fog conditions. *Journal of Intelligent Transportation Systems*, 1-20.

Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. *Accident Analysis & Prevention* 50, 25-33.

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.

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.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013a. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention* 57, 30-39.

Xu, C., Wang, W., Liu, P., 2013b. Identifying crash-prone traffic conditions under different weather

on freeways. *Journal of Safety Research* 46, 135-144.

Xu, J., Rahmatizadeh, R., Bölöni, L., Turgut, D., 2017. Real-Time Prediction of Taxi Demand Using Recurrent Neural Networks. *IEEE Transactions on Intelligent Transportation Systems*.

Yanmaz-Tuzel, O., Ozbay, K., 2010. A comparative Full Bayesian before-and-after analysis and application to urban road safety countermeasures in New Jersey. *Accident Analysis & Prevention* 42(6), 2099-2107.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. *Accident; analysis and prevention* 51, 252-259.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science* 63, 50-56.

Yu, R., Abdel-Aty, M.A., 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. *IEEE Transactions On Intelligent Transportation Systems* 15(1), 205-213.

Yu, R., Quddus, M., Wang, X., Yang, K., 2018. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. *Accident Analysis & Prevention* 120, 304-310.

Yu, R., Wang, X., Abdel-Aty, M., 2017. A Hybrid Latent Class Analysis Modeling Approach to Analyze Urban Expressway Crash Risk. *Accident Analysis & Prevention* 101, 37-43.

Yu, R., Wang, X., Yang, K., Abdel-Aty, M., 2016. Crash risk analysis for Shanghai urban expressways: A Bayesian semi-parametric modeling approach. *Accident Analysis & Prevention* 95(Pt B), 495-502.

Yuan, J., Abdel-Aty, M., 2018. Approach-Level Real-Time Crash Risk Analysis for Signalized

Intersections. *Accident Analysis & Prevention* 119, 274-289.

Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-Time Crash Risk Prediction Using Long Short-Term Memory Recurrent Neural Network, *Transportation Research Record: Journal of the Transportation Research Board*, Washington D.C.

Yuan, J., Abdel-Aty, M., Wang, L., Lee, J., Wang, X., Yu, R., 2018a. Real-Time Crash Risk Analysis of Urban Arterials Incorporating Bluetooth, Weather, and Adaptive Signal Control Data, *Transportation Research Board 97th Annual MeetingTransportation Research Board*, Washington D.C.

Yuan, J., Abdel-Aty, M., Wang, L., Lee, J., Yu, R., Wang, X., 2018b. Utilizing bluetooth and adaptive signal control data for real-time safety analysis on urban arterials. *Transportation Research Part C: Emerging Technologies* 97, 114-127.

Yue, L., Abdel-Aty, M., Wu, Y., Wang, L., 2018. Assessment of the safety benefits of vehicles' advanced driver assistance, connectivity and low level automation systems. *Accident Analysis & Prevention* 117, 55-64.

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

Zhu, X., Yuan, Y., Hu, X., Chiu, Y.-C., Ma, Y.-L., 2017. A Bayesian Network model for contextual versus non-contextual driving behavior assessment. *Transportation Research Part C: Emerging Technologies* 81, 172-187.