11-13-2020

Strategies to Identify and Mitigate Secondary Crashes in Real-time

Angela Kitali
akita002@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

STRATEGIES TO IDENTIFY AND MITIGATE SECONDARY CRASHES IN REAL-TIME

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

by

Angela E. Kitali

2020

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Angela E. Kitali, and entitled Strategies to Identify and Mitigate Secondary Crashes in Real-time, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Albert Gan

_____
Mohammed Hadi

_____
Xia Jin

_____
Wensong Wu

_____
Thobias Sando

_____
Priyanka Alluri, Major Professor

Date of Defense: November 13, 2020

The dissertation of Angela E. Kitali is approved.

_____
Dean John L. Volakis
College of Engineering and Computing

_____
Andrés G. Gil
Vice President for Research and Economics Development
and Dean of the University Graduate School

Florida International University, 2020

DEDICATION

To my parents, Mary Marunda and Edes Trifon, grandparents, Flora Raphael and Raphael Marunda, and my siblings Happy, Sara, and Salama, for always being there for me and inspiring me to dream big.

ACKNOWLEDGMENT

Foremost, I thank the Almighty God, I could not have accomplished this work without His blessings.

I would like to convey my immense gratitude and appreciation to my academic advisor, mentor, and role model, Dr. Priyanka Alluri, for her enormous and tireless support in my research work. Her devotion, patience, and understanding provided me with a platform to grow my professional, leadership, and interpersonal skills. Thank you for inspiring me to work hard and persevere.

Sincere appreciation goes to my mentor and master's supervisor, Dr. Thobias Sando. Thank you for being there for me, both professionally and personally. You motivated me to become the person I am today.

I would like to express my heartfelt appreciation to my committee members, Dr. Albert Gan, Dr. Mohammed Hadi, Dr. Xia Jin, and Dr. Wensong Wu, for showing interest in my research work and never hesitating to provide insight and guidance throughout this journey.

I am appreciative of the collaborative and supportive work of everyone who assisted me in my dissertation. Special thanks to Dr. Emmanuel Kidando, Mr. Haifeng Wang, Dr. Boniphace Kutela, Dr. Wanyang Wu, Mr. Jason Crawford, and Ms. Tasha Cunningham. It is my pleasure to thank my lab mates and research co-authors for challenging me to work hard and strive for excellence.

I would like to recognize the financial support from several organizations, including the Florida International University Dissertation Year Fellowship, Florida Department of Transportation, the Women's Transportation Seminar, Institute of Transportation

Engineers, and the American Society of Civil Engineers. Their support helped me accomplish my doctoral degree and prepare me for my professional journey ahead.

To my family and friends, thank you for always loving, encouraging, and cheering on me on each milestone that I achieved. Many thanks to my uncle and aunt, Mr. and Mrs. Edgar Marunda, for supporting me on this journey. To Ruth Kahatano, thank you for your motherly love and support; I have reached this far because of you.

ABSTRACT OF THE DISSERTATION

STRATEGIES TO IDENTIFY AND MITIGATE SECONDARY CRASHES IN REAL-TIME

by

Angela E. Kitali

Florida International University, 2020

Miami, Florida

Professor Priyanka Alluri, Major Professor

Traffic incidents are the primary source of non-recurring congestion. In addition to affecting roadway operations, traffic congestion resulting from an incident exposes other vehicles to the risk of being involved in additional incidents, typically referred to as secondary crashes. Secondary crashes adversely affect traffic operations and impose risk on the safety of both road users and incident responders. Transportation agencies have been looking for ways to mitigate secondary crashes. However, secondary crash mitigation has several challenges. The length of the queue caused by an initial incident and the amount of time this queue lasts on the road varies, depending on the characteristics of the respective incident. Since identifying potential secondary crashes is difficult, investigating the factors that may influence these crashes becomes even more challenging. Moreover, the limited knowledge of what constitutes a secondary crash and its contributing factors largely impede mitigation strategies.

The goal of this research was to investigate approaches to mitigate secondary crashes on freeways. To achieve this goal, a readily implementable data-driven approach to identify secondary crashes in real-time was developed. This approach is more accurate

in identifying secondary crashes since it better reflects the changes in traffic characteristics caused by the primary incident. Following the identification of secondary crashes, the next step involved developing a secondary crash likelihood model. This model established the relationship between the likelihood of secondary crashes and influential factors, i.e., incident characteristics, traffic flow attributes, temporal attributes, presence of work zone, and other geometric attributes. The model results indicate that the presence of work zones significantly influenced the occurrence of secondary crashes. Overall, as expected, roadway geometric, temporal, traffic flow, incident, and weather attributes were found to influence secondary crashes.

The probabilistic relationship between factors that influence the risk of cascading crashes was also explored. Crashes are termed as "cascading" when the subsequent secondary crashes occur within the impact area of the prior secondary crashes and the primary incident. Cascading crashes were found to be most likely to occur when traffic is in the transition state, i.e., when there is a platoon of vehicles traveling at high differential speeds.

Once an incident has occurred, traffic conditions upstream of the incident change with time, and so does the likelihood of secondary crashes. The likelihood model was implemented to dynamically predict the risk of a secondary crash in real-time. The proposed model accounts for the temporal variation of prevailing conditions that influence the likelihood of secondary crashes. This model could be used to develop an Advanced Traffic Management System (ATMS) to proactively prevent secondary crashes. Through this system, first responders will be more vigilant and better prepared in case secondary crashes occur. In addition, motorists upstream of the primary incident could be warned about the potential for secondary crashes.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

LIST OF ACRONYMS

AADT	Annual Average Daily Traffic
API	Application Programming Interface
AUC	Area Under the Curve
AMS	American Meteorological Society
ATDM	Active Transportation and Demand Management
ATIS	Advanced Traveler Information System
ATMS	Advanced Traffic Management System
BCI	Bayesian Credible Interval
BDeu	Bayesian Dirichlet equivalent uniform
CCTV	Closed-Circuit Television
CI	Confidence Interval
Cloglog	Complementary log-log
CV	Connected Vehicle
dBZ	Decibel relative to Z (Reflectivity)
DMS	Dynamic Message Sign
DSRC	Dedicated Short-Range Communication
FB	Full Bayes
FDOT	Florida Department of Transportation
FHWA	Federal Highway Administration
GHC	Greedy Hill Climbing
GIS	Geographic Information System
GPS	Global Positioning System

HCM	Highway Capacity Manual
HEFT	Homestead Extension of Florida's Turnpike
HMC	Hamiltonian Markov Chain
ID	Identification
ITS	Intelligent Transportation System
MAC	Media Access Control
MCMC	Markov Chain Monte Carlo
MM	Mile Marker
MCS	Mainline Central Section
MSS	Mainline South Section
NCHRP	National Cooperative Highway Research Program
NEXRAD	Next Generation Weather Radar
NB	Negative Binomial
NOAA	National Oceanic and Atmospheric Administration
NUTS	No U-Turn Sampling
OR	Odds Ratio
LASSO	Least Absolute Shrinkage and Selection Operator
USDOT	United States Department of Transportation
RCI	Roadway Characteristics Inventory
ROC	Receiver Operating Characteristics
RSU	Roadside Units
SCDOT	South Carolina Department of Transportation
SR	State Road

TIM	Traffic Incident Management
TMC	Transportation Management Center
TRB	Transportation Research Board
TSM&O	Transportation Systems Management and Operations
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
WCT	Weather Climatic Toolkit

CHAPTER 1
INTRODUCTION

1.1 Background

Transportation agencies strive for an efficient transportation system that is safe and has minimal delays. Nevertheless, congestion and traffic incidents have continuously been deterring the performance of the transportation network. The cost of traffic congestion to Americans in 2019 was estimated to be approximately $88 billion, an average of $1,377 per driver (INRIX, 2019). This congestion is partly caused by an increased traffic volume, particularly during peak hours, and is commonly termed as recurrent congestion. Traffic incidents, which include traffic crashes, disabled vehicles, debris on roadways, etc., are also a significant cause of congestion, generally referred to as non-recurrent congestion. Traffic incidents often lead to capacity reduction and deterioration of the level of service. They account for more than half of all urban traffic delays and almost all rural traffic delays (Baykal-Gürsoy et al., 2009).

Traffic incidents also expose other vehicles to the risk of being involved in additional crashes called secondary crashes (Owens et al., 2010). Figure 1-1 explains secondary crashes using a hypothetical example. In this example, a prior traffic incident (a crash in this scenario) occurred on the northbound lanes at 8:33 AM. This crash, categorized as a primary crash, resulted in a queue backup upstream of the crash location. Two crashes, one near the primary crash location and the other further upstream of the primary crash location, occurred at 8:35 AM and 8:38 AM, respectively. Another crash also occurred in the opposite direction (i.e., on the southbound lanes) at 8:55 AM. While the first crash that occurred at 8:33 AM is considered the primary crash, the remaining

three crashes are considered secondary crashes which occurred as a result of the primary crash.

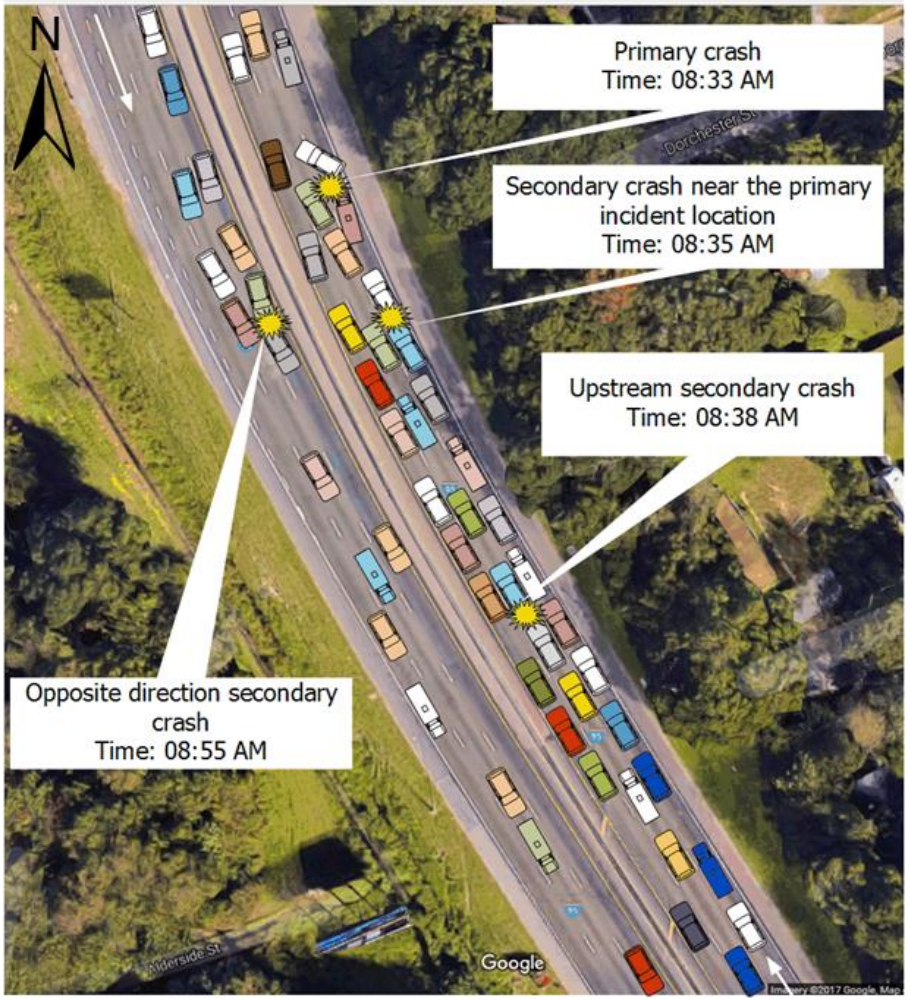


Figure 1-1: Definition of a Secondary Crash

In summary, crashes are considered as secondary crashes if they occur: (a) at the scene of the primary incident (Zhang and Khattak, 2010; Moore et al., 2004); or (b) within the queue upstream of the primary incident (Zhang and Khattak, 2010); or (c) within the queue in the opposite direction of the primary incident caused by driver distraction, i.e., onlookers effect (Yang et al., 2014a).

Secondary crashes have progressively been perceived as a significant issue, particularly on freeways (Hirunyatiwattana and Mattingly, 2006). As such, there has been a growing interest in addressing secondary crash occurrence. Secondary crashes are non-recurring, leading to reduced capacity, additional traffic delays, and increased fuel consumption and emissions. These crashes also affect the safety of both road users and incident responders. The United States Department of Transportation (USDOT) estimated that secondary crashes alone are responsible for approximately 18 percent of all freeway traffic fatalities and 20 percent of all crashes (Owens et al., 2010). Further, compared to primary incidents, secondary crashes have a significant impact on traffic management resource allocation (Vlahogianni et al., 2012; Karlaftis et al., 1999).

Prevention of secondary crashes has, therefore, been highlighted as a high-priority task for traffic incident managers (O'Laughlin and Smith, 2002) and Transportation Management Centers (TMCs) (Owens et al., 2010). The Federal Highway Administration (FHWA) uses the reduction of secondary crashes as one of the performance measures for state incident management systems (National Cooperative Highway Research Program [NCHRP], 2014). The Florida Department of Transportation (FDOT) included secondary crashes as a *Safety* performance measure in its Transportation Systems Management and Operations (TSM&O) Strategic Plan (Florida Department of Transportation [FDOT], 2017). Specifically, to reduce the risk to responders, secondary crashes, and delays associated with incidents, FDOT has an Open Roads Policy of clearing all travel lanes within 90 minutes. Several states also consider secondary crash mitigation strategies in allocating funding for the development of Traffic Incident Management (TIM) programs

and on-road help services, such as FDOT's Road Ranger freeway service patrol (Lou et al., 2011).

1.2 Problem Statement

Secondary crashes adversely affect the operational and safety performance of the transportation network. As such, agencies are looking for ways to mitigate secondary crashes to reduce non-recurrent delays and the adverse safety impacts associated with these crashes. However, some hurdles limit the implementation of approaches to reduce secondary crashes. First and foremost, the process of identifying secondary crashes is itself a challenge since there is no universal definition of a secondary crash. The inconsistency in defining secondary crashes limits the possibility of exploring the underlying relationship between secondary crash occurrences and influential factors. This limitation, in turn, hinders the mitigation efforts. The following subsections provide a detailed discussion on the challenges facing the identification and prediction of secondary crashes. A thorough exploration of these challenges will assist in developing effective policies and countermeasures to mitigate the risk of secondary crashes.

1.2.1 Identify Secondary Crashes

Not all incidents lead to secondary crashes. The likelihood of secondary crashes depends on several factors, including traffic flow characteristics, incident characteristics, weather conditions, roadway geometric conditions, etc. An in-depth understanding of these factors will help agencies on several fronts. First, it will assist in proactively preventing secondary crashes. Second, first responders will be more vigilant and better prepared in

case secondary crashes occur. And finally, motorists upstream of the primary incident could be warned about potential secondary crashes.

From a statistical learning perspective, secondary crash risk modeling can be viewed as a dichotomous classification problem, where 1 indicates that the secondary crash occurred, and 0 indicates that no secondary crash occurred. Secondary crashes are generally infrequent (Kitali et al., 2018; Yang et al., 2018; Xu et al., 2016; Owens et al., 2010). This means that the proportion of incidents that result in secondary crashes (i.e., primary incidents) is much less than the proportion of incidents that do not cause secondary crashes, referred to in this research as normal incidents. This asymmetric nature of the binary response variable makes the modeling of the likelihood of secondary crashes an imbalanced classification problem. Neglecting this imbalance characteristic can lead to serious consequences, both in the model's estimates and prediction accuracy (Kitali et al., 2019b).

Previous studies have considered several incident-related, traffic-related, geometric-related, and weather-related factors when developing secondary crash risk models. However, simply incorporating all variables in the model may lead to biased results, considering the possible significant correlation among the variables. Only a few studies have considered identifying the most important variables before developing secondary crash risk models. Variable subset selection methods, such as a stepwise technique, have been used to add one best-fit variable at a time during model fitting (Mishra et al., 2016; Xu et al., 2016; Zhan et al., 2009). Nevertheless, this criterion has several drawbacks, including the result that each addition of a new feature may render one or more of the already included variables non-significant. Also, because the stepwise variable

selection process is discrete, it often exhibits high variance and may not reduce the prediction error of the full model. In other words, small changes in the data can result in different variables being selected, and this can potentially reduce the model's prediction accuracy (Menard and Torelli 2014; Tibshirani 1996).

In general, three major challenges are encountered when modeling the risk of secondary crashes: (1) infrequent nature of secondary crashes, (2) selection of the most important variables, and (3) identification of variable correlation. Therefore, any candidate model needs to account for these issues.

Occasionally, as indicated in Figure 1-2, some primary incidents result in a series of cascading crashes. Crashes are termed as "cascading" when the subsequent secondary crashes occur within the impact area of the prior secondary crashes and the primary incident. Events consisting of cascading crashes are expected to have longer impact duration and hence larger impacts on traffic. This situation presents additional impedance and increases interference among vehicles, particularly in upstream traffic (Zhang and Khattak, 2010).

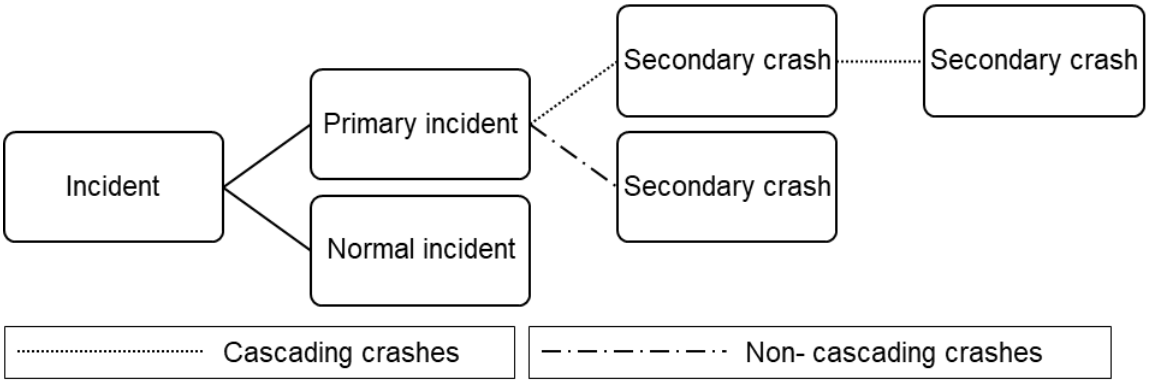


Figure 1-2: Illustration of the Occurrence of Cascading Crashes

1.2.2 Mitigate the Risk of Secondary Crashes

In an earlier study by Karlaftis et al. (1999), the likelihood of secondary crashes was observed to increase by 2.8 percent for each additional minute required to clear the initial crash. Other recent studies also associated an increase in incident clearance duration with a higher likelihood of secondary crashes (Goodall, 2017; Kitali et al., 2019b, 2018; Sando et al., 2018). In this case, managing secondary crashes requires a proactive approach, i.e., an ability to alleviate the risk of secondary crashes before they occur. Only a few studies have focused on drafting and deploying specific countermeasures to mitigate the risk of secondary crashes (Park et al., 2018; Park and Haghani 2016b; Yang et al., 2017; Kopitch and Saphores 2011; Karlaftis et al., 1999).

A proactive approach requires the proposed strategy to accurately identify whether the current traffic incident has a probability of resulting in additional incidents (i.e., probable primary incident). Upon confirming that the present incident has a likelihood of becoming a primary incident, the next steps involve estimating its impact (in terms of time and distance) and timely disseminate safety messages to affected traffic. Recent studies have, therefore, relied on the use of real-time traffic data to identify and predict the likelihood of secondary crashes using different modeling approaches. However, the proposed models were developed and calibrated with static parameters, which do not account for the temporal variation in traffic characteristics influenced by the incident. The prevailing traffic conditions before the occurrence and those following the incident's occurrence may have a significant and varying impact on the likelihood of secondary crashes. Furthermore, the magnitude of the impact of the traffic flow characteristics on the likelihood of secondary crashes is expected to vary with time.

1.3 Research Goal and Objectives

The goal of this research was to investigate approaches to mitigate secondary crashes on freeways. The specific objectives of this research include:

1. Identify potential factors that influence the risk of secondary crashes.
2. Predict the probability of secondary crashes in real-time.

Figure 1-3 presents the main steps used to implement the two objectives. The first step involved identifying secondary crashes using high-resolution traffic data. Specifically, the high-resolution traffic data were used to automatically determine the spatiotemporal impact areas of primary incidents, and hence, detect secondary crashes that occurred within the affected area. This research focused only on secondary crashes that occurred in the upstream direction of the primary incident. This is because secondary crashes in the opposite direction of the primary incident are affected by other factors, such as the visibility of drivers in the opposing direction. In this case, visibility is influenced by several attributes, including median width, median type, and type of median barrier.

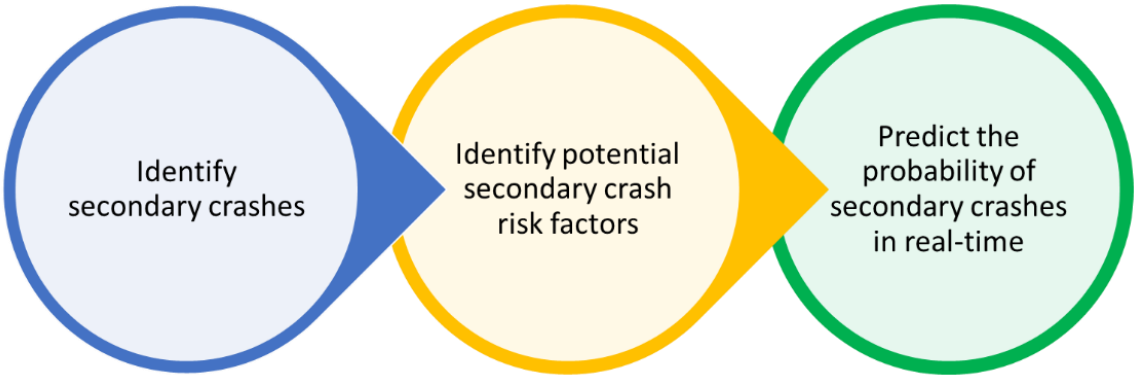


Figure 1-3: Development of Real-time Secondary Crash Risk Prediction Model

After identifying secondary crashes, the second step involved identifying factors influencing the likelihood of secondary crashes. A penalized logistic regression model,

fitted using a bootstrap approach, was used to link the likelihood of secondary crashes with potential factors, including roadway geometric, temporal, traffic flow, incident, and weather attributes.

Following the occurrence of an incident, some of the influential attributes of a potential secondary crash, e.g., traffic flow characteristics and weather conditions, may vary with time. It is hypothesized that the temporal variation in these attributes will be accompanied by the changes in the likelihood of secondary crashes. Thus, this research developed a dynamic model that predicts the likelihood of secondary crashes in real-time. That is the proposed model accounts for the temporal variation of prevailing conditions that influence the likelihood of secondary crashes.

1.4 Dissertation Organization

The remaining chapters of this dissertation are organized as follows:

- Chapter 2 entails a comprehensive synthesis of the literature on the main approaches used to identify secondary crashes. The chapter discusses the methods used to predict the probability of secondary crashes and presents the approaches being adopted to mitigate secondary crashes. Also presented is a summary of the research areas that require further investigation relating to the identification of secondary crashes, understanding of factors influencing the occurrence of secondary crashes, and the prediction of secondary crashes in real-time.
- Chapter 3 focuses on discussing the data used to achieve the research goal. Specifically, the chapter discusses, in detail, the types of data used, data sources, data collection strategy, and data processing steps.

- Chapter 4 discusses the methodologies used to achieve the research objectives.
- Chapter 5 presents the analyses and discusses the results. The secondary crash identification results are first discussed, followed by the results of the likelihood model on the influence of work zones on secondary crashes. Finally, the results of the dynamic real-time secondary crash risk prediction model are discussed.
- Chapter 6 concludes this dissertation by providing a summary of this research, contributions, and recommendations for future research.

CHAPTER 2
LITERATURE SYNTHESIS

This chapter presents a synthesis of previous studies that focused on identifying secondary crashes and analyzing the risk factors influencing the occurrence of these crashes. Section 2.3 of this chapter presents previous literature that explored strategies to mitigate secondary crashes. The areas of research that need further investigation associated with the identification, prediction, and prevention of secondary crashes are discussed in the last section.

2.1 Existing Methods to Identify Secondary Crashes

Secondary crashes are traffic incidents that occur within the spatial and temporal impact area of the primary incidents (Zhang and Khattak, 2010; Moore et al., 2004; Yang et al., 2014a; Karlaftis et al., 1999). Unlike other traffic incidents that are easily identified by incident responders, detection of secondary crashes is not straightforward since the definition itself is subjective. It is difficult to determine visually, either directly at the crash site or through closed-circuit television (CCTV) cameras, if the crash is a result of the backup caused by another incident, especially since the backup may also be due to recurrent congestion. Thus, the first step in identifying secondary crashes is to define the impact area of the prior incident, i.e., its spatiotemporal boundaries.

As summarized in Table 2-1, three major approaches have been used to define the spatiotemporal impact area of primary incidents: (1) manual method, where personnel visually estimate the queue of the primary incident; (2) static method that uses predefined spatiotemporal thresholds; and (3) dynamic approach that estimates the primary incident influence area as a function of its impact on traffic flow characteristics, e.g., speed, volume,

and/or density. An extensive literature review revealed that tremendous efforts have been made to identify secondary crashes. The following subsections provide more details about these three methods.

Table 2-1: Methods Used to Identify Secondary Crashes

Method	Approach	Advantages	Challenges
Manual	Personnel visually identify secondary crashes: • On-site approach using incident responders, e.g., Highway Patrol, etc. • Off-site approach using CCTV, etc.	• Simple • Does not require any data processing	• Subjective • Unreliable • Inconsistent • Random
Static	Identify secondary crashes based on predefined distance and time thresholds for each primary incident (e.g., 2 miles upstream and 2 hours after the primary incident)	• More reliable than the manual method • Relatively easy to implement	• Less reliable compared to the dynamic method
Dynamic*	Identify secondary crashes based on the queue length of the primary incident, estimated based on prevailing traffic conditions	• Most reliable • Accurate	• Resource intensive • Limited by data availability

Note: *Can be reliably implemented in real-time; CCTV = Closed Circuit Television.

2.1.1 Manual Method

As the term "manual" indicates, in this method, secondary crashes are manually identified by either TMC personnel or incident responders (Kitali et al., 2019a). In this case, the impact area of primary incidents is estimated visually based on the judgment of the observer. Identifying secondary crashes on a CCTV camera is considered an off-site approach, while identifying secondary crashes on-site by incident responders, including police, on-road service patrols (e.g., FDOT's Road Rangers), etc., is considered an on-site approach (NCHRP, 2014). The manual method has traditionally been used by agencies to identify secondary crashes. It is simple and does not require any data processing. However, despite being the most used method, it is subjective, unreliable, inconsistent, and random.

2.1.2 Static Method

The static method identifies secondary crashes based on some fixed spatial and temporal thresholds, i.e., the primary incident impact area is pre-determined. Crashes that occur within the spatial and temporal impact range of a primary incident are identified as secondary crashes. Figures 2-1 and 2-2 graphically summarize previous studies that identified secondary crashes based on fixed spatial and temporal thresholds (Chang and Rochon, 2011; Green et al., 2020; Hirunyatiwattana and Mattingly, 2006; Jalayer et al., 2015; Karlaftis et al., 1999; Kopitch and Saphores, 2011; Latoski et al., 1999; Moore et al., 2004; Raub, 1997; Tian et al., 2016; Zhan et al., 2008).

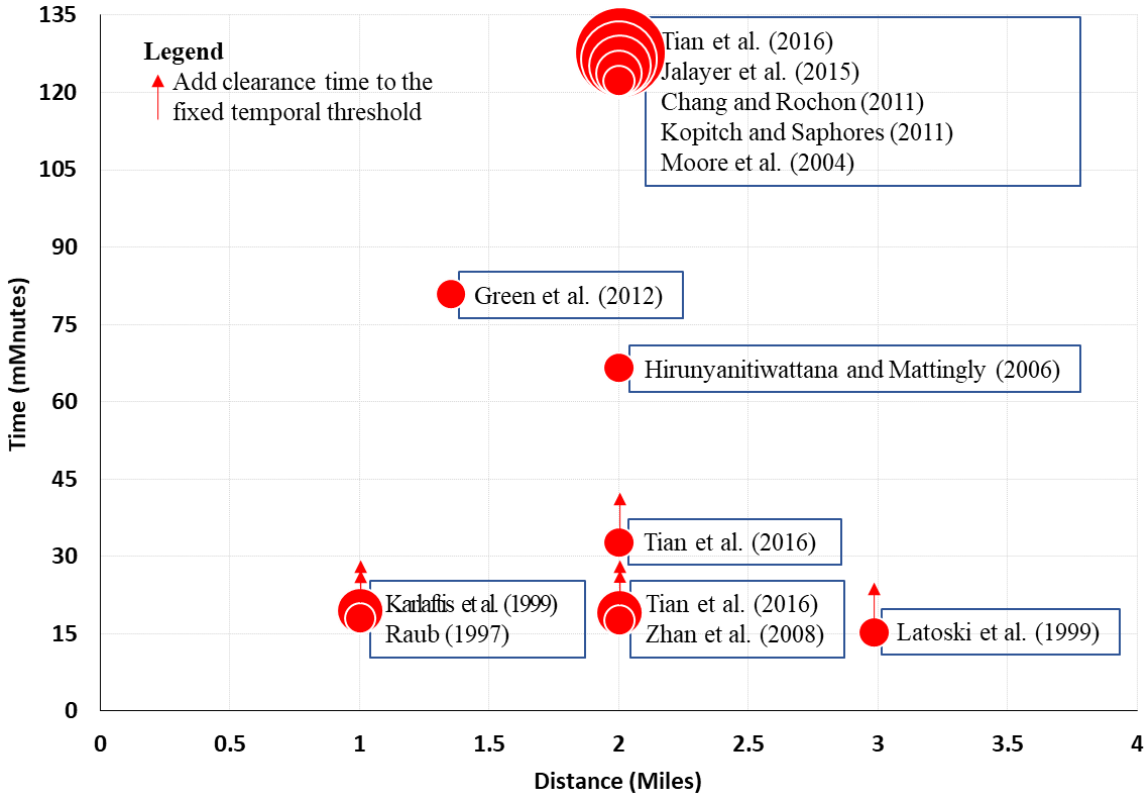


Figure 2-1: Studies that used Static Method to Identify Secondary Crashes in the Upstream Direction

As indicated in Figure 2-1, the spatial and temporal thresholds adopted by these studies range between 1 to 3 miles and 15 minutes to 2 hours, respectively.

Similarly, secondary crashes that occurred in the opposite direction of the primary incident – because of the onlooker effect – were also commonly identified using some different predefined thresholds. Figure 2-2 summarizes the studies that used the static method to identify secondary crashes in the opposite direction of the primary incident (Chang and Rochon, 2011; Green et al., 2012; Kopitch and Saphores, 2011; Moore et al., 2004).

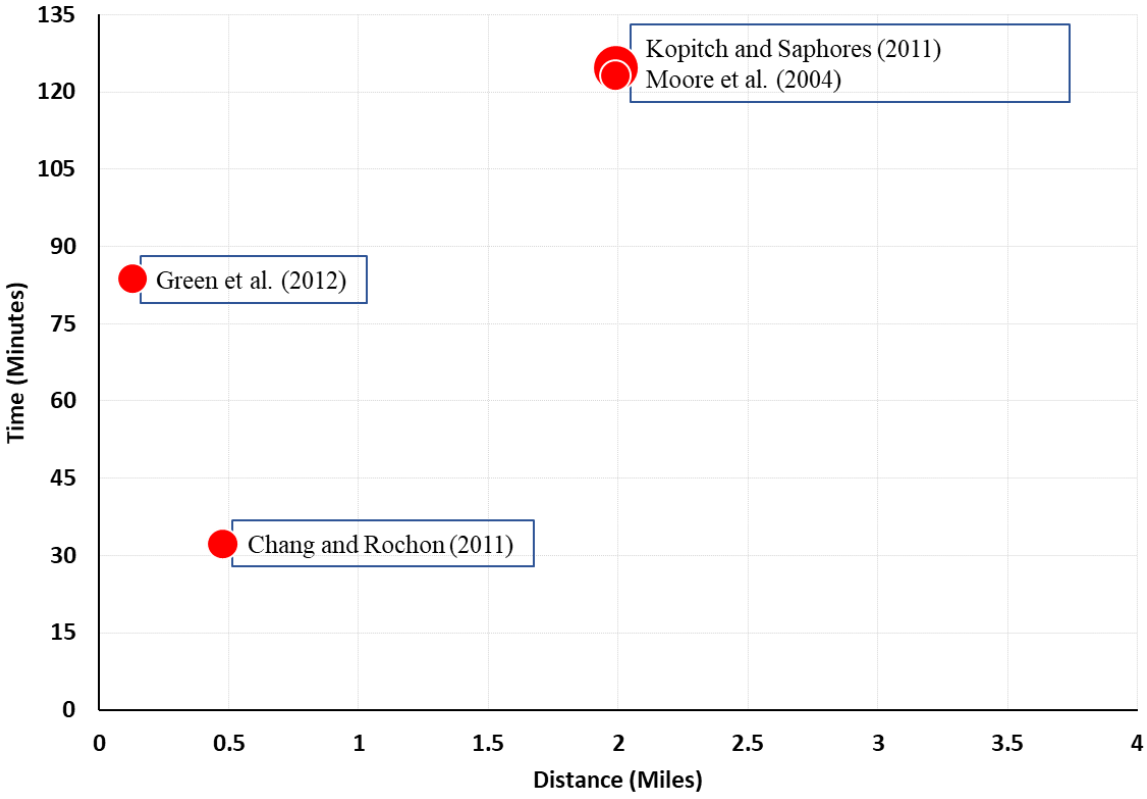


Figure 2-2: Studies that used Static Method to Identify Secondary Crashes in the Opposite Direction

For example, Chang and Rochon (2011) identified secondary crashes using a 30-minute and 0.5-mile threshold in the opposite direction of the primary incident. Meanwhile, Green

et al. (2012) used a spatial threshold of 1,000 feet and a temporal threshold of 80 minutes. Other studies used similar thresholds to identify secondary crashes both on the upstream and opposite directions of the primary incident (Kopitch and Saphores, 2011; Moore et al., 2004).

Unlike the manual method, the static method is more reliable simply because it is a function of predefined spatiotemporal parameters and not based on human judgment. However, the static method's one-size-fits-all approach of using fixed spatiotemporal thresholds do not yield reliable results (Kitali et al., 2019a). In other words, the fixed spatiotemporal thresholds do not effectively reflect the dynamic impact of incidents with varying characteristics, and therefore, may under- or overestimate the impact area (Ou, Xia, Wang, Wang, and Lu, 2020). To accurately identify secondary crashes, the impact area of the primary incidents should be defined based on its impact on traffic flow characteristics.

2.1.3 Dynamic Method

To overcome the limitations associated with the static approach, some studies have focused on identifying secondary crashes based on prevailing traffic flow conditions at the time of the primary incident. In this case, spatiotemporal thresholds are flexibly selected based on the impact of the primary incident on traffic flow parameters, hence the term *dynamic*. The dynamic methods used in previous studies to identify secondary crashes can generally be grouped into three categories: queuing model-based, shockwave-based, and traffic data-based.

Figure 2-3 graphically summarizes the previous studies that explored the use of dynamic models to estimate the impact area of the primary incident (Chung, 2013; Dougald et al., 2016; Goodall, 2017; Imprialou et al., 2014; Kitali et al., 2019a, 2019b, 2018; Mishra et al., 2016; Park and Haghani, 2016a, 2016b; Sando et al., 2019; Sarker et al., 2017; Sun and Chilukuri, 2010, 2006; Vlahogianni et al., 2010, 2012; Wang et al., 2016, 2018; Xu et al., 2016; Yang et al., 2014a, 2014b, 2014c; Zhan et al., 2009; Zhang and Khattak, 2010; Zheng et al., 2014). Sun and Chilukuri (2006) proposed the use of an incident progression curve, a method that relies on incident duration to estimate the queue length, and hence, identify secondary crashes that occurred within the queue. The incident progression curve method indicated a 30% improvement in secondary crash identification accuracy compared to the static method. Zhan et al. (2009) used the cumulative arrival and departure rate approach to estimate the spatiotemporal impact area of incidents with lane blockages. Crashes that occurred within the estimated primary crash incident impact area were marked as secondary crashes (Zhan et al., 2009). Zhang and Khattak (2010) estimated the primary incident impact area based on deterministic queueing models.

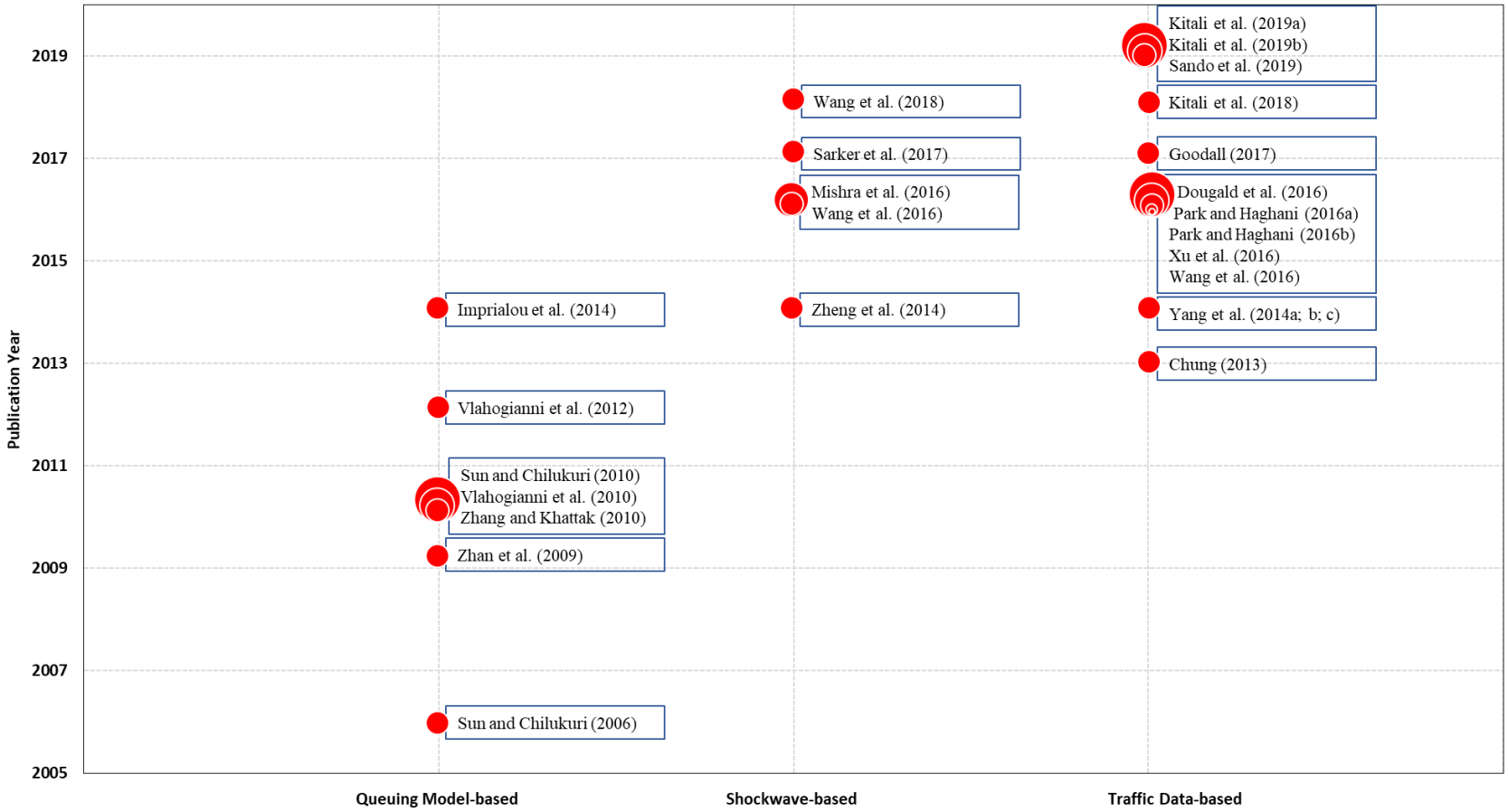


Figure 2-3: Existing Literature on Dynamic Methods

Although queuing methods provide a more realistic representation of incident impact areas, compared to the static approach, they generally rely upon the number and nature of the accessible variables, such as assumed roadway capacities, arrival rates, etc. Different roadway segments are subject to different queuing formation processes because of their unique traffic, geometry, and incident characteristics, as well as prevailing weather conditions.

Apart from queuing approaches, other studies have used shockwave principles to dynamically identify secondary crashes, as shown in Figure 2-3 (Mishra et al., 2016; Sarker et al., 2017; Wang et al., 2019). In this case, the incident impact area is triangular. The spatiotemporal thresholds comprise the backward forming and discharging shockwaves linked with the occurrence and clearance of the incident (H. Yang et al., 2018). The backward-forming shockwave impacts the growth rate of the queue generated by the incident. When the incident is cleared, a forward-recovery shockwave initiates and ultimately reaches the backward-forming shockwave resulting in queue dissipation.

Figure 2-4 demonstrates the use of shockwave principles to identify secondary crashes. In Figure 2-4, the primary incident was assumed to generate three shockwaves, two upstream forming shockwaves, and one upstream dispersing shockwave (Wang et al., 2019). The first shockwave was assumed to be generated when an incident occurs resulting in reduced speeds and increasing density, a situation that creates a bottleneck until the treatment reaction commences. The second shockwave was assumed to be generated after incident responders, such as police and/or tow trucks, arrive at the incident scene causing more deterioration to traffic flow conditions. These first two shockwaves were further assumed to continue until dispersal. The last shockwave occurs after the bottleneck is eased

and the traffic congestion starts to clear (Wang et al., 2019). A crash that falls within the gray area represented in Figure 2-4 is considered as a secondary crash.

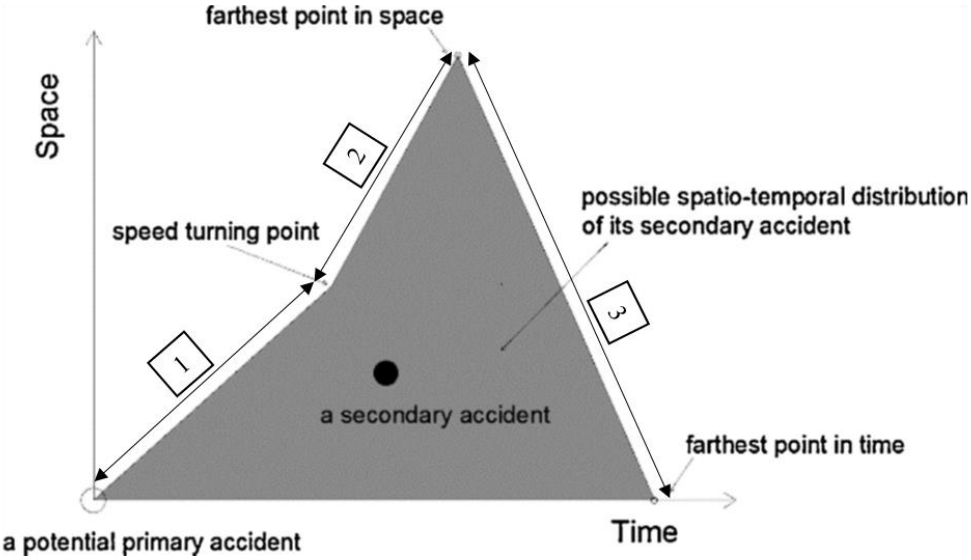


Figure 2-4: Definition of Spatiotemporal Impact Area Using Shockwave Principles (Wang et al., 2019)

Several issues limit the application of the shockwave approach for identifying secondary crashes. The simplified assumption on the prevailing traffic conditions and modeling of the shockwave propagation remains to be an issue since they cannot accurately depict the dynamic progression of traffic states (H. Yang et al., 2018). Non-constant discharge and arrival rates make it difficult to model the complicated shockwaves with the assumption of a constant speed. Overall, both the queuing and shockwave dynamic methods use prior assumptions to simplify the complex characteristics of the traffic conditions, resulting in an incorrect estimation of the incident impact areas. Further, both methods cannot accurately distinguish the recurrent congestion from the non-recurrent congestion caused by the incident (Ou et al., 2020).

To overcome the limitations of the queuing and shockwave methods, recent efforts in identifying secondary crashes have been shifted towards the use of data-driven approaches. Empowered by the advancements in traffic data collection technologies, several studies have explored the use of data-driven approaches to identify secondary crashes. These approaches take advantage of the readily available traffic data retrieved from infrastructure-based traffic sensors, probe vehicles, crowdsourced traffic data from third-party vendors, and connected vehicle (CV) technologies.

The key premise of the data-driven approach is to use prevailing traffic conditions data to accurately estimate the incident impact area. Vehicle speed was the main traffic flow characteristic used by previous studies that identified secondary crashes using a data-driven approach (Sando et al., 2019; Kitali et al., 2019a; 2018; Goodall, 2017; Park et al., 2018, 2017; Dougald et al., 2016; Park and Haghani 2016a; 2016ab; Yang et al., 2014; Chung, 2013). The foundation of data-driven approaches is the determination of a reference speed, and to accomplish this, different methodologies have been proposed. These approaches mostly rely on the use of historical speed data.

Yang et al. (2014a) identified the incident-induced impact area by comparing the prevailing speed data from microwave detectors with the pre-defined percentile speed of historical speed data. That is, the speed is stated to be affected by the incident if it drops below the 50^{th} percentile of the historical speed measurement. This information was used to develop speed contour plots which were then used to identify the pairs of primary crashes and secondary crashes. Other previous studies used a similar approach to define the reference speed (Xu et al., 2016; Yang et al., 2014b, 2014c; Chung, 2013). Nonetheless,

estimating the congestion and non-congestion thresholds empirically may be time-consuming and hinder the transferability of the results to other locations.

Dougald et al. (2016) and Goodall (2017) extended the approach proposed by Yang et al. (2014) by adjusting the assumption used to establish the binary contour plots. Meanwhile, other studies (Park et al., 2018, 2017; Park and Haghan, 2016a, 2016b) used a Bayesian structure equation to estimate the impact area of a primary incident. Kitali et al. (2019a, 2019b; 2018) used a 95% confidence interval to define the upper and lower bounds of the speed profile.

By synthesizing the real-time traffic data and traffic incident data to identify the prevailing traffic conditions, the data-driven approach greatly improved the process of identifying the impact area of the primary incident (Yang et al., 2018). Reference speed is a foundational component of the data-driven methods used to identify secondary crashes, and accurate estimation of reference speeds depends on the completeness of the available traffic data. Similar to other dynamic methods, another issue to be considered while developing data-driven approaches to identify secondary crashes is the approach used to classify congestion and non-congestion patterns. Further consideration must be made to mimic how congestion builds up and dissipates along the segments impacted by the primary incident. This is an important step towards the accurate identification of secondary crashes. Failure to properly estimate the incident impact area may lead to over or underestimation of the impact area, and hence, the number of secondary crashes caused by the respective incident.

Results from previous studies indicate that the proposed dynamic methods provide better accuracy in identifying secondary crashes than conventional static methods (H. Yang

et al., 2018). Compared to the static or manual method, dynamic methods are more advanced and reliable since they identify secondary crashes based on prevailing traffic flow characteristics. It is worth noting, however, that the implementation of dynamic approaches depends on the availability of reliable traffic data.

2.2 Prediction of Probability of Secondary Crashes

Following the identification of secondary crashes, the next step towards developing strategies to mitigate secondary crashes is to explore the causal relationship between secondary crashes and potential explanatory variables. Identifying risk factors that influence the likelihood of secondary crashes is critical to the development and implementation of efficient and resilient traffic management strategies. An effective strategy will assist in proactively preventing secondary crashes and allow first responders to be more aware of potential secondary crashes and be better prepared should they occur. In addition, motorists upstream of the primary incident could be warned about potential secondary crashes. The following subsection presents the methods used to identify factors contributing to the occurrence of secondary crashes. The last subsection discusses the issues that arise when developing secondary crash risk models and ways to address them.

2.2.1 Secondary Crash Risk Prediction Models

A comprehensive literature review revealed that only a few studies have explored secondary crash risk models. Both parametric and non-parametric models have been used to analyze the likelihood of secondary crashes. Most of these studies adopted the respective models following the binary nature of secondary crash occurrence given the presence of a primary incident or normal incident. Primary incidents refer to incidents that resulted in a

secondary crash, while normal incidents refer to incidents that did not result in a secondary crash. In these studies, geometric, weather, traffic conditions, and incident characteristics associated with primary incidents were compared with those of normal incidents.

2.2.1.1 Parametric Models

Most of the studies that developed parametric models used binary regression models, such as logit, probit, or complementary log-log models, to analyze the likelihood of secondary crashes (Goodall 2017; Wang et al., 2016; Karlaftis et al., 1999; Kopitch and Saphores 2011; Zhan et al., 2008, 2009). In these studies, the response variable was dichotomous, with a "*yes*" category representing incidents that resulted in a secondary crash and a "*no*" category representing incidents that did not result in a secondary crash. As mentioned earlier, these two categories of incidents are generally referred to as primary incidents and normal incidents, respectively.

In secondary crash risk models, the independent variables include a list of potential factors that may contribute to the likelihood of secondary crashes. The coefficients obtained by estimating the relationship between the probability of a secondary crash following a primary incident, based on a set of explanatory variables, can hence be used to quantify the impact of each contributing factor on the secondary crash risk. Table 2-2 presents a summary of studies that used parametric modeling approaches to explore risk factors that influence the likelihood of secondary crashes. Included in Table 2-2 are secondary crash identification methodologies, secondary crash risk prediction models, and significant variables in each study.

Table 2-2: Summary of Literature on Parametric Secondary Crash Risk Models

Reference	Secondary crash identification method	Method	No. of var.	Variable selection method	Significant variables
Karlaftis et al., 1999	Static (1 mile and 15 min)	LR	18	Not Applicable	Season, clearance time, type of vehicle involved, and lateral location
Zhan et al., 2008	Static (2 miles and 15 min + clearance)	LR	18	Not Applicable	Number of lanes, primary incident duration, time-of-day, number of vehicles, and vehicle rollover.
Zhan et al., 2009	Cumulative arrival and departure	LR	19	Forward conditional criterion	Primary incidents type and lane-blockage duration, time of day, and direction where the incident occurred
Kopitch and Saphores 2011	Static (2 miles and 2 h)	LR	9	Not Applicable	Number of vehicles, number of trucks, changeable message sign, and road work project
Khattak et al., 2012	Static (1 mile and duration of primary incident (+15 min if lane blocked))	LR	13	Not Applicable	Incident duration, crashes, peak hours, number of vehicles, and AADT
Yang et al., 2014b	Data-driven approach	LR (rare events)	10	Statistically significance level (0.1)	Daytime off-peak hours, daytime peak hours, duration, rear-end crashes, lane closure, and winter season
Wang et al., 2016	Shockwave principles	LR	12	Not Applicable	Shockwave originating in the wake of a primary incident, duration, unsafe speed, and weather
Mishra et al., 2016	Shockwave principle	Linear probability model, LR, and MNL	16	VIF correlation factor and significance level	Average speed of upstream traffic, upstream flow, AADT, incident type, number of vehicles, weather condition, and functional class
Wang et al., 2019	Shockwave principle	LR	13	Not Applicable	Shockwave speed that occurred at the time of the primary incident, shockwave speed generated when incident responders arrive at the scene to control traffic, shockwave speed during dissipation, incident processing duration, unsafe speed, and rain.

Note: AADT = Annual Average Daily Traffic; Cloglog = complementary log-log, LASSO = Least Absolute Shrinkage and Selection Operator, LR = Logistic Regression, MNL = Multinomial Logistic Regression, No. of var. = Number of variables, VIF = Variance Inflation Factor.

Table 2-2: Summary of Literature on Parametric Secondary Crash Risk Models (continued)

Reference	Secondary crash identification method	Method	No. of var.	Variable selection method	Significant variables
Xu et al., 2016	Data-driven approach	Bayesian LR	24	Pearson correlation and stepwise logit	Average speed, traffic volume, standard deviation of detector occupancy, volume difference between adjacent lanes, crash severity, crash type, day of the week, road surface condition, and number of lanes
Goodall, 2017	Data-driven approach	LR	3	Not Applicable	Congestion and incident duration
Sarker et al., 2017	Shockwave principle	Generalized ordered response probit	15	Not Applicable	AADT, traffic composition, land use, number of lanes, right side shoulder width, posted speed limits, and ramp indicator
Kitali et al., 2018	Data-driven approach	Bayesian cloglog	21	Random Forest	Average occupancy, incident severity, percent of lanes closed, incident type, incident clearance duration, incident impact duration, and incident occurrence time.
Kitali et al., 2019b	Data-driven approach	Penalized LR (with resampling)	23	LASSO	Mean of detector occupancy, coefficient of variation of equivalent hourly volume, mean of speed, incident type, percent lane closed, incident occurrence time, shoulder blocked, number of responding agencies, incident impact duration, incident clearance duration, and roadway alignment

Note: AADT = Annual Average Daily Traffic; Cloglog = complementary log-log, LASSO = Least Absolute Shrinkage and Selection Operator, LR = Logistic Regression, MNL = Multinomial Logistic Regression, No. of var. = Number of variables, VIF = Variance Inflation Factor.

Using five years of incident data from the Borman Expressway in Northwest Indiana, Karlaftis et al. (1999) developed logistic regression models to explore the influence of primary incident characteristics on the likelihood of a secondary crash. The study found that a number of factors can significantly influence the likelihood of secondary crash occurrence, including clearance time of the primary incident, season, type of vehicle involved, day of the week, and lateral location of the primary incident.

Zhan et al. (2009) reported the time of day, the primary incident type, and primary incident lane-blockage duration as the most influential factors that affect the occurrence of secondary crashes. The study results further indicated that the incident duration had the greatest influence on secondary crash occurrence. Kopitch and Saphores (2011) observed that for each additional vehicle involved in a primary incident event, the odds of having a secondary crash increase by a factor of 1.161. In addition to the number of vehicles involved in a primary incident, Kopitch and Saphores (2011) found that primary incident injury type and severity also significantly influence the risk of a secondary crash. Specifically, compared to other primary incident types such as road hazards, crashes were observed to increase the likelihood of secondary crashes by 1.936. Compared to other severity levels, fatal and severe injury incidents were found to increase the odds of a secondary crash by a factor of 3.177 (Kopitch and Saphores, 2011).

Similarly, Goodall (2017) and Wang et al. (2016) used a logistic regression model to predict the likelihood of a secondary crash. Shockwaves that originated in the wake of a primary incident were observed to significantly impact the probability of a secondary crash occurrence than traffic volume (Wang et al., 2016). Based on this observation, the study suggested that incident responders that arrive at the scene of an incident to control traffic

should not suddenly block, decrease, or release the traffic flow, but rather to control traffic in a smooth and controlled manner (Wang et al., 2016).

Khattak et al. (2012; 2009) extended the conventional logistic regression model to account for the interdependence between primary incidents and secondary crashes. Secondary crashes were observed to be more likely to occur if the duration of the primary incident was long; accordingly, the duration is expected to be longer if secondary crashes occur. The interrelationship between incident durations and the occurrence of secondary crashes was modeled using a two-level hierarchical prediction approach. First, the incident duration was estimated using an ordinary least square regression model. Next, a logistic regression model was fitted using estimated duration time and other factors, such as weather, road information, and Annual Average Daily Traffic (AADT), to analyze the occurrence of secondary crashes.

Using the proportional test, Hirunyanitiwattana and Mattingly (2006) assessed the significant differences in the characteristics of secondary crashes and primary crashes with respect to time of day, area type (urban or rural), collision type, primary collision factor (e.g., speeding, failure to yield, alcohol, etc.), road classification (freeways, multi-lane, and two-lane), and crash severity. A primary crash that occurred during the peak period was found to be more likely to result in secondary crashes than during other periods. On the other hand, the probability of secondary crashes in urban districts was observed to be higher than in rural districts. Urban freeways with more than four lanes were reported as the type of roadway with the highest number of primary and secondary crashes. Speeding was identified as the highest collision factor for primary and secondary crashes (Hirunyanitiwattana and Mattingly, 2006). Khattak et al. (2009) estimated the likelihood

of secondary crash occurrence using a binary probit model. The study results indicated that the primary incident duration, number of involved vehicles, and AADT had a significant positive impact on the likelihood of secondary crashes (Khattak et al., 2009).

While most primary incidents result in one secondary crash, some primary incidents result in multiple secondary crashes, and others result in cascading crashes. As mentioned earlier, crashes are termed as "cascading" when the subsequent secondary crashes occur within the impact area of the prior secondary crashes and the primary incident. As illustrated in Figure 2-5, crashes are identified as "multiple secondary crashes" (and not as "cascading crashes") when two or more secondary crashes caused by the same primary incident are not necessarily within the impact area of either of the secondary crashes. A few studies have modeled the risk of multiple secondary crashes (Xu et al., 2019; Sarker et al., 2017; Mishra et al., 2016; Zhang and Khattak, 2010). The ordered logit model, multinomial logit model, and zero-inflated ordered probit regression model are some of the models used to model multiple secondary crashes caused by a single primary incident.

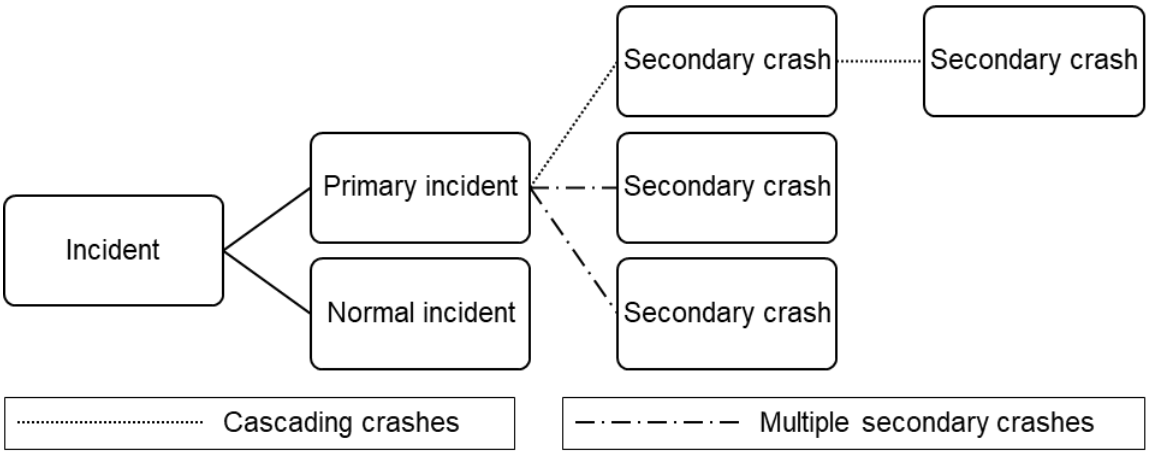


Figure 2-5: Illustration of Difference Between Cascading Crashes and Multiple Secondary Crashes

Zhang and Khattak (2010) used ordered logit model to investigate the factors contributing to multiple secondary crashes. Incidents were categorized using a three-point ordinal scale: (1) a normal incident; (2) one primary incident-secondary crash pair; and (3) one primary incident with two or more secondary crashes. This scale was created to capture event adversity from a traffic management perspective, with the last category capturing multiple secondary crashes (Zhang and Khattak, 2010). The results suggested that the probability of multiple secondary crashes increased with an increase in the number of involved vehicles and lane blockage.

Sarker et al. (2017) developed a Poisson model, negative binomial (NB) model, NB model with heterogeneous dispersion, and NB model with heterogeneous dispersion and unobserved heterogeneity to predict the frequency of secondary crashes. The following factors were reported to significantly affect secondary crash occurrence: posted speed limit higher than 55 miles per hour, AADT, urban land use, number of lanes, right shoulder width, and presence of ramp.

Xu et al. (2019) used a zero-inflated ordered probit regression model to study the effects of prevailing traffic characteristics on the likelihood of multiple secondary crashes caused by a single primary incident. Other potential factors considered include incident characteristics, weather conditions, and roadway geometric attributes. Two states were considered in modeling the frequency of secondary crashes. The first state, the *secondary-crash-free* state, predicted whether the initial incident will lead to secondary crashes, and the second state, referred to in the study as the *secondary-crash-prone* state, determined the frequency of secondary crashes caused by one primary incident. The following factors were found to be influential in the secondary-crash-free state: average traffic volume,

average speed, and the difference between the number of on-ramps and off-ramps in a segment. In the secondary-crash-prone state, the significant factors that were found to influence the likelihood of multiple secondary crashes included hit-and-run primary crashes, average detector occupancy, rainy weather, and primary crash severity.

2.2.1.2 Non-parametric Models

Non-parametric models, such as Bayesian neural networks and decision trees, have also been used to model secondary crash risk (Vlahogianni et al., 2010, 2012). A fundamental difference between non-parametric models and parametric models is that the non-parametric models lack an inherent mechanism for explicitly describing the significance of input variables, and hence, considered to be a black box (H. Yang et al., 2018). The need for developing non-parametric models with explanatory power is related to the decision-making process in transportation. Instinctively, any decision in transportation and traffic operations ought to be founded on a strong comprehension of the mechanism by which various variables interface with and impact transportation phenomena (Vlahogianni et al., 2012).

Vlahogianni et al. (2010) used a Bayesian network approach to identify characteristics of primary incidents that affect the likelihood of secondary crashes. The observed traffic conditions at the time of an incident and the time required to respond to and clear the incident was identified as the most significant determinants in defining the upstream impact area of an incident (Vlahogianni et al., 2010).

Vlahogianni et al. (2012) developed a multi-layer perceptron neural network model to identify potential risk factors that may influence the occurrence of secondary crashes. As shown in Table 2-3, traffic-related, primary-incident-related, geometric-related, and

weather-related attributes were found to significantly impact the likelihood of secondary crashes. Contrary to other studies, Vlahogianni et al. (2012) found a negative relation between secondary crash likelihood and the number of lanes blocked due to the primary incident, severity of the primary incident, existence of a curved section at the location of the primary incident, and involvement of heavy vehicles.

Table 2-3: Summary of Literature on Non-Parametric Secondary Crash Risk Models

Reference	Secondary crash identification method	Method	Explanatory function	No. of var.	Significant variables
Vlahogianni et al., 2010	Method based on spatiotemporal impact area of primary crash	Bayesian Neural Network	Mutual information	8	Maximum queue length, queue duration, and primary crash duration
Vlahogianni et al., 2012	Automatic tracking of moving traffic jams	Bayesian Neural Network	Mutual information and partial derivatives	11	Traffic speed, changes in traffic speed and volume, duration of the primary crash, hourly volume, rainfall intensity, number of vehicles involved, blocked lanes, percentage of trucks, and upstream geometry
Park and Haghani, 2016	Data driven approach based on Gaussian Mixture Model and Bayesian structure equation model	A principles Bayesian learning approach to Neural Network and Logit model	Multilayer perceptron	13	Location area, incident type, and time of day
Park et al., 2017; 2018	Data driven approach based on Gaussian Mixture Model and Bayesian structure equation model	A principles Bayesian learning approach to Neural Network and Stochastic Gradient Boosted Decision Trees	A pedagogical rule extraction	13	Unexpected traffic congestion caused by a primary incident and onlooker factors

Note: No. of var. = Number of variables

Instead of using the conventional neural network models, Park and Haghani (2016a) proposed a principled Bayesian learning approach to neural networks to predict secondary crashes more accurately and robustly. A pedagogical rule extraction approach was used to improve the understanding of secondary crashes by extracting comprehensible rules from the neural networks. Unlike the neural network risk model proposed by Vlahogianni et al. (2012), Park and Haghani (2016a) used a sequentially predicted clearance duration to predict the probability of having a secondary crash. In addition to the Bayesian neural network approach, Park et al. (2018; 2017) used a Stochastic Gradient Boosted Decision Trees method to predict the probability of secondary crashes in real-time.

In general, regarding the prediction of secondary crashes, both parametric and non-parametric models were used to link the secondary crash risks with geometric, primary incident, weather, and traffic characteristics. In these studies, the features of geometric, weather, traffic conditions, and incident characteristics associated with primary incidents were compared with those of normal incidents. Understanding factors contributing to the occurrence of secondary crashes will provide some useful managerial tools for alleviating the effects of primary incidents, and thus, reduce the likelihood of secondary crashes.

2.2.2 Issues Accompanying Modeling of Secondary Crash Risk

Modeling the risk of secondary crashes is accompanied by several challenges. The infrequent nature of secondary crashes is one of the significant issues that needs to be addressed when modeling the risk of secondary crashes. Selection of the most important variable, detection of variable correlation, use of more representative traffic variables, and

missing information are among the issues encountered with explanatory variables used in secondary crash risk models. The following subsections discuss these issues in detail.

2.2.2.1 Imbalanced Data

As indicated earlier, secondary crashes are infrequent in nature. A majority of secondary crash risk models developed using either logit or probit link functions are symmetrical, i.e., the likelihood of secondary crash occurrence is presumed to rise to a probability of 0.5, then decrease toward the asymptote at one (1). In other words, in secondary crash likelihood prediction, symmetric models, such as logit or probit models, yield more reliable results when the proportion of normal incidents (~50%) is equal to the proportion of primary incidents (~50%). However, secondary crashes account for less than 20% (Owens et al., 2010; Sando et al., 2019) of total incidents, meaning that the proportion of primary incidents is much less than the proportion of normal incidents (i.e., primary incidents and normal incidents are asymmetrically distributed).

To account for the imbalanced nature of the response variable in a secondary crash risk model, Yang et al. (2014b) introduced the rare-event logistic regression model, and Kitali et al. (2019b) used a Synthetic Minority Over-sampling TEchnique-Nominal Continuous (SMOTE-NC) technique. Kitali et al. (2018) used a complementary log-log model (cloglog) as an alternative prediction model over the conventional logit and probit models. Unlike the logit and probit models, the cloglog model is asymmetrical with a fat tail as it departs from zero (0) and sharply approaches one (1) (Kitali et al., 2017; Martin and Wu, 2017).

2.2.2.2 Variables Selection

As indicated in Figure 2-6, researchers have considered several incident-related, traffic-related, temporal-related, geometric-related, and weather-related factors when developing secondary crash risk models. In actuality, it may not be possible to include all variables in the model due to the possible significant correlation among the factors. Moreover, the use of less important variables will introduce noise in the model and hence, reduce its accuracy.

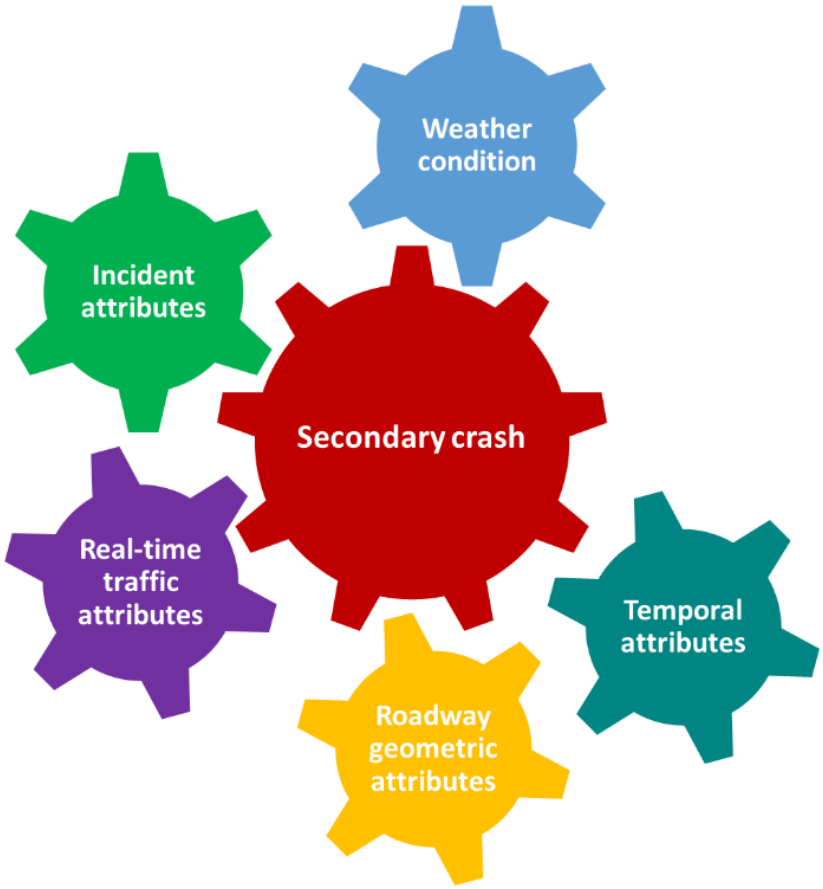


Figure 2-6: Factors Contributing to Secondary Crash Occurrence

One way to address this issue is to select and include only the most important variables. Variable subset selection methods, such as a stepwise technique, were used in

several studies to add one best-fit variable at a time during model fitting (Mishra et al., 2016; Xu et al., 2016; Zhan et al., 2009). Nevertheless, this criterion has several drawbacks, including the result that each addition of a new variable may render one or more of the already included variables non-significant. Also, because the stepwise variable selection process is discrete, it often exhibits high variance and may not reduce the full model's prediction error. In other words, small changes in the data can result in different variables being selected, and this can potentially reduce the model's prediction accuracy (Menard and Torelli, 2014; Tibshirani, 1996).

As an alternative to stepwise variable selection, Kitali et al. (2018) used random forests, a non-parametric approach, to select the most important variables for inclusion in the secondary crash risk prediction model. In a later study, Kitali et al. (2019b) applied the Least Absolute Shrinkage and Selection Operator (LASSO) penalized likelihood, a regression analysis method that performs both variable selection and regularization. The LASSO method enhances the prediction accuracy and interpretability of the statistical model (Tibshirani, 1996). LASSO shrinks some coefficients of a regression model, in this case, logistic regression, and sets others to zero (0) to obtain variables with a substantial effect on the outcome (Tibshirani, 1996). LASSO also performs important variable selection and variable correlation simultaneously. That is, between a pair of highly correlated variables, LASSO tends to pick the most important variable and discard the other by shrinking them towards zero.

Because the LASSO method performs variable selection through a continuous process, it does not suffer as much from high variability, i.e., it simultaneously does both continuous shrinkage and automatic variable selection. The penalty term introduced by

LASSO during the variable selection process ensures better estimation of the prediction error while avoiding overfitting. Selecting an optimal subset of explanatory variables is expected to improve the classification accuracy and make the model's interpretation easier. Since some of the variables will be minimized to zero, model thriftiness is achieved as well.

2.2.2.3 Use of Aggregated Traffic Flow and Weather Characteristics

Traditional traffic data, such as AADT and speed limit, have often been included as explanatory variables in secondary crash risk models (Chimba and Kutela, 2014; Khattak et al., 2012; Mishra et al., 2016; Zhang and Khattak, 2010). These data limit the reliability of results simply because they are aggregated values and do not reflect the prevailing traffic conditions at the time of an incident. With the availability of large-scale high-resolution traffic flow data in recent years, high-resolution traffic data, instead of AADT and speed limit, have been increasingly used in developing secondary crash risk prediction models (Kitali et al., 2018, 2019b; Park and Haghani, 2016a, 2016b; Sando et al., 2019; Vlahogianni et al., 2012; Xu et al., 2016). The high-resolution traffic flow data provides a more accurate measurement of traffic flow conditions before the occurrence of primary incidents and secondary crashes, compared with the traditional aggregated static traffic data, such as AADT and speed limit.

Xu et al. (2016) used the random-effect logistic regression to develop a secondary crash risk prediction model using the high-resolution traffic flow data before the occurrence of primary incidents. The results suggested that the inclusion of high-resolution traffic variables significantly increases the model's predictive performance. Traffic

volume, average speed, occupancy variation, and volume difference between adjacent lanes are the main traffic variables contributing to the increased risk of secondary crashes.

Inclement weather conditions, particularly rainfall, is one factor that could potentially exacerbate the occurrence of secondary crashes. Rainfall decreases the driver's sight distances and increases the vehicle's stopping distance (Haule et al., 2020; Kidando et al., 2019a). During rainy conditions, approaching vehicles may not have an adequate opportunity to make emergency maneuvers, leading to an increased possibility of secondary crashes (Li et al., 2014). It is imperative to incorporate weather conditions as one of the potential variables in the secondary crash likelihood model.

Previous research that included rainfall as one of the secondary crash influential factors obtained the data either from an incident database (Wang et al., 2016; Khattak et al., 2012, 2009; Xu et al., 2016; Zhan et al., 2008) or rain gauges (Kopitch and Saphores, 2011; Vlahogianni et al., 2012). Incident report-based rainfall data is qualitatively recorded by incident responders only once and mostly at the incident notification time. As such, this value of rainfall information may not reflect the prevailing rainfall intensity, especially in locations that experience short duration rainfall, when the incident impact duration is relatively long (Andrew, 2019). Gauge-based rainfall data are retrieved from weather stations that are usually sparsely distributed (Andrew, 2019). Similar to traffic flow characteristics, rainfall intensity varies both spatially and temporally. However, both incident-based and gauge-based rainfall data do not account for the spatiotemporal nature of rainfall.

2.2.2.4 Missing Potential Variables

While previous studies have considered numerous variables in secondary crash likelihood models, some variables have rarely been considered. Some of these variables include the presence of work zone, vertical curves, merging, and diverging ramps within the incident impact area. Work zone areas are associated with unexpected congestion due to a combination of factors, including daily changes in traffic patterns, narrowed rights-of-way, and complex arrangements of traffic control devices and signs (Federal Highway Administration [FHWA], 2007). These situations may influence the likelihood of secondary crashes. However, work zones were rarely considered in previous studies as one of the sources of the explained variation in the likelihood or severity of secondary crashes (Balke, 2009; Kopitch and Saphores, 2011; Yang et al., 2014b).

Unlike other roadway sections, merge and diverge influence areas are accompanied by more lane changes and high speed differentials by drivers attempting to enter or exit the freeway. This situation may increase the risk of secondary crashes. Thus, it is essential to incorporate merge and diverge influence areas in secondary crash risk models. Few studies have considered ramps as a potential variable that may influence the likelihood of secondary crashes (Karlaftis et al., 1999; Khattak et al., 2012, 2009; Park and Haghani, 2016b). Of those studies, the influence of ramps on secondary crash occurrence was not found significant.

In summary, researchers have used both parametric and non-parametric models to link secondary crash risks with geometric, incident, weather, and traffic characteristics. Understanding factors that contribute to the occurrence of secondary crashes will help

devise effective strategies to alleviate the effects of primary incidents, and thus, reducing the likelihood of a secondary crash.

2.3 Strategies to Mitigate Secondary Crashes

Mitigating the risk of secondary crashes is a crucial goal for effective traffic incident management. Deploying strategies that focus on clearing incidents as quickly as possible will have a significant impact on reducing the risk of secondary crashes. However, mitigation strategies may be challenging to deploy, due to limited available resources, e.g., patrol vehicles, personnel, traffic surveillance systems, etc. Moreover, each primary incident may occur during different conditions, resulting in various impacts. For example, an incident responder may be hindered by a long queue, thus delaying the process of incident clearance (H. Yang et al., 2018).

The variable speed limit control strategy is one of the countermeasures that has been explored by previous studies as an alternative to reduce the risk of secondary crashes. A variable speed limit is a mainline traffic control strategy that has been increasingly used for improving traffic safety on roadways (Zhao et al., 2019; Li et al., 2014). Introducing a variable speed limit when the risk of a secondary crash is high can help achieve the desired speed reduction to minimize hard-braking and high deceleration conditions that can lead to secondary crashes. Li et al. (2014) proposed using variable speed limits to reduce the risks of secondary crashes during inclement weather conditions. By analyzing the risk of secondary crashes, the variable speed limit strategy can adjust the speed limits according to the prevailing traffic and weather conditions. Based on safety surrogate measures, the

proposed variable speed limit system was found to reduce the risk of secondary crashes by 40-50 percent (Li et al., 2014).

Numerous studies have indicated incident duration as the most significant factor influencing the occurrence of secondary crashes (Kitali et al., 2018; Goodall, 2017; Wang et al., 2016; Zhan et al., 2009). Khattak et al. (2012) observed a significant correlation between incident duration, the likelihood of a secondary crash, and the primary incident characteristics. A 10-minute increase in the primary incident duration was found to be associated with a 0.2 percent increase in the likelihood of secondary crashes (Khattak et al., 2009).

Similarly, Goodall (2017) found the probability of a secondary crash occurrence to increase by approximately one (1) percentage point for each additional two to three minutes spent on the scene under congested traffic. Compared with other traffic incidents whose occurrences are quite stochastic, the occurrence of secondary crashes is more deterministic as they are mostly caused by either turbulent traffic conditions initiated by the primary incident or the onlooker effect (Xu et al., 2019). The impact of incident duration on the risk of secondary crashes was found to increase even further when traffic transitioned from a free-flow state to a congested state (Park and Haghani, 2016).

It is essential to prevent secondary crashes in advance with an effective prevention strategy (Park et al., 2018). A proactive secondary crash mitigation strategy can be implemented by disseminating advanced warning messages to inform upstream drivers of the potential secondary crash risk. The disseminated information will provide motorists with an opportunity to take necessary precautions to avoid being involved in a secondary crash, such as slowing down, changing lanes in advance, and/or diverting to alternate

routes. The upstream communication approach often consists of an incident warning, in addition to the speed advisory, which may increase the likelihood of driver compliance and minimize secondary crashes.

Before implementing an advanced warning strategy, the occurrence of secondary crashes has to be predicted in real-time. Recent studies, therefore, have relied on the use of high-resolution traffic data to identify and predict the likelihood of secondary crashes using different modeling approaches. However, the proposed models were developed and calibrated with fixed model parameters, which cannot account for the different traffic patterns with spatial and temporal variability. For instance, the prevailing traffic conditions before and following the occurrence of the incident may have a significant and varying impact on the likelihood of secondary crashes. Furthermore, the magnitude of the impact of the traffic flow characteristics on the likelihood of secondary crashes is expected to vary with time and space. Prediction of the risk of secondary crashes as a function of time and distance will increase the accuracy and efficiency of advanced warning strategies.

Several methods that can be used to broadcast warning messages to upstream motorists include: Dynamic Message Signs (DMSs) (Kopitch and Saphores, 2011); Advanced Traveler Information Systems (ATIS), such as Florida's FL511 service; navigation applications, such as Waze; and emerging technologies, such as Connected Vehicles (CVs) (Soloka 2019; Yang et al., 2017). The following subsections discuss these communication avenues to inform drivers upstream of a primary incident that may help to mitigate potential secondary crashes.

2.3.1 Dynamic Message Signs

DMSs are programmable devices that can display any combination of letters and/or symbols/graphics to deliver messages to motorists. They can provide real-time information and are used for traffic warnings, regulations, routing, and traffic management (Montes et al., 2008). Some messages provided by DMSs suggest a course of action to motorists, such as change travel speed, change lanes, or divert to a different route. Other messages may serve to inform motorists of changes in current or future traffic conditions (e.g., Congestion Ahead), or state regulations (e.g., *Buckle Up It's the Law*, *Click It or Ticket*, etc.).

DMS messages may reduce potential secondary crashes and downstream speed differentials by informing motorists of downstream traffic conditions (e.g., congestion caused by a crash) and encouraging safer driving (Chatterjee et al., 2002; Mounce et al., 2007). Kopitch and Saphores (2011) used the distance from the primary incident location to the nearest upstream DMS as a proxy to quantify the impacts of DMS messaging on secondary crash prevention. Specifically, this variable was included in the form of a quadratic function of the distance from the primary incident to the nearest upstream DMS. The DMS location was assumed to be at least two miles away from the primary incident for it to be effective (Kopitch and Saphores, 2011). The authors estimated the probability of secondary crash reduction within the DMS influence area at the 85% confidence interval (see Equation 2-1). The function in Equation 2-1 is negative between two miles and 22.3 miles, and it becomes increasingly negative from two miles to 11.5 miles, where it reaches its minimum. The function then increases between 11.15 miles and 22.3 miles, as shown in Figure 2-7. In other words, the effect of the incident information decreases with the increase in its propagation range. Although DMSs were found to influence the probability

of secondary crashes, this finding was not statistically significant (Kopitch and Saphores, 2011).

$$f(DMS) = -001002 \times DMS + 0.0045 \times DMS^2 \quad (2\text{-}1)$$

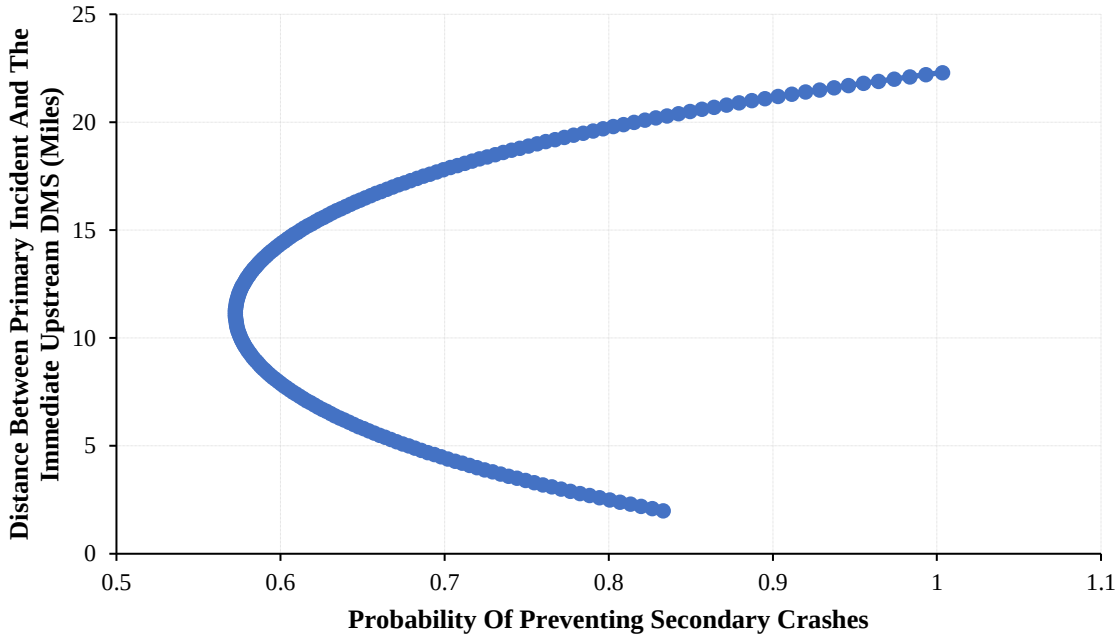


Figure 2-7: Impact of Dynamic Message Sign Messages on Secondary Crash Occurrence (Kopitch and Saphores, 2011)

2.3.2 Advanced Traveler Information Systems

In addition to DMSs, other platforms that could be used to disseminate proactive safety messages to upstream drivers include ATISs, such as Florida's FL511 service, navigation applications, especially those that leverage crowdsourced user reports for providing service, such as Waze, and CV technology. As illustrated in Figure 2-8, an ATIS can allow users to create and send highway advisory messages from their smartphone at the incident scene. The utilization of this correspondence innovation enables drivers to know what is happening on the road, alerts them in a split second about traffic conditions, incidents, police presence, construction, and even route change suggestions to save time

(Imani, 2019). The Waze platform has already been integrated into the SunGuide® software used by many TMCs for traffic management (Glotzbach, 2014). The incidents reported on Waze are linked directly to SunGuide® in real-time. Likewise, the Waze database collects the incidents reported in the SunGuide® system (Glotzbach, 2014).

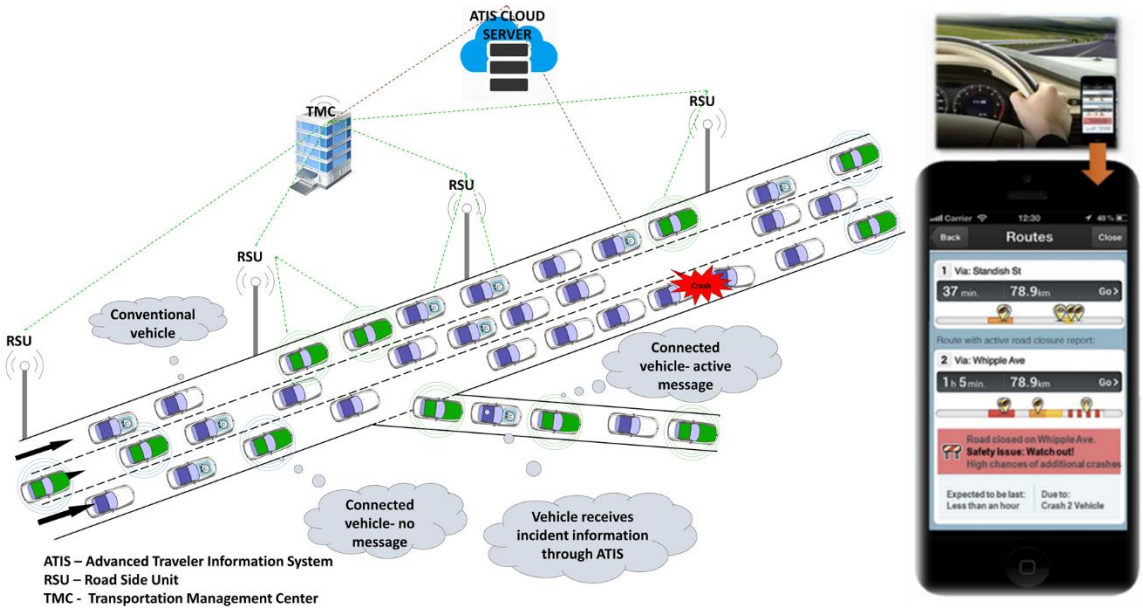


Figure 2-8: Application of Advanced Traveler Information System in Mitigating Secondary Crashes

A study by Amin-Naseri et al. (2017) evaluated the reliability, coverage, and added value of crowdsourced traffic incident reports from Waze in Iowa. The study concluded that the crowdsourced data stream from Waze is an invaluable source of information for broad coverage of traffic monitoring systems, covering 43.2% of Iowa's Advanced Traffic Management System (ATMS) crash and congestion reports. The Waze application also provided timely reporting, 9.8 minutes earlier than the probe-based alternative, on average, and with reasonable geographic accuracy. The Waze reports currently make significant contributions to incident detection and further complement the ATMS coverage of traffic conditions.

Given the emerging CV technologies, it is likely that many vehicles will soon connect with the surrounding infrastructure. CVs are equipped with certain technologies that help them communicate with their environment. This connected environment allows the CVs to communicate (i.e., send and receive messages) with other vehicles, known as V2V communication, as well as communicate with the surrounding infrastructure, known as vehicle-to-infrastructure (V2I) communication (Harding et al., 2014). Yang et al. (2017) explored the possibility of using CV technology for preventing secondary crashes by improving the situational awareness of drivers. A simulation-based modeling framework that enabled V2V communication was developed to assess connectivity's impact on the risk of secondary crashes. The results indicated that CVs could be a viable way to reduce the risk of secondary crashes. Secondary crash risk, measured by the number of simulated conflicts, was found to be significantly reduced if the market penetration rate of CVs on a highway was relatively high (e.g., 15%) in dense traffic conditions.

2.4 Summary

FHWA has established the reduction of secondary crashes as one of the performance measures for incident management programs. Proper identification of secondary crashes is pivotal to accurate reporting of the effectiveness of the programs deployed to mitigate secondary crashes. However, the limited knowledge of secondary crashes' nature and characteristics has largely impeded their mitigation strategies. The following subsections discuss the research gap pertaining to the identification of secondary crashes, understanding factors influencing the likelihood of secondary crashes, and the prediction of secondary crashes.

2.4.1 Challenges in the Identification of Secondary Crashes

Primarily three methods have been used to identify secondary crashes: (1) manual method; (2) static method; and (3) dynamic method. In the “manual” method, secondary crashes are manually identified by either TMC personnel or incident responders. In this case, the impact area of primary incidents is estimated visually based on the observer's judgment. However, it is subjective, unreliable, inconsistent, and random despite being the most used method.

Instead of relying on the manual method to identify secondary crashes, some studies defined the primary incident's impact area based on fixed spatiotemporal thresholds and detected secondary crashes within the predefined area. Although the static method is better than the manual method, the one-size-fits-all approach of using fixed spatiotemporal thresholds does not yield reliable results. This is because the primary incident's impact area heavily depends on the prevailing traffic conditions, i.e., uncongested or congested conditions. To overcome the limitations of the manual and static methods, recent studies have adopted a dynamic method. This method identifies the spatiotemporal thresholds flexibly based on the impact of the primary incident on traffic flow parameters. Although the dynamic method is proven to yield accurate and reliable results, applying it requires traffic data, which are only available at limited locations. To better identify secondary crashes, this approach needs to be able to distinguish non-congestion patterns from congestion patterns. Further consideration must be made to emulate how congestion conditions develop and disseminate.

2.4.2 Challenges in the Identification of Secondary Crash Influential Factors

After identifying secondary crashes, understanding the contributing factors is crucial to developing strategies to mitigate them. Both parametric and non-parametric models have been adopted for estimating the secondary crash likelihood. The response variable, which is the probability of a secondary crash, is modeled as a binary variable, given a primary incident or normal incident. Traffic flow characteristics, primary incident characteristics, weather conditions, and geometric characteristics have been considered as possible contributing factors. Primary incident characteristics and traffic flow characteristics have been observed to have a significant impact on the likelihood of secondary crashes.

Modeling the risk of secondary crashes has the following challenges: (1) accounting for the infrequent nature of secondary crashes; (2) selecting the most important variables with minimal correlation; (3) considering prevailing traffic conditions; and (4) including other potential variables that are rarely considered in the literature, e.g., presence of work zones, vertical curves, merging ramps, and diverging ramps within the incident impact area.

2.4.3 Challenges with Deploying Secondary Crash Mitigation Strategies

It is important to devise proactive strategies to promptly reduce the risk of secondary crashes because their occurrence is largely influenced by the severity of the primary incident and how quickly the incident is cleared. Previous research that explored strategies to mitigate secondary crashes used traffic data to identify and predict the likelihood of secondary crashes in real-time (Kitali et al., 2018; Xu et al., 2016). However, these studies neglected the influence of prevailing traffic conditions on the likelihood of a secondary crash following the occurrence of the initial incident.

CHAPTER 3
DATA NEEDS

This chapter discusses the data required to achieve the research goal. The first section provides a detailed discussion of the data used to accomplish the research goal. The second section describes the study location and the criteria used to select the study corridors. Finally, the third section summarizes the data needs.

3.1 Data Requirements

Four main types of data were required to achieve the research goal: (1) incident data; (2) high-resolution traffic data; (3) roadway geometric data, including work zone information; and (4) high-resolution rainfall data. Incident data were obtained from the SunGuide® database. High-resolution traffic data were retrieved from HERE Technologies, and work zone data were obtained from the FDOT Open Data Hub. Other roadway geometric characteristics were extracted from the Roadway Characteristics Information (RCI) database, Google Earth Pro, and Google Maps. High-resolution rainfall data were retrieved from the National Oceanic and Atmospheric Administration (NOAA) database. These data were collected for 5.5 years, from January 2014 to June 2019. The following subsections discuss each of these data sources.

3.1.1 SunGuide®

SunGuide® is an ATMS software used by the FDOT to process and archive incident data on freeways. The database stores incident attributes including incident identification (ID), roadway name, latitude and longitude of the incident location, incident notification

time, incident type, number and categories of responding agencies, lane closure information, incident severity, weather condition, and road surface condition.

The categories of incident events included in the SunGuide® database are crash, disabled vehicles, debris on roadway, emergency vehicles, police activity, vehicle fire, flooding, pedestrian, abandoned vehicles, construction, wrong-way driver, etc. For this study, these categories were further summarized into four groups: crashes, vehicle problems, hazards, and other events. Accordingly, the *crashes* group contained crash events. *Vehicle problems* included all events that were not crashes, but were vehicle-related, e.g., disabled vehicles, abandoned vehicles, etc. *Hazards* included all objects on the roadway with the potential of causing crashes, e.g., debris on roadway, wildlife, etc. *Other events* encompassed all events that do not fit in the three aforementioned event categories, e.g., other, bridge work, amber alert, wrong-way driver, etc. These event types were excluded from the analysis.

Incidents that occurred on ramps also were not included in the analysis. Compared to mainline segments, ramps have a complex geometry that significantly affects the traffic transition states, i.e., from free-flow to breakdown, congested, recovery, and eventually back to free-flow. For this reason, incidents occurring on ramps require a separate analysis approach (Sando et al., 2019).

3.1.2 HERE Technologies

HERE Technologies record the space mean speed for roadways by dividing them into segments referred to as Traffic Message Channels. As discussed in detail in Chapter 4, the 5-minute speed data from HERE Technologies were first used to identify secondary

crashes. Next, speed data (i.e., mean and standard deviation (SD)) in the Traffic Message Channel where the incident occurred and within 10 minutes before the occurrence of the incident were collected to capture the traffic conditions before the occurrence of the incident. To determine the prevailing traffic conditions, speed data within the Traffic Message Channels impacted by the incident, from the time the incident was detected to the time when the traffic flow returned to normal, were used. Since the incident impact duration along different Traffic Message Channels may differ, the incident impact area was individually defined for each Traffic Message Channel.

3.1.3 Roadway Geometric Characteristics and Work Zone Data Sources

Roadway geometric characteristics that may significantly impact traffic flow characteristics were considered potential variables that may influence the risk of secondary crashes. The following geometric variables were considered: shoulder width, horizontal curves, vertical curves, merging segment, and diverging segment. Other potential geometric variables that were considered in this study include service plazas and toll plazas. Since there are few service plazas and toll plazas within the study area, these variables were excluded from the analysis.

Shoulder width, horizontal curve, and vertical curves variables were collected from the RCI database for the years 2014 through 2019. The shoulder width variable was derived for the outside shoulder located adjacent to the outside travel lane. Outside shoulders provide for the accommodation of stopped vehicles, emergency use, and lateral support of the roadbed (FDOT, 2016). Since the entire roadway section considered in this study has a median, the shoulder width variable was collected from two roadsides. The final shoulder

width corresponding with each incident was calculated as a weighted value of all the shoulder widths within the incident impact area:

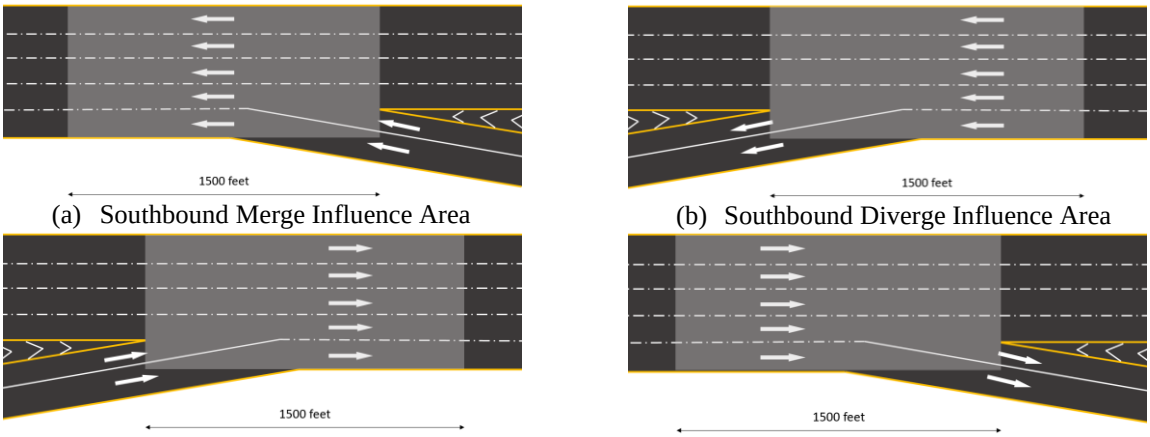
$$Weighted\ shoulder\ width = \frac{\sum_i^n Shoulder\ width_i \times Incident\ impact\ area_i}{Total\ incident\ impact\ area} \quad (3\text{-}1)$$

where, $Shoulder\ width_i$ is the shoulder width within the incident impact area and $Incident\ impact\ area_i$ is the portion of the incident impact area with $Shoulder\ width_i$. The subscript i represents the different shoulder width values within the incident impact area.

The horizontal curve variable was aggregated into two categories: incidents with a horizontal curve within their impact area and incidents without a horizontal curve within their impact area. Similarly, the vertical curve variable was aggregated in the same manner as the horizontal curve.

The merge and diverge influence areas were derived from Google Earth Pro and Google Maps using the Historical Imagery and the Street View tools. The Historical Imagery tool was used to verify the location of the identified ramps during the study period. The merge and diverge influence areas were defined based on the Highway Capacity Manual (HCM) (Transportation Research Board [TRB], 2016). A *merge influence area* spans from the point where the edges of the travel lanes of the merging roadways meet to a point 1,500 feet downstream of that point. Similarly, a *diverge influence area* spans from the point where the edges of the travel lanes of the merging roadways meet to a point 1,500 feet upstream of that point. While the HCM defines the ramp influence area as one that includes only lanes 1 and 2, in this research, both merge and diverge influence areas cover

the entire roadway section (i.e., all travel lanes) since they are measured within the impact area of an incident (see Figure 3-1).



(a) Southbound Merge Influence Area (b) Southbound Diverge Influence Area

(c) Northbound Merge Influence Area (d) Northbound Diverge Influence Area

Figure 3-1: Definition of Merge and Diverge Influence Areas

The final merge/diverge influence area considered in this research was also a dichotomous variable, like the horizontal curve variable. That is, incidents with a merge/diverge influence area within their impact area were grouped into the "*yes*" category and incidents without merge/diverge influence area within their impact area were categorized as "*no*". Note that the "presence of merge influence area" and the "presence of diverge influence area" were treated as separate variables.

The work zone activities data were retrieved from the Active Construction Project database service that is updated nightly in the FDOT Open Data Hub. The database provides the work zone construction location and duration. The Google Map Historical Tool was used to verify the direction where the construction activity was reported. Using this information, the work zone variable was aggregated into two categories: incidents with work zone activity within their impact area (i.e., the "*yes*" category), and those without work zone activity within their impact area (i.e., the "*no*" category).

3.1.4 NOAA Database

The NOAA database preserves, monitors, and assesses climate and historical weather data. One of the systems maintained by NOAA is the Next Generation Weather Radar (NEXRAD). NEXRAD is a network of 160 high-resolution Doppler radar sites that detect precipitation and atmospheric movement and disseminate near real-time data in approximately 5-minute intervals from each site (Barr, 2018). With these high-resolution data, it is possible to obtain the actual rainfall intensity over the road network in short time intervals.

Original data from NEXRAD, referred to as NEXRAD Level-II data, were used in this research. These data included reflectivity, one of the meteorological base data quantities. Radar measures rainfall intensity using radiations reflected on a target surface, in this case, a roadway network. The proportion of a target's productivity in capturing and returning radiofrequency energy is alluded to as reflectivity. Reflectivity can simply be defined as a measure of fractions of radiations reflected by a given surface. It is expressed as the ratio of the radiant energy reflected and the total amount of energy incident upon that surface (Andrew, 2019).

As indicated in Figure 3-2, in this research, reflectivity data were downloaded from the radar located in Miami, Florida (KAMX – Miami, FL). This radar is positioned at latitude: 25.61056, longitude: -80.41306, and has been operational since April 20, 1995. Specifically, the NEXRAD Level-II data were accessed from Amazon S3 through the following link https://noaa-nexrad-level2.s3.amazonaws.com. Similar to other high-resolution Doppler radars under NEXRAD, the KAMX radar covers a 248.5-mile radius.

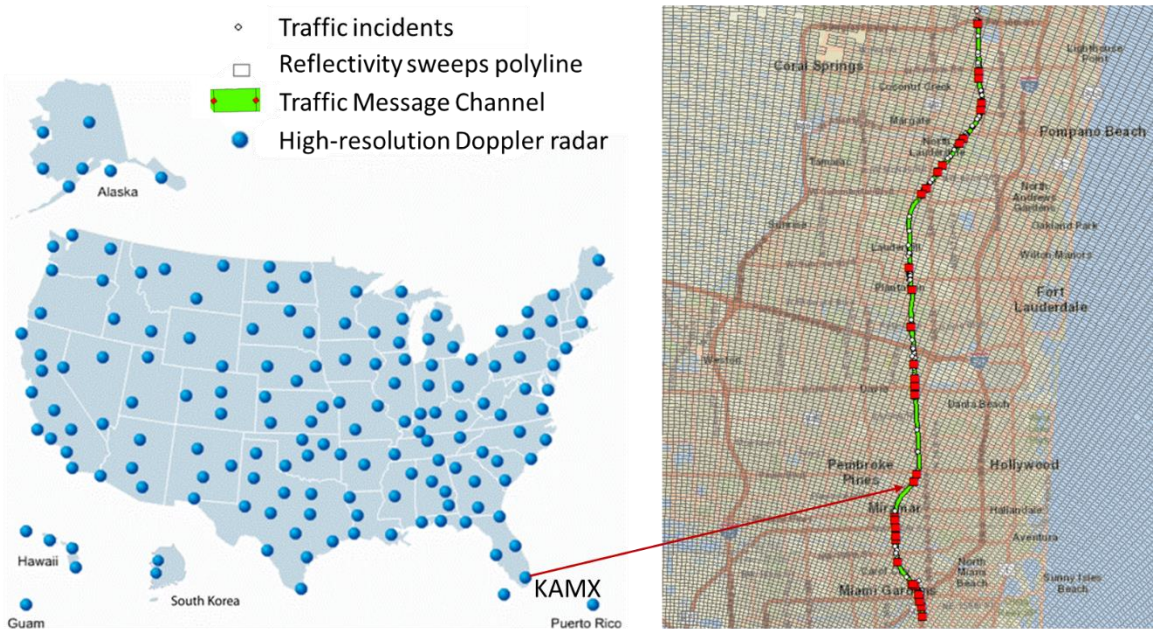


Figure 3-2: Location of Radar used to Collect Rainfall Data (NOAA, n.d.)

Figure 3-3 describes the approach used to retrieve rainfall data from NEXRAD. Reflectivity data were obtained for incidents that occurred during inclement weather conditions, as indicated in the incident database. The data were retrieved at 5-minute intervals, from the time when the incident began impacting traffic to the time when (1) a secondary crash occurred for primary incidents, and (2) when the traffic flow returned to normal for normal incidents. The downloaded radar data from Amazon S3 are in a unique digital binary format. Thus, as indicated in Figure 3-3 (Step 2), the NOAA Weather Climatic Toolkit (WCT) was used to visualize and convert data into a conventional scientific format, a shapefile in this case. ArcGIS software was then used to merge the downloaded radar data with the Traffic Message Channels impacted at time interval (*t*).

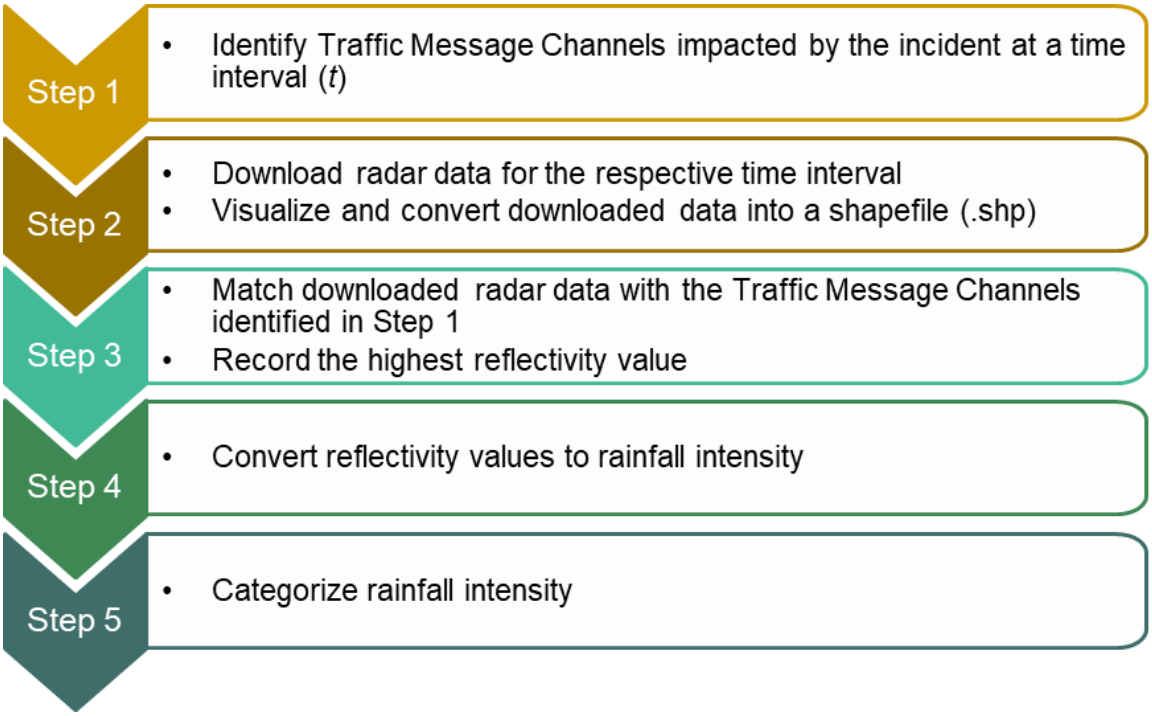


Figure 3-3: Workflow for Collecting and Processing Reflectivity Data

The recorded reflectivity values were converted to rainfall intensity using the following reflectivity-rainfall intensity relationship (Andrew, 2019):

$$R = \frac{10^{\frac{dBZ^{\frac{1}{2}}}{10}}}{250} \quad (3\text{-}2)$$

where, R is the rainfall intensity expressed in millimeters per hour (mm/hr), and dBZ is an abbreviation for decibel relative to reflectivity (Z). The dBZ is used to compare the reflectivity of a target surface in mm^6 per m^3 to the return of a droplet of rain with a diameter of 1 mm. In other words, it measures the strength of the energy reflected to the radar by the target surface, in this case, the roadway segment. Finally, the rainfall intensity data were grouped into three categories according to the American Meteorological Society (AMS) rainfall intensity classification (American Meteorological Society [AMS], n.d.). The three groups include light rainfall (Trace – 0.10 in/hr), moderate rainfall (0.10 – 0.30

in/hr), and heavy rainfall (> 0.30 in/hr). Table 3-1 shows a sample of rainfall data retrieved from KAMX radar in June 2019. The near- high-resolution rainfall data were obtained from the NOAA database.

Table 3-1: Sample Rainfall Data from NEXRAD

Sweep Time	Rainfall (mm/hr)	Rainfall (in/hr)	Rain Category
10:30:14 AM	0.010043937	0.000395431	Heavy
10:35:54 AM	0.034947831	0.0013759	Heavy
10:41:24 AM	0.01003853	0.000395218	Moderate
10:46:55 AM	0.002384304	9.38703E-05	Moderate
10:52:50 AM	0.011050536	0.000435061	Heavy
10:58:28 AM	0.010038571	0.00039522	Light
11:04:35 AM	0.007661987	0.000301653	Heavy

3.2 Study Area

The study corridors were selected from the Florida's Turnpike System Mainline. As shown in Figure 3-4, the Turnpike Mainline is a 312-mile corridor consisting of two main roadways: the Florida Turnpike Mainline (or SR-91), and the Homestead Extension of Florida's Turnpike (HEFT) (or SR 821). The two roadways are 265 mile and 48 miles, respectively. The following subsections discuss the criteria considered while selecting the study corridors.

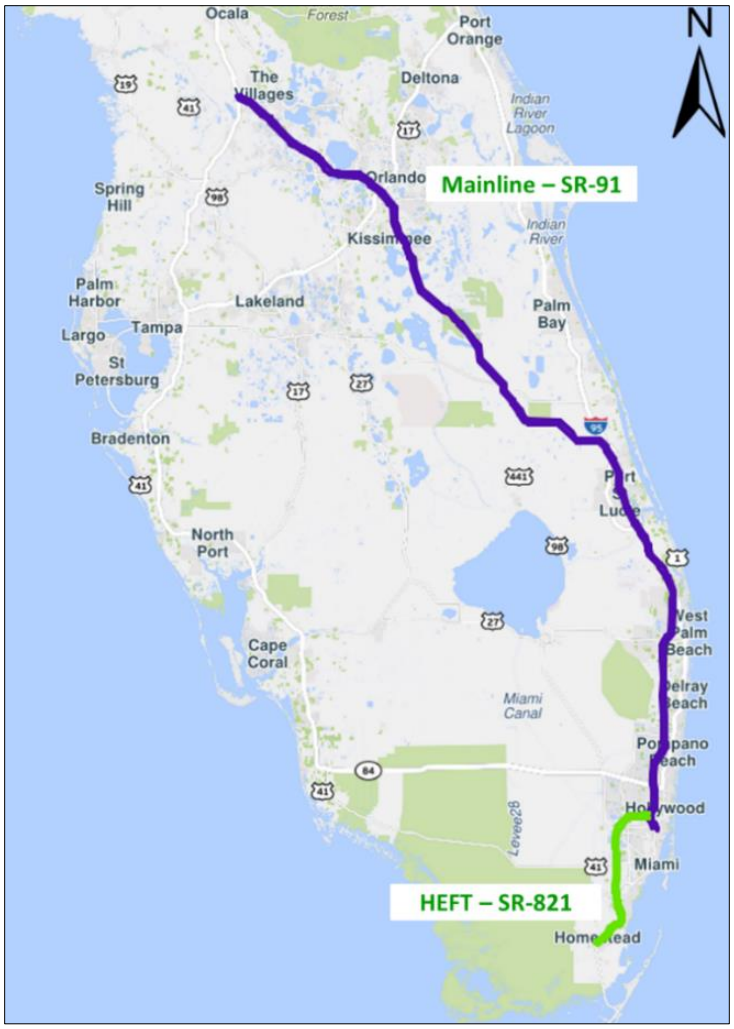


Figure 3-4: Florida's Turnpike Mainline

3.2.1 Study Corridors for Secondary Crash Identification

Two main data sources are required to estimate an incident impact area: (1) traffic incidents; and (2) high-resolution traffic data. The HERE Technologies record the speed for roadways by dividing them into Traffic Message Channels. Traffic Message Channels generally span a stretch from one exit or entrance ramp to the next.

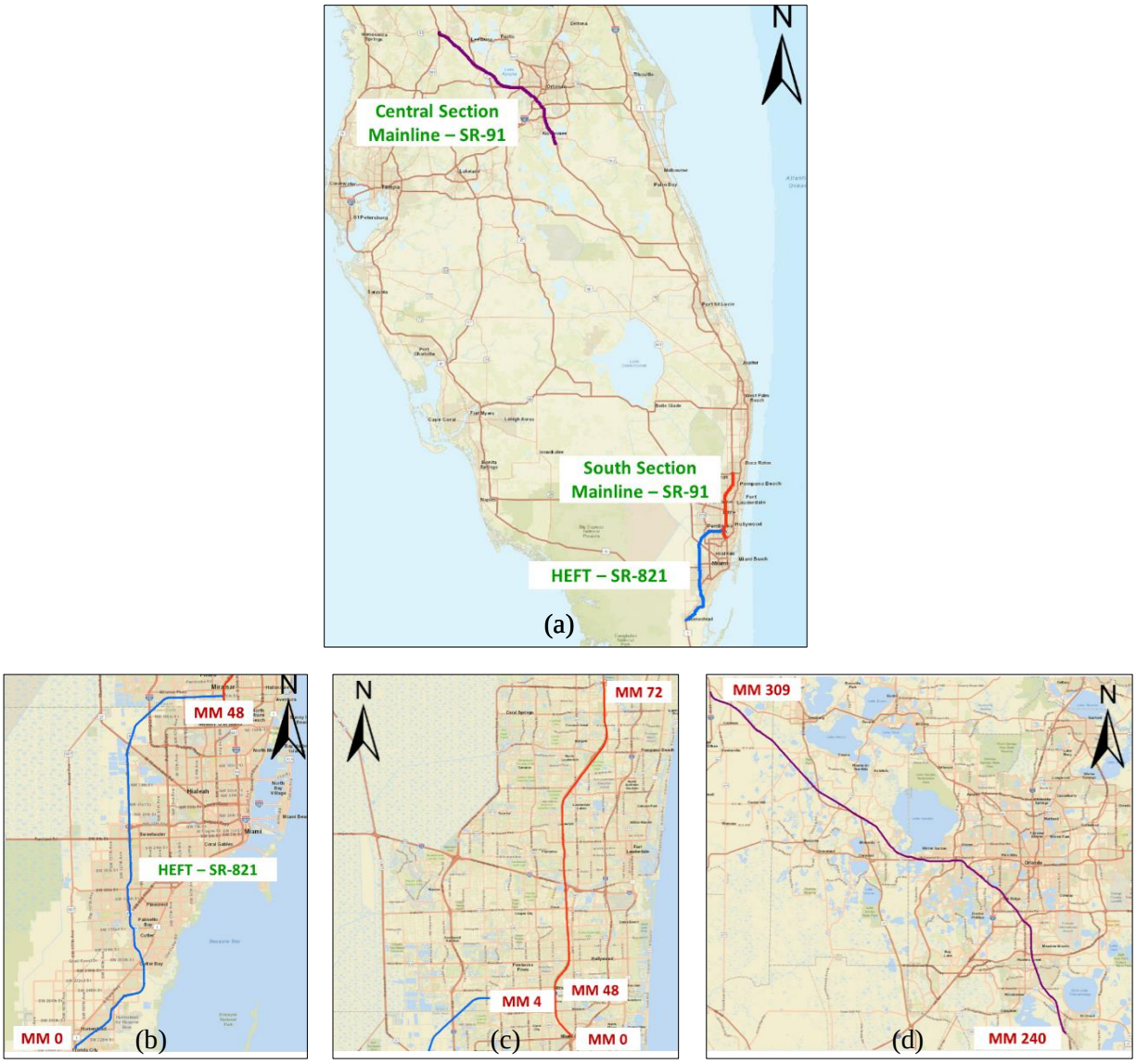
There are a total of 406 Traffic Message Channels along the Florida's Turnpike Mainline, with 284 along the Mainline, and 122 along the HEFT. On average, Traffic

Message Channels along the study corridor span a distance of 1.9 miles and 0.7 miles on the Mainline and HEFT, respectively. About 65% of the Traffic Message Channels along the Mainline are shorter than 1.4 miles. On the other hand, 88% of Traffic Message Channels along the HEFT are shorter than 1.5 miles. Only 7% of the Traffic Message Channels along the HEFT are longer than 2 miles.

Since the longest Traffic Message Channel along the HEFT is approximately 4 miles, and the longest Traffic Message Channel along the Mainline is 15 miles, a minimum 4-mile length was considered as a criterion to include a Traffic Message Channel in the analysis. Notably, only 15% of the Mainline Traffic Message Channels are longer than 4 miles. The use of traffic data from overly long Traffic Message Channels may result in an inaccurate estimation of traffic flow changes caused by the incident.

The final study area included the full 48-mile length of the HEFT and a 97-mile section along Florida's Turnpike Mainline. The 97-mile section included a 69-mile section of the Mainline Central Section (MCS), and 28 miles of the Mainline South Section (MSS).

Figure 3-5(a) shows the location of the study area. The HEFT section is from mile marker (MM) 0 to MM 48 (see Figure 3-5(b)). The MSS is located from MM 0 through MM 4, which is the Turnpike Mainline Spur, and from MM 48 through MM 72, which is the junction between SR-91 and SR-869 (Sawgrass Expressway) (see Figure 3-5(c)). The MCS is located from MM 240 through MM 309 (see Figure3-5(d)). Table 3-2 summarizes the HERE Traffic Message Channels along the selected study corridors.



HEFT | Mainline South Section (MSS) | Mainline Central Section (MCS)

Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, OpenStreetMap contributors, and the GIS User Community.

Figure 3-5: Selected Roadway Sections along Turnpike Mainline

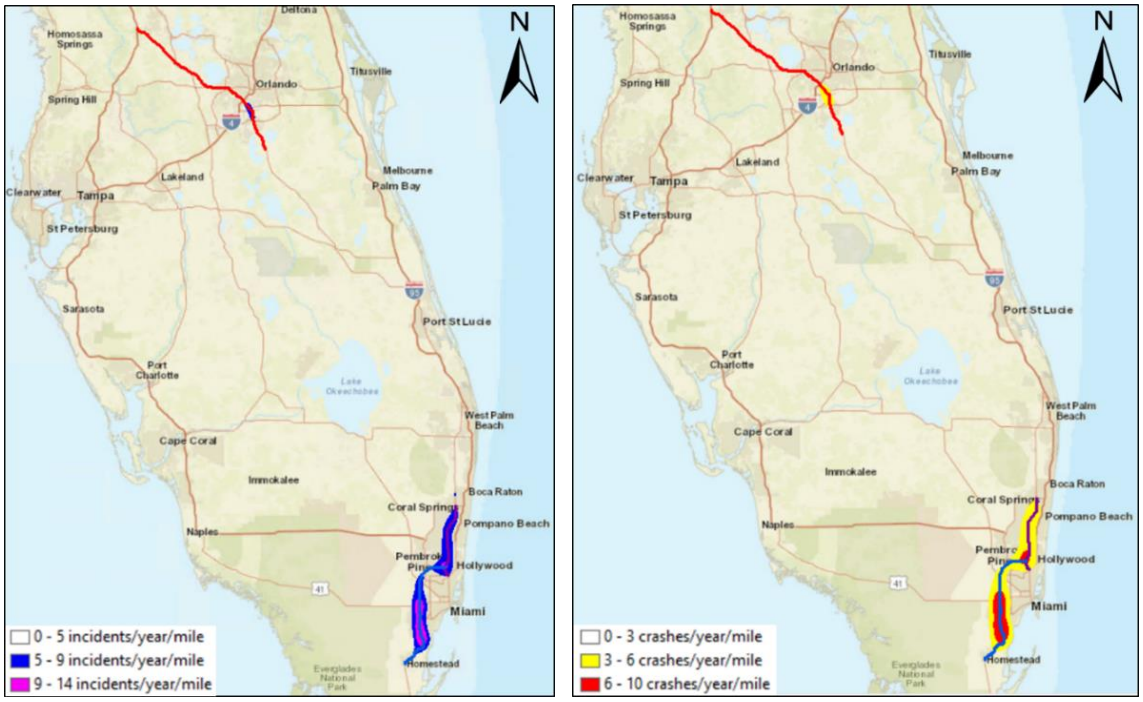
Table 3-2: Distribution of HERE Traffic Message Channels along the Study Corridors

Roadway	Number of Traffic Message Channels			Length of Corridor (miles)
	Northbound	Southbound	Total	
Mainline Central Section	46	47	93	69
Mainline South Section	34	35	69	28
HEFT	61	61	122	48

3.2.2 Study Corridors for Secondary Crash Likelihood Model

Of the three corridors used to identify secondary crashes, only the HEFT and the MSS were used to develop the secondary crash likelihood models. These corridors are located within the same jurisdiction (Miami, Florida) and serve traffic with comparable driving behaviors, patterns, and volume. Also, the group of incident responders that attend to incidents on the two corridors is similar.

The two corridors, HEFT and MSS, were selected based on the availability of speed data from HERE Technologies, incident hotspots, and major construction activities, such as lane widening, bridge maintenance, etc. The Kernel Density function in ArcGIS was used to identify high incident segments within the Florida Turnpike System. The hotspot analysis was conducted based on traffic incidents that occurred along the study corridors during the study period. As indicated in Figure 3-6, both the incidents hotpot analysis and crash hotspot analysis identified the HEFT and MSS as corridors that experienced the highest number of traffic incidents and crashes.



(a) Incident hotspot (b) Crash hotspot

Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, OpenStreetMap contributors, and the GIS User Community.

Figure 3-6: Corridors with High Incidents along Florida's Turnpike

3.2.3 Study Corridors for Secondary Crash Risk Prediction Model

Of the two corridors used in the likelihood model, only MSS was used to implement the prediction model. The exploratory analysis of the Active Construction Projects shapefile indicated that lane widening construction activities took place along the HEFT during the study period. Alternatively, on the MSS, there were no such activities during the study period.

3.3 Summary

The goal of this research was to investigate approaches to mitigate secondary crashes on freeways. This goal was implemented using the following three main steps:

1. identify secondary crashes;

2. identify significant factors contributing to the occurrence of secondary crashes; and

3. predict the probability of secondary crashes in real-time.

Table 3-3 summarizes the data needs for each of the tasks required to achieve the goal of this research.

Table 3-3: Data Needs for Predicting Secondary Crashes in Real-time

Data Source	Data Type	Identify SC	SC Likelihood Model	SC Prediction Model
SunGuide®	Incident			
HERE Technologies	Speed			
RCI	Shoulder width, horizontal curve, and vertical curve			
Google Maps	Merge ramps, diverge ramps, and work zone data			
Google Earth Pro	Merge ramps, diverge ramps, and work zone data			
FDOT Open Data Hub	Work zone data			
NOAA	Rainfall intensity			

Note: FDOT = Florida Department of Transportation; RCI = Roadway Characteristics Information; NOAA = National Oceanic and Atmospheric Administration; SC = Secondary Crash

CHAPTER 4
METHODOLOGY

This research explored approaches to mitigate secondary crashes on freeways. This goal was achieved using the following three components: (1) identify secondary crashes using a dynamic approach, (2) identify factors influencing the likelihood of secondary crashes, and (3) develop a real-time dynamic secondary crash risk prediction model. This chapter discusses the methodology and data preparation efforts used to achieve the research goal and objectives.

4.1 Identify Secondary Crashes

A data-driven approach was used to identify secondary crashes in this research. This method accurately estimates the impact area of the primary incident using speed data from HERE Technologies and identifying secondary crashes occurring within the impact area of the primary incident. The proposed approach aims to capture better traffic flow characteristics, such as speed, that change over space and time and affect the queue formation caused by the primary incident. As discussed in Section 3.2, the study area included the HEFT corridor, a 48-mile extension of the Florida Turnpike, and a 97-mile section on Florida's Turnpike System Mainline, i.e., a 69-mile Mainline Central Section (MCS) and a 28-mile Mainline South Section (MSS). This research used three major steps to identify secondary crashes using the proposed data-driven approach.

4.1.1 Extract and Process Speed Data from HERE Technologies

The 5-minute speed data from HERE Technologies were retrieved from the 284 Traffic Message Channels along the study corridor from January 2014 through June 2019.

These data were used to establish the recurrent speed profile of the section under normal traffic conditions. The average speed in each 5-minute interval was used to establish the recurrent speed profile. Additionally, a confidence interval of two standard deviations was established to define the speed profile's lower and upper bounds (i.e., speed bandwidth) to account for the variation in speeds on a roadway segment. For each Traffic Message Channel, seven speed profiles were generated, one for each day of the week. Independent speed profiles for different days of the week and times of the day were established to account for the recurrent traffic congestion. Figure 4-1 shows a typical speed profile for a 24-hour period on a weekday. As expected, there is a significant drop in speed during the morning peak hours, while the average speeds were the highest between midnight and 5:00 AM.

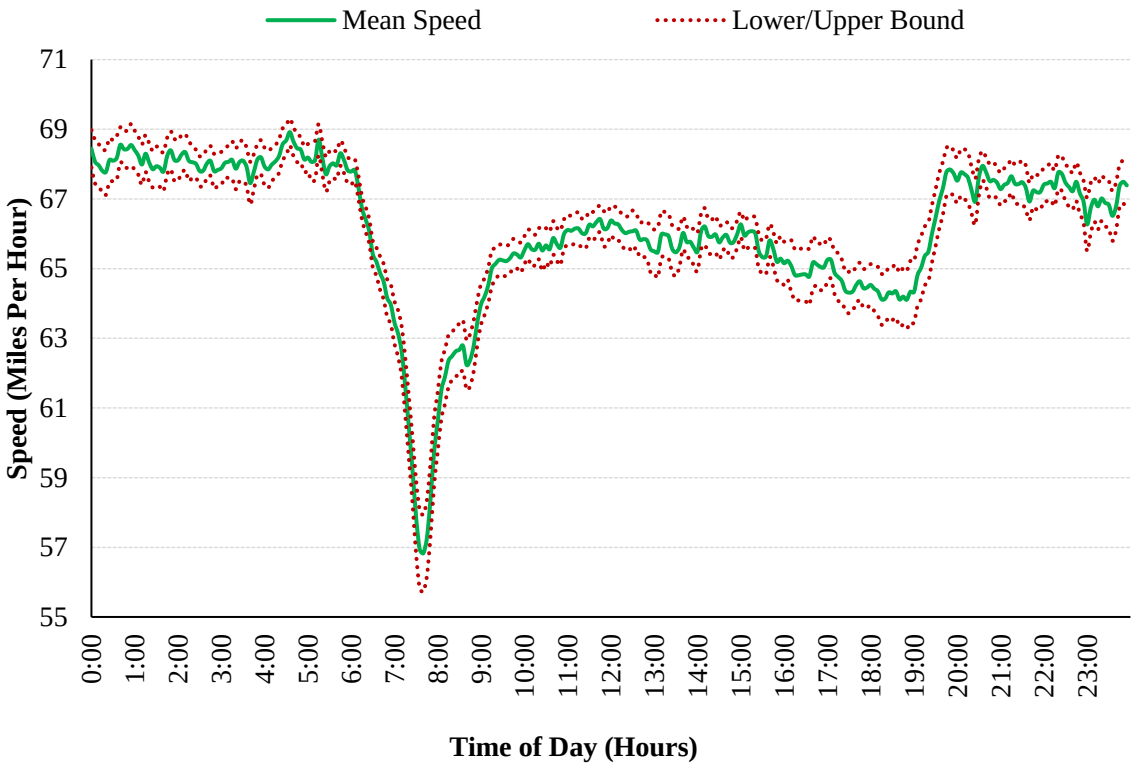


Figure 4-1: Sample Speed Profile for Estimating Normal Traffic Conditions

4.1.2 Match Incidents to a Traffic Message Channel

The geographic location of both the incidents and the Traffic Message Channels is the most critical information required for matching an incident to the Traffic Message Channel. In this research, the mile markers (MMs) of incidents and Traffic Message Channels (start and end) were used instead of the geographic coordinates, i.e., longitudes and latitudes. Through the ArcGIS tool, the *Toll Roads polyline shapefiles* extracted from the FDOT Transportation Data and Analytics Office website were used to assign MMs to the incidents and the start and end of the Traffic Message Channels. This approach ensures that roadway alignment characteristics, especially on curved segments, do not affect the accurate computation of the spatial relationship between incidents and Traffic Message Channels.

Using the assigned MMs, each incident was matched to a Traffic Message Channel at the incident location. For northbound incidents, since MMs increase in the northbound direction, the MM of the northbound incident must be greater than or equal to the MM of the start of the Traffic Message Channel and less than or equal to the MM of the end of the Traffic Message Channel. Similarly, for southbound incidents, since MMs decrease in the southbound direction, the MM of the incident must be greater than or equal to the MM of the start of the Traffic Message Channel and less than or equal to the end of the Traffic Message Channel. In other words, the start and end of each Traffic Message Channel is direction dependent. The date, day, and reported time of incidents that were successfully matched with the Traffic Message Channels were extracted and used in the next steps.

4.1.3 Estimate Incident Impact Area and Identify Secondary Crashes

Traffic incidents and real-time traffic data were required to estimate the incident impact area. The impact area was computed for incidents that were successfully matched to the Traffic Message Channels, as elaborated in the previous step. This process was achieved by tracking the reported speeds at the segment where the incident occurred, from the time the incident was detected to the time when the traffic flow returned to normal. An incident was considered to have affected the traffic flow characteristics of the segment when the average speed along the segment was less than the lower boundary of the speed profile. The same procedure was repeated for all the upstream Traffic Message Channels affected by the incident. Next, the time taken for the traffic to return to normal, following the occurrence of an incident, was recorded for each affected Traffic Message Channel. Since the incident impact duration along different Traffic Message Channels may differ, the incident impact area was defined for each Traffic Message Channel individually.

In summary, this process enabled the accurate estimation of the spatiotemporal impact area of the incident. That is, for each impacted Traffic Message Channel, the temporal thresholds were defined by the incident impact duration, i.e., from the time the incident was first detected to the time traffic returned to normal.

Figure 4-3 shows an example of the impact area caused by an incident **I-1**, where the x- and y-axes represent the time and length of the affected roadway segments, respectively. Note that each cell in Figure 4-3 represents a speed measurement by the Traffic Message Channel at the t^{th} time interval, i.e., 5 minutes in this case. As indicated in Figure 4-2, the impact duration and impact length vary across the different Traffic Message Channels impacted by the incident. While the segment where the incident occurred, i.e.,

Traffic Message Channel 0, has the most extended impact duration, the farthest segment impacted by incident **I-1**, i.e., Traffic Message Channel 6, has the shortest impact duration.

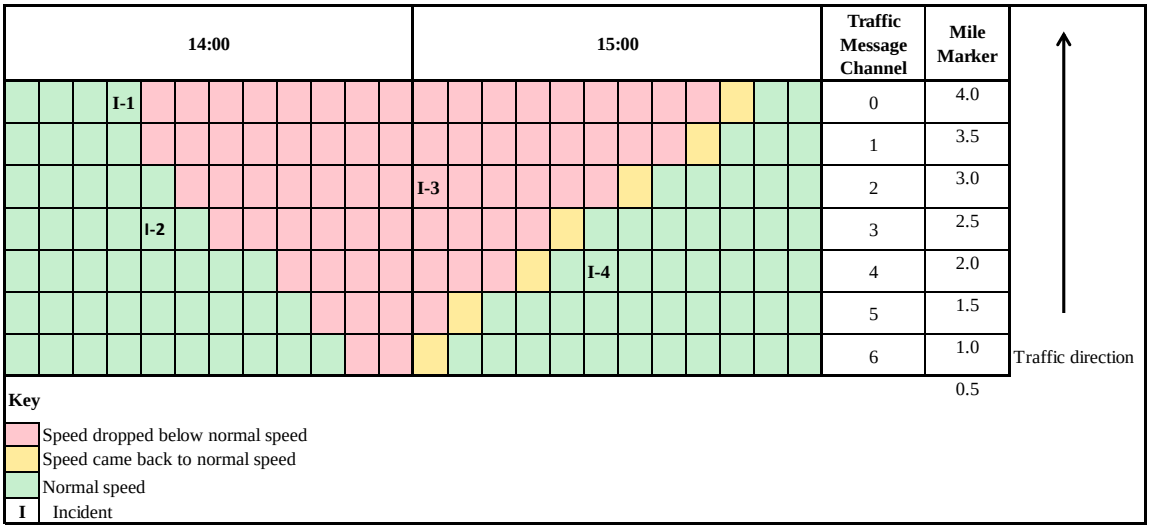


Figure 4-2: Approach to Estimate Incident Impact Area

Following the establishment of the area impacted by each incident, the last step was to identify secondary crashes. A traffic incident was considered a secondary crash if it occurred within the prior incident's spatiotemporal impact area. Referring to Figure 4-2, since incident **I-1** occurred earlier than incidents **I-2, I-3**, and **I-4**, the main task was to determine whether these three incidents occurred because of incident **I-1**. Considering the impact area in Figure 4-3, incident **I-3** was considered a secondary crash to incident **I-1** since it occurred within the impact area of incident **I-1**.

4.2 Identify Factors Influencing the Occurrence of Secondary Crashes

Not all incidents lead to secondary crashes. The likelihood of secondary crashes depends on several factors, including characteristics of primary incidents, weather conditions, geometric conditions, traffic flow characteristics, etc. Besides these factors, work zones have the potential of causing secondary crashes (Kitali, 2019b). However,

simply incorporating all variables in the model may lead to biased results, considering the possible significant correlation that exists among the variables. The proportion of primary incidents is normally lower than the proportion of normal incidents, a situation that makes the response variable in the likelihood model imbalanced. Thus, a modeling approach that accounts for the response variable's imbalanced nature, identifies the most important variables, and detects variable correlation was considered an ideal model for this case. In addition to addressing these issues, this research explored the influence of variables that were rarely considered in previous studies on the risk of secondary crashes. The explored variables include the presence of work zone, merge influence area, diverge influence area, and vertical curves within the incident impact area.

Occasionally, secondary crashes tend to become primary incidents for other crashes, conventionally referred to as cascading crashes. In other words, some primary incidents result in a series of cascading crashes. Although generally uncommon, cascading crashes present a significant challenge to transportation agencies. They are expected to be attended by multiple responding agencies at different time stamps and locations. Moreover, incidents attended to by multiple incident responders may require lane closures, a situation that further reduces the capacity of the roadway resulting in more congestion. Identifying factors associated with the likelihood of cascading crashes is the first step towards devising strategies to mitigate them. The following subsections discuss the methodologies used to identify factors influencing the likelihood of secondary crashes and cascading crashes.

4.2.1 Identify Factors Influencing the Likelihood of Secondary Crashes

As indicated earlier in this dissertation, the research was based on data collected for two corridors, the HEFT and the MSS. Unlike the MSS corridor, lane widening activities were occurring along the HEFT during the study period. Considering the scarcity of studies that evaluated the impact of work zones on the occurrence of secondary crashes, this research extends the previous research on secondary crash likelihood models by evaluating the impact of work zones on the occurrence of secondary crashes.

Instead of using a conventional logistic regression, the adaptive LASSO penalized logistic regression, fitted using the bootstrap resampling approach, was used to model the likelihood of secondary crashes in work zones. Specifically, the adaptive LASSO penalized estimator was used to extract the most important variables, with minimal correlation. Since the proportion of primary incidents was smaller than the proportion of normal incidents, the bootstrap resampling method was used to fit the penalized logistic regression. The following subsections describe in detail the penalized logistic regression and the bootstrap resampling approach.

4.2.1.1 Penalized Logistic Regression

Secondary crash risk models estimate the probability that a secondary crash will occur given a prior incident. From a statistical point of view, secondary crash risk modeling can be viewed as a binary classification problem. Suppose that the dataset for incidents has n observations $(\boldsymbol{X}^i, y_i), i\epsilon 1,2,\dots,n,$ with p explanatory variables, then $\boldsymbol{X}^i = x_{i1}, x_{i2}, \dots, x_{ip} = x_i^T$. Let $y = (y_1, \dots, y_n)^T$ becomes the response variable, which is binary in nature, i.e., y_i represents the secondary crash indicator (1 indicates a secondary crash is caused by a primary incident (i), and 0 indicates that no secondary crash occurred).

Researchers have used several methods to identify factors influencing the risk of a secondary crash. Of the previously adopted methods, logistic regression has an exceptional advantage since it provides a direct estimate of class probability and does not require a tuning parameter. As shown in Equation 4-1, the logistic regression model presents the class-conditional probabilities through a linear function of the predictors.

$$log\frac{\Pr(y_i=1|x_i)}{\Pr(y_i=0|x_i)} = \beta_0 + x_i^T\beta \quad (4\text{-}1)$$

where $\beta = (\beta_1, \dots, \beta_p)^T$ is the vector of coefficients of the p predictors to be estimated, excluding the intercept β_0, and $\Pr(y_i = 1|x_i)$ and $\Pr(y_i = 0|x_i)$ denote the conditional probabilities of the class labels 1 and 0, respectively. A maximum likelihood approach is commonly used in calculating the coefficients, and the log-likelihood can be written as:

$$l(\beta_0, \beta) = \sum_{i=1}^{N}\{y_i logProb(Y = 1; \beta) + (1 - y_i)log(1 - logProb(Y = 1; \beta))\}$$

$$= \sum_{i=1}^{N}\left\{y_i(\beta_0 + x_i^T\beta) - log\left(1 + e^{(\beta_0 + x_i^T\beta)}\right)\right\} \quad (4\text{-}2)$$

LASSO penalized logistic regression is a regression analysis method that performs both variable selection and regularization to enhance the statistical model's prediction accuracy and interpretability (Tibshirani, 1996). The LASSO penalized estimator shrinks some coefficients of a regression model and sets others to zero (0) to obtain variables with a substantial effect on the outcome (Tibshirani, 1996). LASSO performs important variable selection and variable correlation simultaneously. That is, between a pair of highly correlated variables, LASSO tends to pick the most important variable and discard the other by shrinking it toward zero.

Because the LASSO method performs variable selection through a continuous process, it does not suffer as much from high variability, i.e., it simultaneously does both

continuous shrinkage and automatic variable selection. The penalty term introduced by LASSO during the variable selection process ensures better estimation of the prediction error while avoiding overfitting. Selecting an optimal subset of explanatory variables is expected to improve the classification accuracy and make the model interpretation easier. Since some of the variables will be shrunk to zero, model thriftiness is also achieved.

The logistic regression model in Equation 4-1 can further be extended into the LASSO logistic regression model by adding the L_1 constraint on β parameters (Equation 4-3). The L_1 constraint is added to minimize the negative log-likelihood function with the penalty term. The generated coefficients can be expressed as a sparse linear combination of p number of predictor variables when solving the following optimization problem:

$$\min_{(\beta_0, \beta)} \left\{ \sum_{i=1}^{N} -\frac{1}{n} \left[y_i (\beta_0 + x_i^T \beta) - log \left(1 + e^{(\beta_0 + x_i^T \beta)} \right) \right] + P_\lambda(\beta) \right\} \quad (4\text{-}3)$$

where $P_\lambda(\beta)$ is the penalty term that depends on λ, a vector of non-negative regularization parameters, commonly referred to as a tuning parameter. The tuning parameter λ controls the strength of shrinkage in the explanatory variables, i.e., when λ takes larger values, more weight will be given to the penalty term and vice versa (Tibshirani, 1996). In this way, both shrinkage and variable selection are done simultaneously, and it is also this property that makes LASSO generally easier to interpret. Depending on the LASSO penalty's property, some coefficients in β will be exactly equal to zero. Further, it is also because of the penalty term λ that a LASSO model can include any number of variables.

While there are numerous penalty terms, a good penalty produces an estimator that is not biased or over-penalize large parameters (Algamal and Lee, 2015a). Thus, the adaptive LASSO penalty was selected in this research because it applies adaptive weights

when penalizing parameters (Zou, 2006). The adaptive LASSO imposes a higher weight to the small coefficients and a lower weight to the large coefficients to reduce the selection bias and fit the model consistently (Algamal and Lee, 2015b). Thus, this approach is said to have an oracle property. It is the main advantage of adaptive LASSO, compared to other penalty terms, such as the conventional LASSO, ridge penalty, and elastic net (Algamal and Lee, 2015a). Thus, the estimation of the vector βj is obtained by minimizing Equation 4-4, where w_j is a vector of data-driven weights. Although various methods have been used to estimate the weights (e.g., LASSO estimates), this research used ridge regression to estimate initial weights (SAS Institute Inc., 2019) because of the limitations of LASSO, as pointed out by Algamal and Lee (2015b).

$$\hat{\beta} = \arg\min \beta \left[-L(\beta|Y) + \lambda \sum_{j=1}^{p} w_j |\beta_j|\right] \quad (4\text{-}4)$$

4.2.1.2 Bootstrap Resampling

The bootstrap resampling method was used to estimate the logistic regression parameters to resolve the data imbalance caused by a disproportionally high percentage of normal incidents compared to primary incidents. Bootstrap resampling involves estimating parameters by repeatedly and randomly sampling subsets of data, and hence, providing more accurate estimates (Hastie et al., 2009; Kassambara, 2017; Pei et al., 2016). The conventional bootstrapping approach involves drawing a sample randomly and evenly with replacement. The resampling focused on neutralizing the effect of a significantly low percentage of primary incidents.

A three-step resampling approach was applied to the dataset. First, the incident dataset was divided into two groups: normal incidents and primary incidents. Then, *k* samples (where *k* equals the number of primary incidents) were randomly drawn from all

groups in each bootstrap replication. The resulting subset of data contained an equal number of normal incidents and primary incidents. The new dataset was then used to fit the penalized logistic regression. Finally, the procedure of drawing samples of *k* observations and fitting the model was repeated 5,000 times (arbitrarily selected as a trade-off between prediction accuracy and computation time), and the standard errors and confidence intervals of the estimates were calculated based on these 5,000 estimates.

The model coefficients were obtained by calculating the mean of all the estimates of the bootstrap samples. The odds ratio (OR), which represents how the dependent variable varies with the predictor variable, was computed relative to the base category. The odds ratio was calculated as:

$$OR = e^{coefficient} \tag{4-5}$$

4.2.2 Identify Factors that Influence the Likelihood of Cascading Crashes

This research used a Bayesian network to understand the probabilistic relationship among variables influencing the likelihood of cascading crashes. Before fitting the Bayesian network, a data-driven approach was first adopted to identify incidents that did not result in cascading events, referred to in this research as *non-cascading crashes*, and incidents that resulted in cascading events, referred herein as *cascading crashes*.

Figure 4-3 summarizes the approach used to investigate the probabilistic relationship between factors contributing to the occurrence of cascading crashes. Specifically, the methodology presented in Figure 4-3 is divided into three main steps: (a) fitting the penalized logistic regression model, (b) building the Bayesian network structure,

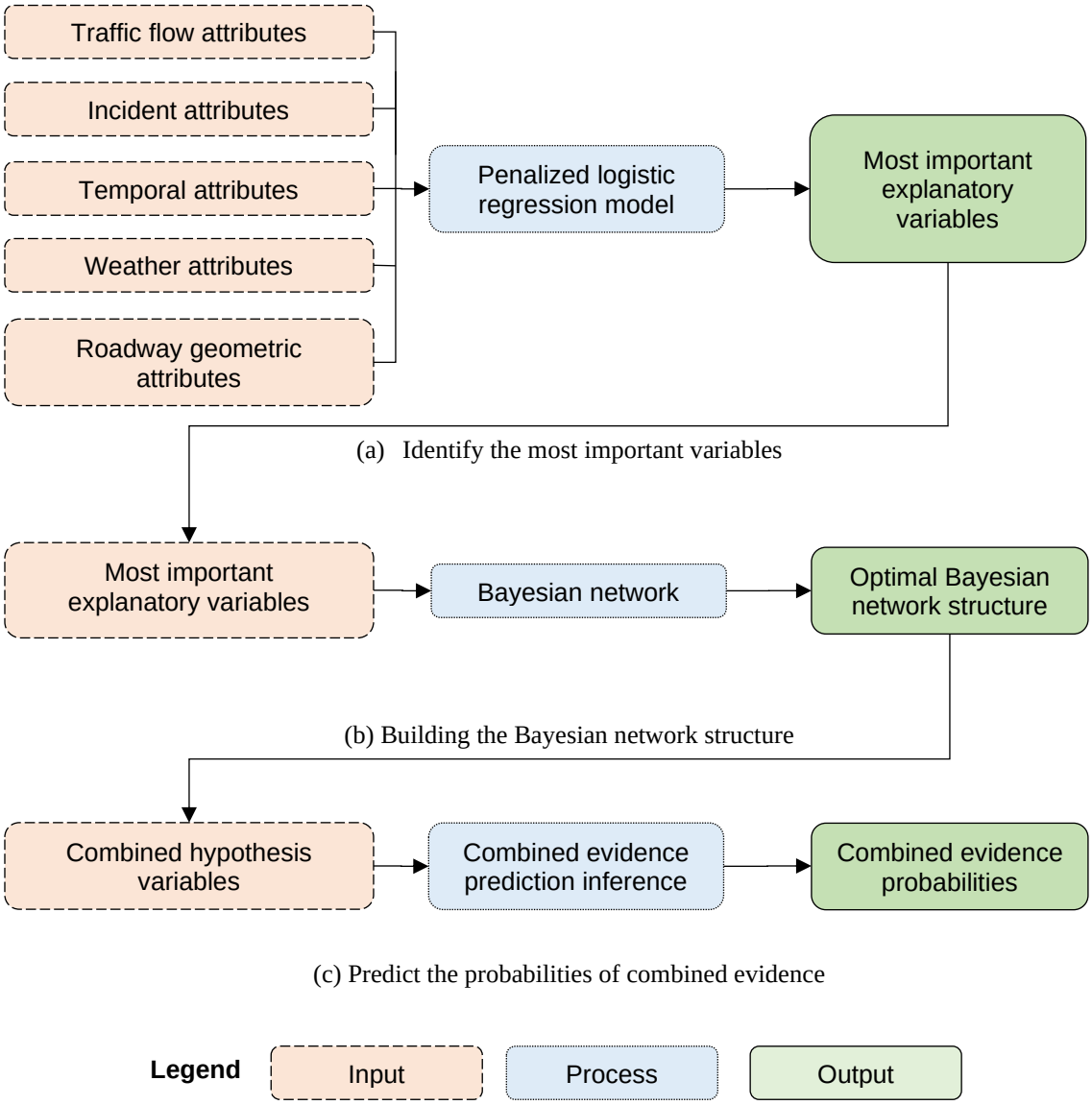
and (c) predicting the probabilities of combined evidence. The following subsections discuss each of the three steps in detail.

4.2.2.1 Penalized Logistic Regression

The penalized logistic regression was considered in this approach for the same reasons it was used to model the likelihood of secondary crashes. This approach has the advantage of simultaneously estimating the model coefficients, performing variable selection, and accounting for multi-collinearity (James et al., 2013). Since the proportion of cascading crashes was smaller than the proportion of non-cascading crashes, a bootstrap resampling method was used to fit the penalized logistic regression. For the cascading crash likelihood model, $\Pr(y_i = 1|x_i)$ and $\Pr(y_i = 0|x_i)$ denote the conditional probabilities of the *cascading crashes* and *non-cascading crashes*, respectively.

4.2.2.2 Bayesian Network

A Bayesian network was used to understand the probabilistic relationship among variables influencing the likelihood of cascading crashes. The Bayesian network model's choice was based on the interpretability of the Bayesian network, which explicitly presents the probabilistic relationships among variables in the model (Kidando et al., 2019b; Kutela and Teng 2019).



Note: Evidence = A condition that has been observed, e.g., incident type; Hypothesis variable = A variable that has a direct probabilistic relationship with the occurrence of cascading crashes.

Figure 4-3: Methodology Workflow for Cascading Crash Likelihood Model

To integrate subjectivity and reveal hidden probabilistic relationships among variables, the structure learning of the Bayesian network was conducted using an algorithm and expert knowledge. The Bayesian network structure was trained, using the Bayesian Dirichlet equivalent uniform (BDeu) as the search algorithm. After the Bayesian network structure was developed, the expert knowledge and findings from previous studies were

applied to refine the trained Bayesian network structure by only changing some of the arrow directions, such as the cause-effect direction. A similar approach was adopted in several previous studies (Cong et al., 2018; Stylianou and Dimitriou, 2018; Xie and Waller, 2010).

The greedy hill-climbing (GHC) algorithm was adopted as the search strategy to retrieve the optimal network structure from the data. The GHC algorithm iteratively adds, removes, and reverses edges to find a network with the highest score (Kidando et al., 2019b). The best network structure is obtained once the score cannot be improved further in the search process. Assume dataset *T* is used to train the network structure *B*, the Bayesian network structure then obtains the best network structure *B* by maximizing the scoring value, $\mathrm{BDeu}(B,T)$. The BDeu metric can be expressed as:

$$\mathrm{BDeu}(B,T) = \log\big(P(B)\big) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(log\left(\frac{\Gamma\left(\frac{N'}{q_i}\right)}{\Gamma\left(N_{ij}+\frac{N'}{q_i}\right)}\right) + \sum_{k=1}^{r_i} log\left(\frac{\Gamma\left(N_{ijk}+\frac{N'}{r_i q_i}\right)}{\Gamma\left(\frac{N'}{r_i q_i}\right)}\right)\right) \tag{4-6}$$

where,

N'	=	equivalent sample size,
N_{ij}	=	number of instances in the data *T*, where variable $\prod_{X_i}$ takes their *j-th* configuration, such that $\sum_{k=1}^{r_i} N_{ijk} = N_{ij}$,
N_{ijk}	=	number of instances in the data *T*,
r_i	=	number of states of the finite random variable, $\boldsymbol{X}_i$,
$q_i = \prod_{X_i \in \boldsymbol{X}_i} r_i$	=	number of possible configurations of the parent set $\prod_{X_i}$ of $\boldsymbol{X}_i$, and
n	=	number of observations.

Given the estimated optimal Bayesian network structure, and the evidence associated with the hypothesis variables, the model parameters, which are the discrete probability values

in the conditional probability tables, were estimated using the maximum likelihood estimation method.

4.2.2.3 Combined Evidence Prediction Inference

Using the optimal network retrieved in the analysis, the probabilistic inference was conducted through combined evidence prediction reasoning. The combined evidence predictive inference involves valuing the probability of an event's occurrence, e.g., cascading crash given some evidence. This process attempts to answer questions, such as what is the probability of a cascading crash occurring during peak hours when the road surface is wet?

The predicted probability of an incident resulting in cascading crashes, based on the combined evidence, was estimated as:

$$Predicted\ probability_i = P\left(Incident = i | e_{x_1} = x_1, e_{x_2} = x_2\ , \dots e_{x_h} = X_h\right) \quad (4\text{-}6)$$

where, e_x is the evidence of a hypothesis variable x, and x_h is the observed evidence of hypothesis variables X. Similar to individual hypothesis variable analyses, for the combined evidence, each observed evidence was assigned a certainty value of 1, i.e., $P\left(Incident = i | e_{x_1} = x_1, e_{x_2} = x_2 = 1\right)$. The conditional probability distributions of the trained Bayesian network structure were estimated using the maximum likelihood approach. Both the Bayesian network structure training and inferences were implemented using the pyAgrum 0.15.2 program, a Python open-source package (Wuillemin, 2019).

4.3 Predict the Probability of Secondary Crashes in Real-time

When an incident occurs, traffic conditions upstream of the incident vary with time, and so does the probability of secondary crashes. This research developed a real-time

dynamic prediction model to account for the temporal variation of secondary crash likelihood. Dynamic updating refers to the continuous updating of the secondary crash likelihood model over time. As a result, the model coefficients are continuously updated with time.

This research proposed a dynamic binary classifier which dynamically accounts for model uncertainty and allows within-model parameters to change over time. In contrast to the frequentist approach, the Bayesian approach takes the probability of a binary event as a random variable instead of a fixed value. This approach allows the incorporation of uncertainty in parameter estimates, which is particularly useful when forecasting.

A 5-minute time interval was used when the primary incident started impacting traffic to when the secondary crash occurred and when the normal incident started affecting traffic to the time the traffic returned to normal. To illustrate this, consider a normal incident and a primary incident that started impacting traffic from 8:00 AM to 9:00 AM. A secondary crash occurred at 8:30 AM within the queue caused by the primary incident. In this example, the first model will include information on both normal and primary incidents from 8:00 AM to 8:05 AM. Subsequent models will include information from 8:05 AM to 8:10 AM, 8:10 AM to 8:15 AM, etc. The last model for the normal incident will be from 8:55 AM to 9:00. Meanwhile, the last model for the primary incident will be from 8:25 AM to 8:30 AM, the time when the secondary crash occurred.

Before developing the prediction model, the following steps were first implemented: (a) define prior distributions, (b) extract prevailing explanatory variables, (c) impute the missing data points in the explanatory variables, (d) fit the Bayesian model, and

(e) generate posterior distributions. These steps are illustrated in Figure 4-4 and discussed in the following subsections.

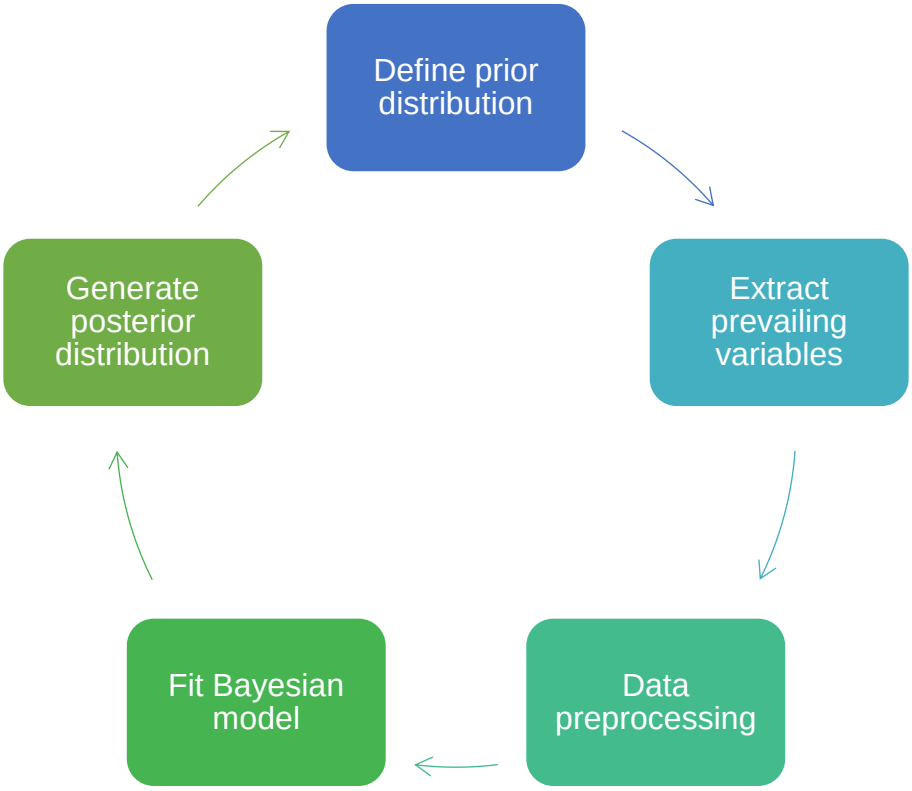


Figure 4-4: Methodology Workflow for Secondary Crash Risk Prediction Model

4.3.1 Define Prior Distribution

In Bayesian inference, the distributions of parameters are estimated using a maximum a posteriori probability method for which the prior distribution for all unknown parameters has to be defined (Kitali et al., 2017). Normally, two categories of priors are used in the Bayesian approach: informative and non-informative. Informative priors are based on the literature, expert knowledge, or information retrieved explicitly from previous data analysis (Kitali et al., 2018). On the other hand, non-informative priors, also called "vague" priors, are often used in the absence of reliable prior information regarding model

parameters. Assigning the non-informative priors to model parameters is common in Bayesian modeling, especially in the absence of informative priors (Kruschke, 2013). The non-informative priors impose minimal influence over the estimates and allow the data characteristics to dominate instead (Ntzoufras, 2009).

Non-informative priors were specified only in the first model since no previous information was available to generate the informative prior distributions. In particular, the normal distributions with a mean of zero and a standard deviation of 10 were assigned as the non-informative priors in the first model. For the subsequent models, the prior distributions were estimated using the posterior distributions of the immediately previous model. This process was implemented to improve the model output's robustness by accounting for the spatial and temporal variation of the secondary crash likelihood.

4.3.2 Extract Prevailing Explanatory Variables

Several factors may affect the likelihood of secondary crashes. Some of these factors are constant to the specific prior incident and do not vary with time. An excellent example of these factors includes pre-incident variables, which are variables that can be measured before the occurrence of the incident, e.g., traffic flow characteristics, incident type, incident occurrence time, incident severity, etc. Other variables that influence the likelihood of secondary crashes vary with time. These may include traffic flow characteristics upstream of the incident and rainfall. Thus, in this step, the prevailing traffic flow characteristics and rainfall data are prepared.

4.3.3 Data Preprocessing

Few incidents, i.e., less than 0.05%, were missing some prevailing traffic flow characteristics. In this case, the K-nearest neighbor method was used to replace the missing information with substituted values. K-nearest neighbor imputation is carried out by finding the k closest samples (Euclidian distance) in the data (Kuhn, 2019). The missing value of the predictor is computed by averaging the respective k-nearest samples. The value of k was chosen to be five (5). In this step, the data were also centered and scaled to ensure the robustness of the models.

4.3.4 Fit Bayesian Model

The response variable is binary, representing a secondary crash likelihood indicator, where 0 indicates that no secondary crash occurred (normal incident) and 1 signifies a secondary crash occurred (primary incident). Since the response variable is binary and asymmetric, the cloglog model was used to predict the probability of secondary crashes. Unlike other conventional classification regression models, such as logistic and probit models, the cloglog model is asymmetrical around the inflection point, a situation that favors the prediction of rare events (secondary crashes in this case) (Kitali et al., 2017).

In dynamic Bayesian cloglog regression, recursive estimation allows for sequential processing and is done in two steps: updating and predicting (McCormick et al., 2012; K. Yang et al., 2018). Consider a secondary crash occurrence as a binary response, y_t, and a set of predictors $X_t = \{x_{1,t}, x_{2,t}, x_{3,t}, \dots, x_{k,t}\}$. The predicted secondary crash occurrence at time t, denoted by y_t, is estimated with a vector of explanatory variables $\boldsymbol{X}_t$ using the cloglog regression, which is expressed as:

$$y_t = \begin{cases} 1 \; if \; the \; i^{th} \; incident \; is \; primary \; incident \\ 0 \qquad else \; normal \; incident \end{cases} \tag{4-8}$$

$$y_t \sim Binomial(p_t) \tag{4-9}$$

$$cloglog\ (p_t) = \log\big(-\log\,(1 - p_t)\big) = (\,\boldsymbol{X}_t)^T \theta_t \tag{4-10}$$

where, θ_t is a k-dimensional vector of regression coefficients, including intercept and explanatory variables, at time t, i.e., $\theta_t = \{x_{1,t}, x_{2,t}, x_{3,t}, \dots, x_{k,t}\}$. At a given time, t, the procedure takes the posterior mode of θ from time $t-1$ and uses it to construct the prior for time t. This is implemented by first using the information up to time $t-1$ to construct an estimate of the parameters for time t, yielding the *prediction equation*. This equation predicts the value of the observations at time t based on the estimated parameter using data up to time $t-1$. The prediction equation is then combined with the observed data at time t, and the new information factors into updated parameter estimates via the *updating equation* (K. Yang et al., 2018). As shown in Equation 4-11, the updating equation is proportional to the product of a Binomial density (Likelihood) and the prediction equation (Prior), so that the entire procedure has a Bayesian interpretation.

$$p(\theta_t|Y_t) \propto p(y_t|\theta_t)p(\theta_t|Y_{t-1}) \propto Likelihood \times Prior \tag{4-11}$$

To calibrate the model's parameters presented in Equation 4-10, a full Bayes (FB) approach, based on Markov Chain Monte Carlo (MCMC) simulations, was used. The No U-Turn Sampling (NUTS) technique was adopted in the analysis. NUTS is based on the Hamiltonian Monte Carlo (HMC) that avoids the random walk behavior, which has a greater advantage over convergence during sampling than other sampling techniques, such as Metropolis. More information regarding the comparison of NUTS and other techniques for sampling the posterior distribution can be found in Hoffman and Gelman (2014) study.

As with the Bayesian estimation, the convergence of the MCMC simulations was assessed using the Gelman-Rubin Diagnostic statistic. Also, a visual diagnostics approach was used to assess the convergence of the chains, including the use of the autocorrelation plot and the trace plot of each parameter. A total of 80,000 iterations, including 40,000 for warmup and 40,000 for inference, were sufficient to produce the desirable Gelman-Rubin statistic, which shows that the convergence has been reached.

4.3.5 Generate Posterior Distributions

As described in Section 4.3.4, the posterior distributions of the model parameters were obtained by combining the prior information with the likelihood function following the Bayes rule. These distributions were used to extract the model coefficients. In addition, the posterior distributions can be used to update the next model. This was achieved by using posterior distributions as priors. The posterior distribution for each explanatory variable was plotted in a histogram, also using the Kernel density. From these two plots, a parametric distribution, e.g., normal distribution, t-distribution, etc. that closely follows the posterior distribution was assumed.

4.4 Summary

This chapter described the approach used to identify secondary crashes, identify factors influencing the likelihood of secondary crashes, and dynamically predict the risk of secondary crashes in real-time. This research proposed a data-driven approach to better estimate the primary incident impact area, and hence, identify secondary crashes that occurred within the impacted area. To accomplish this, traffic incidents from the SunGuide® database and high-resolution speed data from HERE Technologies were used.

Following the identification of secondary crashes, the next step involved identifying factors influencing the likelihood of secondary crashes. A penalized logistic regression fitted using the bootstrap resampling approach was used to identify risk factors that influence the occurrence of secondary crashes. The proposed model is considered to improve the secondary crash risk model's predictive accuracy because it accounts for the asymmetric nature of secondary crashes, performs variable selection, and removes correlated variables. This research extends the previous research on secondary crash likelihood models by evaluating the impact of work zones on the occurrence of secondary crashes. The Bayesian network model used to explore the concurrent factors that influence the probability of cascading crashes was also discussed in detail. After identifying secondary crash influential factors, the final task involved dynamically predicting the risk of secondary crashes in real-time. The dynamic Bayesian cloglog model was proposed to accomplish this task and is described in detail in this chapter.

CHAPTER 5
RESULTS AND DISCUSSION

This chapter is divided into five major sections. The first section presents the results and discusses the secondary crash identification process. The second section discusses the results of the secondary crash likelihood models. The third section presents the leading causes of cascading crashes. The fourth section presents the results of the real-time secondary crash risk prediction model. The final section provides a summary of the research findings.

5.1 Secondary Crash Identification

To identify secondary crashes, 322,259 incidents from the SunGuide® database and high-resolution speed data from HERE Technologies were evaluated. Table 5-1 provides a summary of the secondary crashes identified along the study corridors. As indicated in Table 5-1, a total of 4,549 secondary crashes were identified from 3,977 primary incidents. This is an equivalent of 5.7 secondary crashes per mile per year along the 148-mile study corridor. In other words, about six secondary crashes per mile occurred annually along the study corridors.

Table 5-1: Secondary Crashes Identified Using the Improved Approach

Seg.	Seg. Len. (miles)	NI	PI	SC	All Inc.	All Crash	SC/ mile/year	Prop. of SC/Inc. (%)	Prop. of SCs/Crash (%)
HEFT	48	111,274	2,516	2,964	116,521	19,369	11.2	2.5	15.3
MSS	28	93,709	932	1,008	95,583	9,020	6.5	1.1	11.2
MCS	69	109,090	529	577	110,155	8,818	1.5	0.5	6.5
Overall	145	314,073	3,977	4,549	322,259	37,207	5.7	1.4	12.2

Note: HEFT = Homestead Extension of Florida's Turnpike; MSS = Mainline South Section; MCS = Mainline Central Section; Seg. Len. = Segment Length; NI = Normal Incident; PI = Primary Incident; SC = Secondary Crash; Inc. = Incident.

The identified secondary crashes account for 1.4% of all traffic incidents. While the proportion of secondary crashes, when compared to all incidents, may not seem

alarming at first glance, secondary crashes account for 12.2% of all crashes included in the analysis. As indicated in Table 5-1, the highest proportion of secondary crashes were identified along the HEFT corridor, followed by the MSS, and finally, the MCS.

5.1.1 Spatiotemporal Distribution of Secondary Crashes

Figures 5-1 and 5-2 show the spatial and temporal characteristics of secondary crashes in relation to primary incidents. The median distance between primary incidents and secondary crashes was found to be 2.5 miles. About half of secondary crashes occurred within 40 minutes after the primary incident. Almost half of the secondary crashes (47%) occurred within 2 miles upstream of the primary incident. Meanwhile, more than three-quarters of secondary crashes (93%) occurred within two hours.

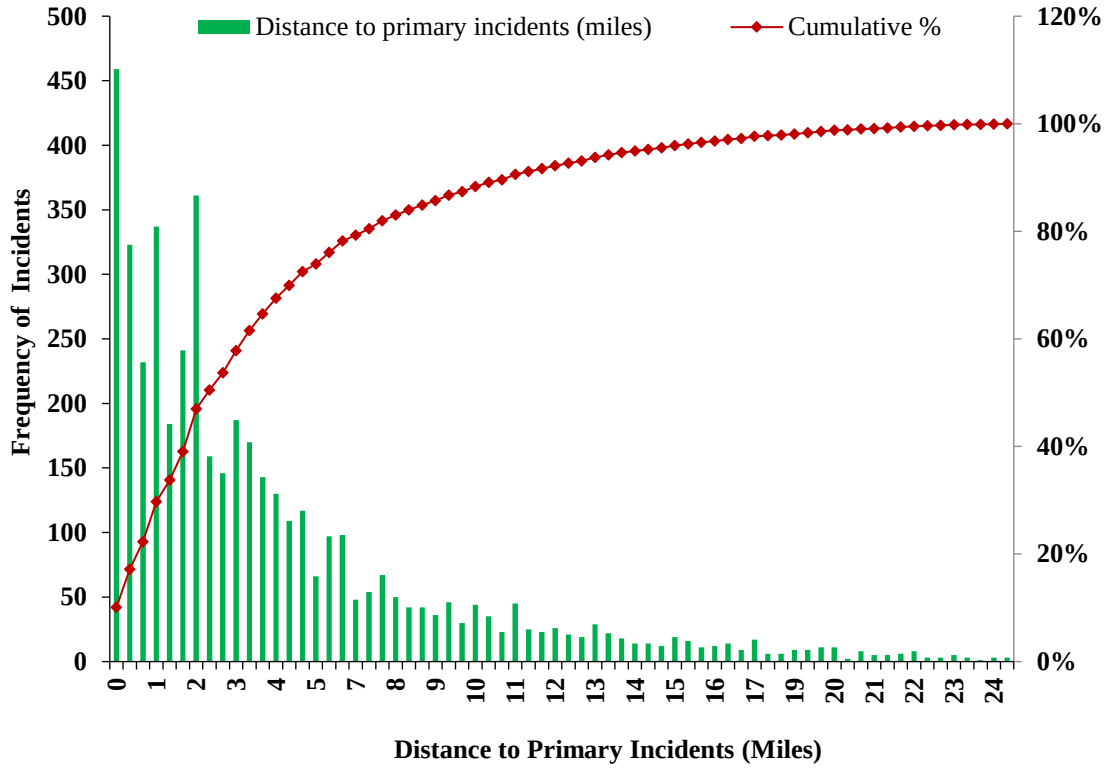


Figure 5-1: Spatial Distribution of Secondary Crashes in Relation to Primary Incidents

Overall, 40% of secondary crashes occurred within two hours of the onset of a primary incident and two miles upstream of the primary incident, the most commonly considered static spatiotemporal threshold.

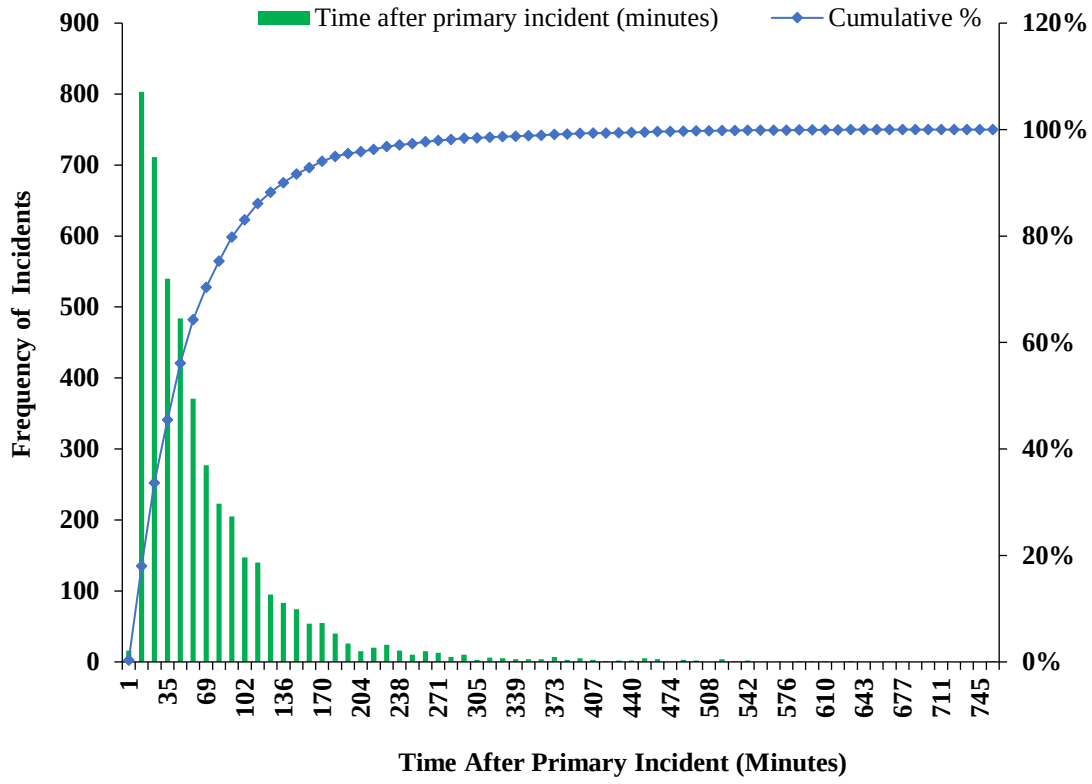


Figure 5-2: Temporal Distribution of Secondary Crashes in Relation to Primary Incidents

5.1.2 Time of Day and Day of Week Distribution

Figure 5-3 shows the distribution of the 4,549 secondary crashes, 3,977 primary incidents, and 314,073 normal incidents by different periods. More than three-quarters of secondary crashes (85%) occurred during peak hours, i.e., morning peak, 6:00 AM to 10:00 AM, and evening peak, 3:00 PM to 8:00 PM. Specifically, 33% of secondary crashes occurred during the morning peak, while the remaining 52% occurred during the evening peak. The highest proportion of secondary crashes during morning peak hours occurred

from 8:00 AM to 9:00 AM (11%), while the highest proportion of secondary crashes during the evening peak period (13%) occurred between 5:00 PM and 6:00 PM, summing to a total of 24% of all secondary crashes that occurred along the study corridors. In total, the proportion of secondary crashes that occurred during peak hours accounted for 85% of total secondary crashes that occurred on the study corridors.

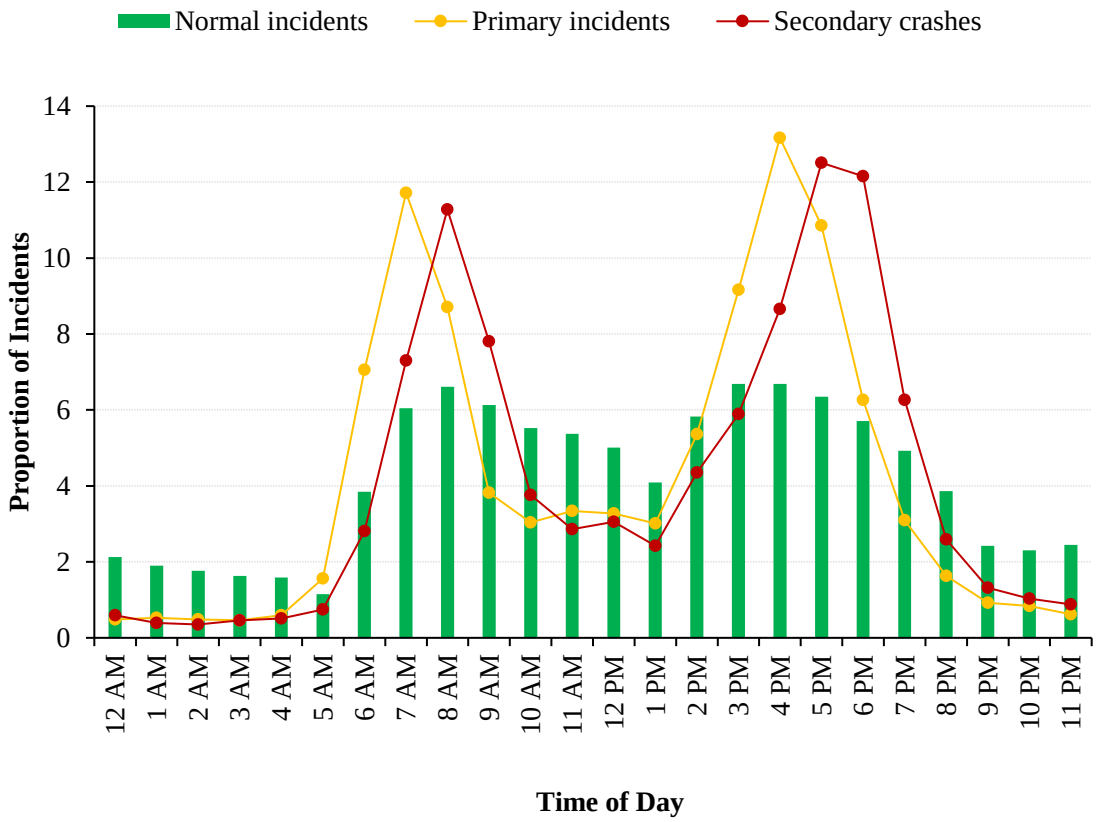


Figure 5-3: Distribution of Traffic Incidents by Time of Day

The highest proportion of primary incidents was observed during the evening peak period between the hours of 2:00 PM and 8:00 PM, accounting for 50% of all primary incidents. As can be inferred from Figure 5-3, the peaks of primary incidents and secondary crashes are one hour apart. Unlike primary incidents and secondary crashes, there is no significant distinction in the distribution of normal incidents during peak hours. More than

three-quarters of normal incidents (94%) occurred between the hours of 6:00 AM and 8:00 PM. As can be observed from Table 5-2, approximately half of normal incidents occurred during peak hours (53%), while the remaining half occurred during off-peak hours.

Table 5-2: Distribution of Traffic Incidents by Time of Day

Temporal Characteristic	Category	Incident Category (%)		
		Normal Incidents	Primary Incidents	Secondary Crashes
Time of Day	Peak hours	68	84	85
	Off-peak hours	32	16	15

More than three-quarters of both primary incidents (84%) and secondary crashes (85%) occurred during peak hours. Compared to off-peak hours, peak-hour traffic flow characteristics were found to contribute more to the occurrence of secondary crashes. Smaller gaps between vehicles characterize congested traffic, providing less maneuvering room for drivers to avoid a crash (Mishra et al., 2016; Kitali et al., 2019b).

Figure 5-4 presents the distribution of incidents by day of the week. It can be inferred from Figure 5-4 that the proportion of normal incidents, primary incidents, and secondary crashes is much higher on weekdays than on weekends. Compared to other days of the week, Friday was found to experience the highest proportion of secondary crashes (20%). Only 13% of secondary crashes were found to occur on weekends.

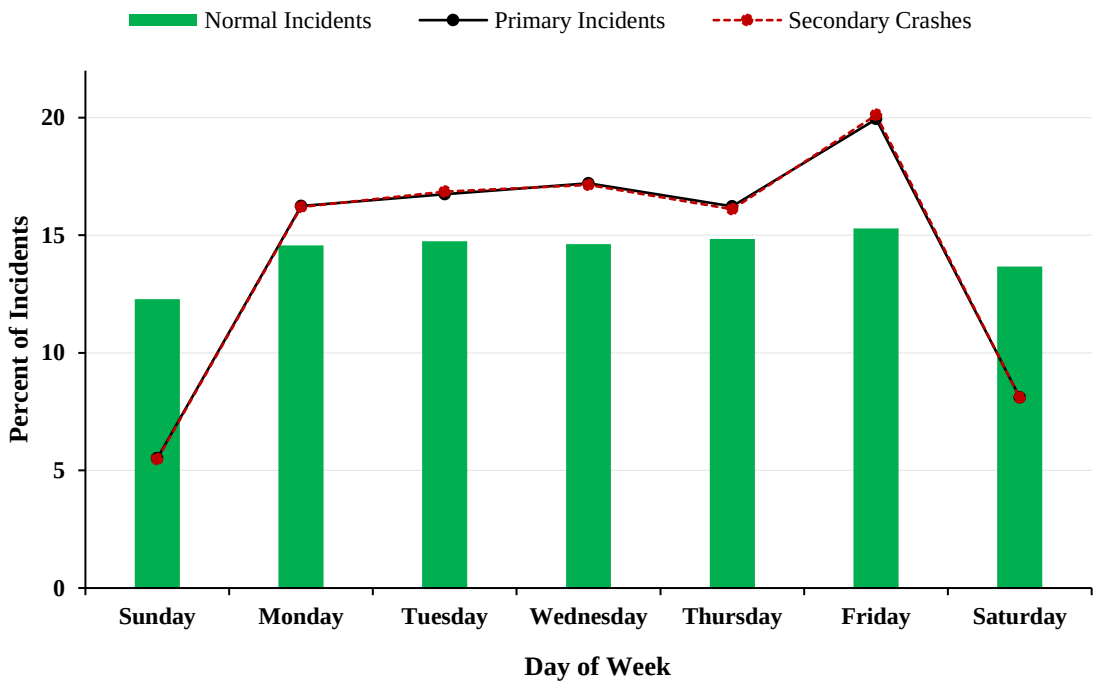


Figure 5-4: Distribution of Normal Incidents and Secondary Crashes by Day of Week

5.1.3 Incident Characteristics

Figure 5-5 provides the distribution of the incident clearance duration for towing-involved and no-towing-involved incidents. From Figure 5-5, it can be inferred that 94% of traffic incidents that did not involve towing were cleared within 95 minutes, while only 64% of traffic incidents that involved towing were cleared within 90 minutes, a value adopted from the FDOT's Open Road Policy.

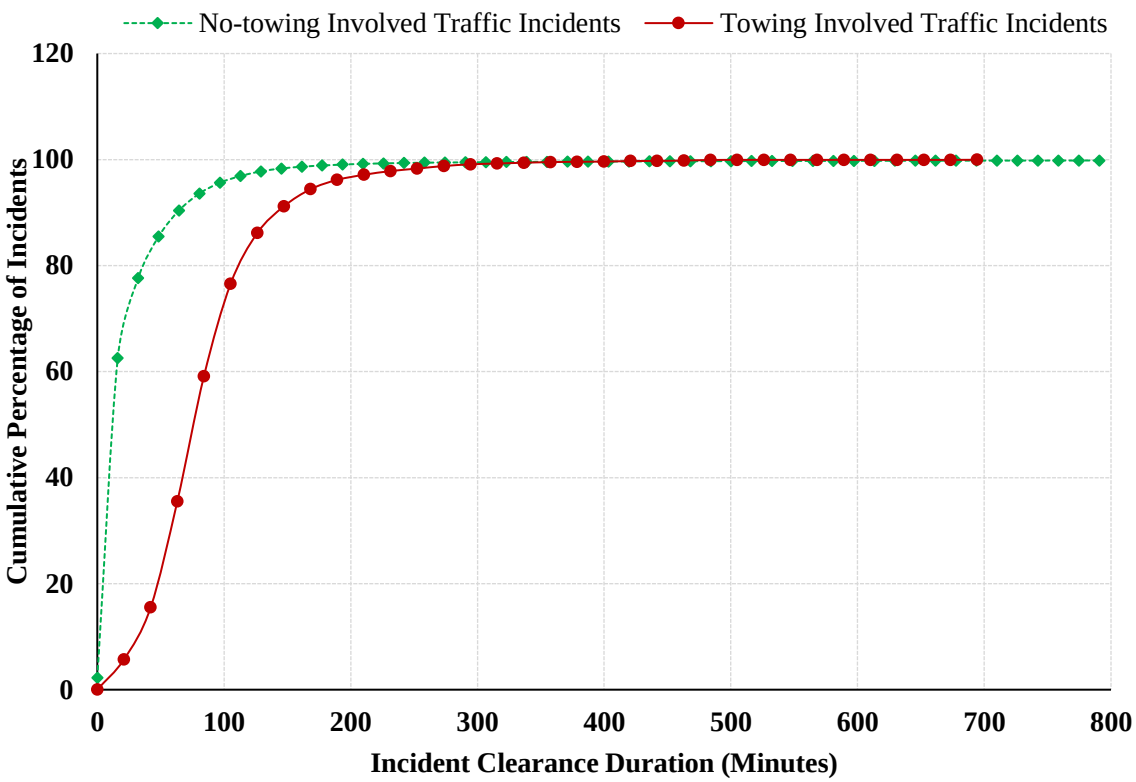


Figure 5-5: Distribution of Incident Clearance Duration for Towing-Involved and No-Towing Involved Incidents

In addition to towing, the Emergency Medical Services (EMS) presence at the incident scene was also identified as one of the factors that increase the incident clearance duration. This observation is evident in Figure 5-6, where 95% of traffic incidents that did not involve EMS were cleared within 90 minutes, while only 64% of traffic incidents that involved EMS were cleared within 90 minutes.

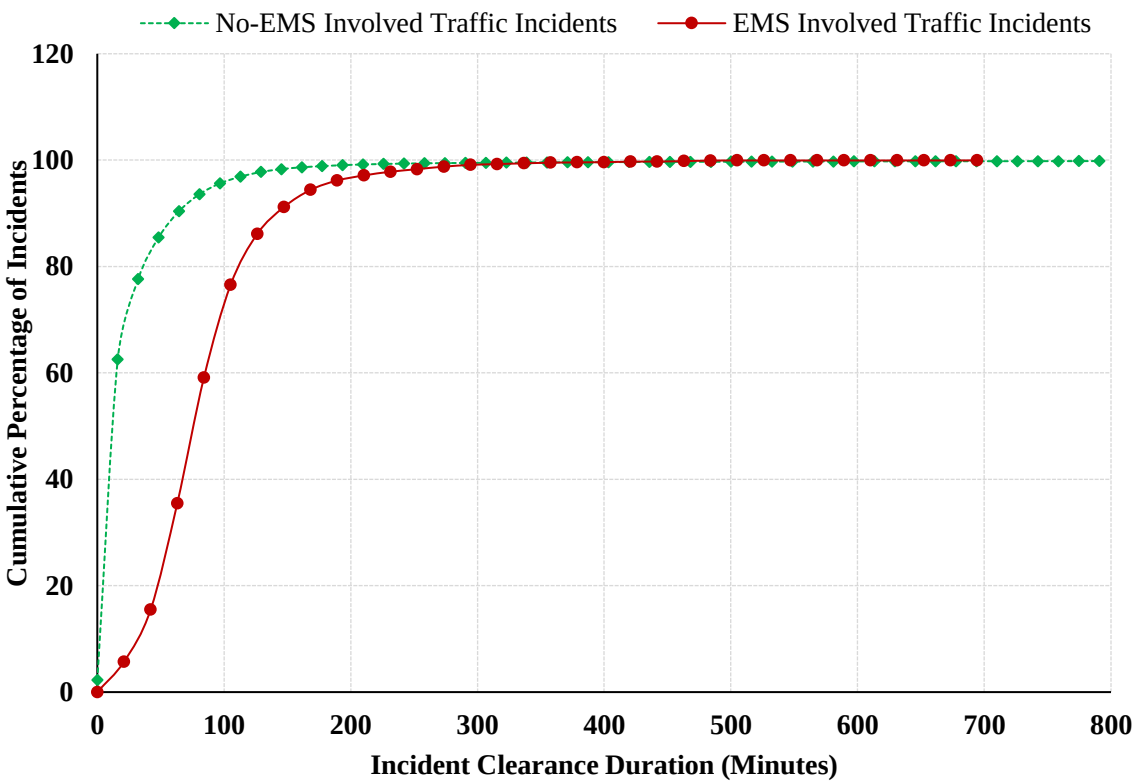


Figure 5-6: Distribution of Incident Clearance Duration for EMS-Involved and No-EMS Involved Incidents

As expected, traffic incidents involving towing and EMS resulted in longer incident clearance durations as they tend to require more time to be cleared. As indicated in previous studies, the likelihood of secondary crashes increases with an increase in incident clearance duration (Xu et al., 2016; Kitali et al., 2018). This is evident from the data, as 13% of primary incidents required towing, while only 3% of normal incidents required towing (see Table 5-3). Similarly, a higher percentage of incidents involving EMS resulted in secondary crashes (11%). Furthermore, while only 28% of normal incidents involved more than one responding agency, 51% of primary incidents and 55% of secondary crashes involved multiple responding agencies. These statistics suggest that incidents involving a greater number of responding agencies increase the likelihood of secondary crashes.

Table 5-3: Incident Distribution Based on Responders' Characteristics

Incident Characteristics	Category	Incident Category (%)		
		Normal Incidents	Primary Incidents	Secondary Crashes
Towing Involved	No	97.0	86.6	85.2
	Yes	3.0	13.4	14.8
Emergency Involved	No	98.2	89.4	89.1
	Yes	1.8	10.6	10.9
Number of Responding Agencies	1	71.9	49.0	45.2
	2	24.3	31.7	33.8
	3	1.8	7.0	9.0
	4	0.9	4.7	5.3
	5	0.8	4.9	5.0
	6+	0.3	2.6	1.7

As can be observed from Table 5-4, 97% of normal incidents did not result in a lane closure, while 21% of primary incidents resulted in a lane closure. The percentage of lanes closed is an indicator of the severity of the primary incident, as severe incidents tend to result in an increased number of lanes closed (Kitali et al., 2018). About 9% of primary incidents resulted in moderate to severe impacts on traffic, while only 1% of normal incidents were moderate to severe.

Table 5-4: Incident Characteristics

Incident Characteristics	Category	Incident Category (%)		
		Normal Incidents	Primary Incidents	Secondary Crashes
Percentage of Lanes Closed	0	97.0	79.3	99.7
	0-50	0.4	2.7	0.2
	50-100	2.6	18.0	0.1
Incident Severity*	Minor	98.9	90.6	93.4
	Moderate	0.7	5.9	4.9
	Severe	0.4	3.5	1.7

Note: *Incident severity refers to the extent to which the incident impacted the traffic.

As indicated in Figure 5-7, only 10% of normal incidents were crashes, a proportion similar to all incidents (12%), while approximately half of the primary incidents were crashes (47%). In other words, the probability of secondary crashes was found to be higher

when primary incidents were crashes. Note that the category "Other" in Figure 5-7 includes emergency vehicles, vehicle fire, and police activity. All incidents include normal incidents, primary incidents, and secondary crashes.

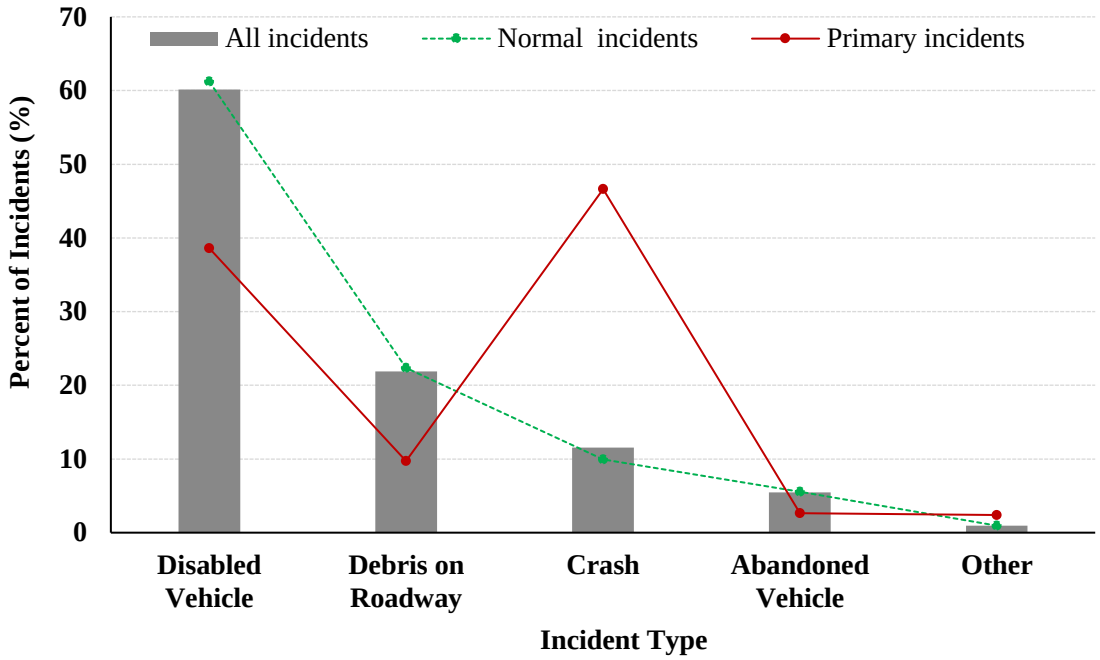


Figure 5-7: Distribution of Incidents by Incident Type

Figure 5-8 shows the distribution of the incident clearance duration for normal incidents and primary incidents. Overall, normal incidents were cleared more quickly than primary incidents; approximately 94% of the normal incidents were cleared within 90 minutes, while only 82% of the primary incidents were cleared within 90 minutes. The longer clearance time of the primary incidents could be considered one of the factors that may have contributed to the occurrence of secondary crashes.

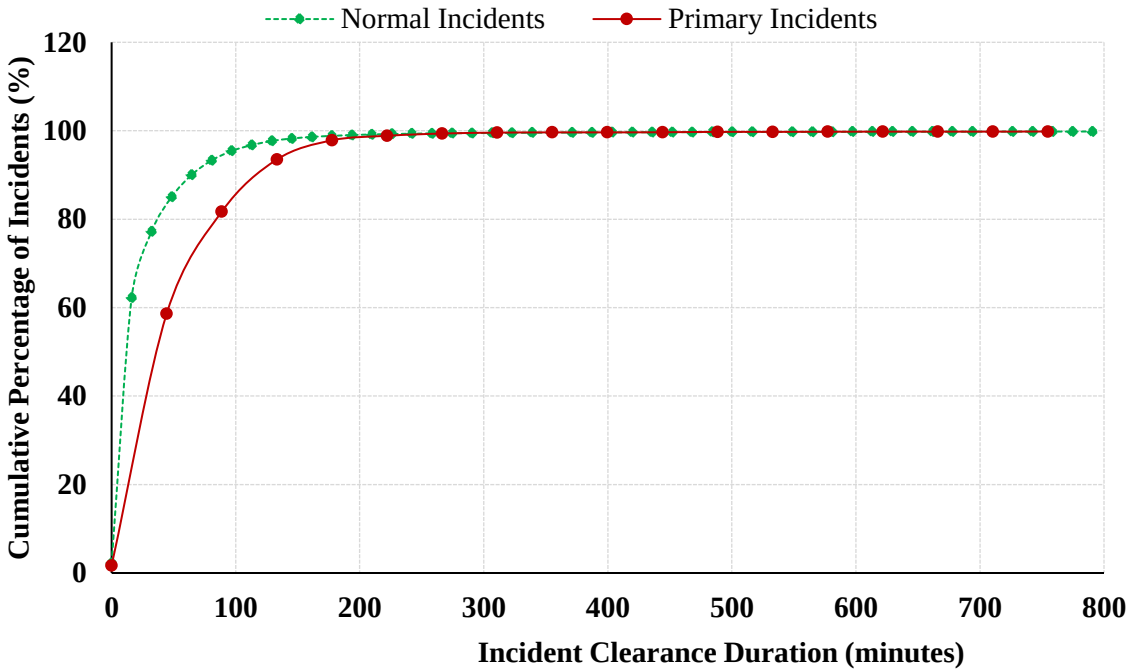


Figure 5-8: Distribution of Incident Clearance Duration for Normal and Primary Incidents

Figure 5-9 presents the distribution of the incident clearance duration for the identified primary incidents and secondary crashes. Approximately 77% of the secondary crashes were cleared within 90 minutes, while 82% of the primary incidents were cleared within 90 minutes. This observation implies that primary incidents were cleared somewhat faster than secondary crashes.

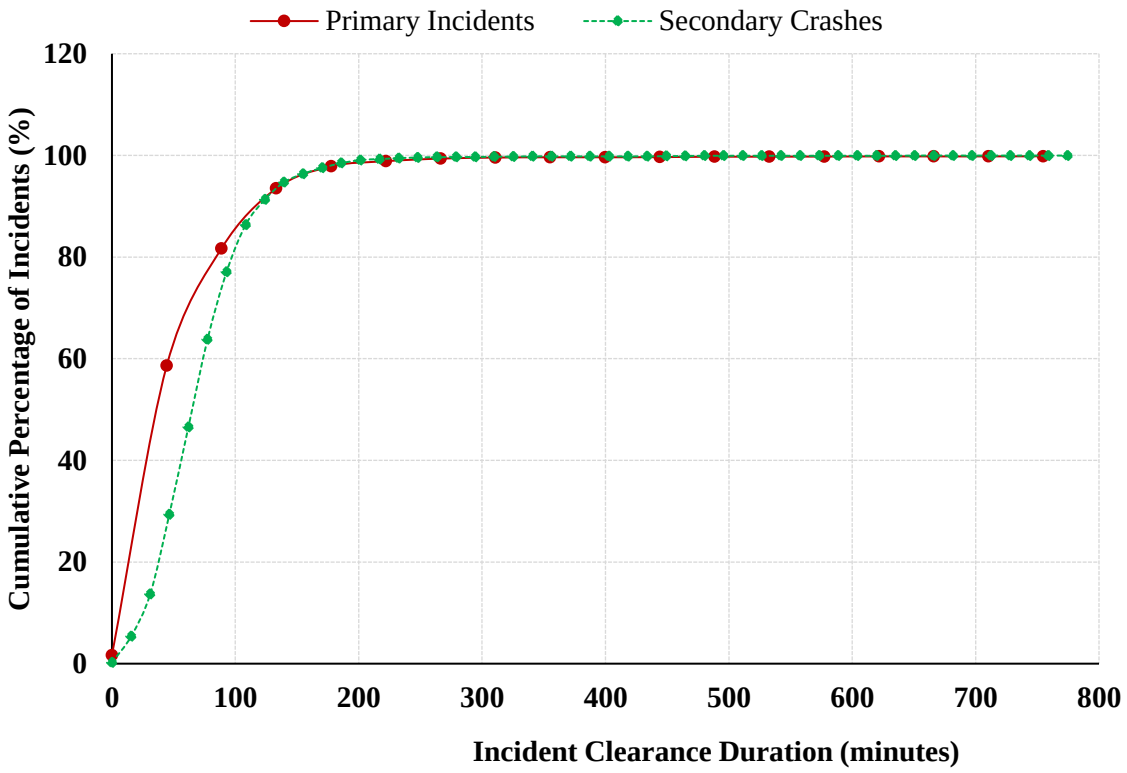


Figure 5-9: Distribution of Incident Clearance Duration for Primary Incidents and Secondary Crashes

5.1.4 Environmental Conditions

Environmental conditions (i.e., weather, roadway surface, and lighting) have been identified as some of the factors that influence the likelihood of secondary crashes (Vlahogianni et al., 2012). Table 5-5 summarizes the variation of weather condition, roadway surface condition, and lighting condition by incident category, i.e., normal incident, primary incident, and secondary crash. Regarding weather condition, as indicated in Table 5-5, more than three-quarters of all the three incident categories occurred under clear weather condition. Compared to normal incidents (2%), a higher proportion of primary incidents (13%) occurred during cloudy/fog/rainy conditions. Similarly, a higher percentage of primary incidents (11%) and secondary crashes (18%) occurred on wet

surface conditions. These statistics imply that inclement weather conditions and adverse road surface conditions are among the factors that increase the probability of secondary crashes.

Table 5-5: Environmental Conditions

Environmental Condition	Category	Incident Category (%)		
		Normal Incidents	Primary Incidents	Secondary Crashes
Weather	Clear	97.9	87.3	79.9
	Cloudy/Fog/Rain	2.1	12.7	20.1
Roadway Surface Condition	Dry	98.7	88.7	81.6
	Wet	1.3	11.3	18.4
Lighting Condition	Daylight	71.3	80.2	77.5
	Dark/Dusk/Down	28.7	19.8	22.5

5.2 Secondary Crash Influential Factors

5.2.1 Descriptive Statistics

A total of 116,521 incidents on the HEFT corridor and 95,583 incidents on the MSS were used to identify secondary crashes. Altogether, 2,964 secondary crashes were identified on the HEFT corridor, accounting for 3% of the 116,521 HEFT incidents that were included in the analysis. A total of 1,008 secondary crashes were identified from the 95,583 MSS evaluated. These secondary crashes account for 1% of all MSS incidents. Descriptive statistics indicated that more secondary crashes occurred on the HEFT than on the MSS. Although the proportion of secondary crashes to all incidents on both corridors may not seem initially alarming, proportionally, roughly 11 secondary crashes/mile/year and seven secondary crashes/mile/year occurred on the HEFT and the MSS, respectively. Secondary crashes also accounted for 15.3% and 11.2% of all HEFT and MSS crashes, respectively.

Following a careful review, the only factor that was different along the two corridors was the presence of work zones. The HEFT had lane widening activities throughout the study period, while the MSS had very little construction activity during the study period. Furthermore, similar incident response procedures are in place for the two corridors. The same incident responders attend to incidents and report to the same agency (Florida's Turnpike), and the same TMC is responsible for managing incidents for both sections. Thus, the higher proportion of secondary crashes on the HEFT may be attributed to the presence of major construction activities.

The likelihood model response variable is dichotomous, consisting of normal incidents and primary incidents. Note that normal incidents are those that did not lead to any secondary crashes, and primary incidents are those that led to secondary crashes. After removing secondary crashes and incidents that were missing information for some of the attributes, the final number of incidents included in the likelihood model consisted of 105,479 and 88,340 incidents for the HEFT and the MSS, respectively.

Tables 5-6 and 5-7 summarize the data and the variables used in the analysis. The variables were categorized into the incident, temporal, weather, traffic flow, and roadway geometric characteristics. For the normal incidents, the mean speed before they occurred and the mean prevailing speed were comparatively similar; while for the primary incidents, the mean prevailing speed was relatively lower than the mean speed before the incidents occurred. As expected, the variation in speed for primary incidents was higher than that of normal incidents.

Most normal incidents were vehicle-related (see Table 5-6). The proportion of primary incidents that were crashes was higher than that of normal incidents that were

crashes. The proportion of primary-crash incidents on the HEFT was higher than that of the MSS. Many normal incidents were responded to by one agency, while the proportion of primary incidents attended to by one or more than one agency were almost equal for both study corridors. Compared to normal incidents, a higher proportion of primary incidents had emergency medical services (EMS) as one of the responding agencies. A similar observation can be made for incidents where towing was involved and on moderate to severe incidents. While an equal proportion of normal incidents occurred during peak and off-peak hours, more than two-thirds of primary incidents occurred during peak hours. Compared to normal incidents, a higher proportion of primary incidents occurred during adverse weather conditions and on wet road surfaces.

Table 5-6: Descriptive Statistics of Continuous Variables

Variable	Incident Category	HEFT					MSS				
		Min	Mean	Med	SD	Max	Min	Mean	Med	SD	Max
Shoulder width (feet)	All incident	6.0	11.3	10.0	2.9	25.0	0.9	10.7	10.4	2.7	32.0
	Normal Incident	6.0	11.2	10.0	2.9	23.0	0.9	10.7	10.4	2.7	32.0
	Primary Incident	8.0	11.3	11.0	2.3	25.0	4.0	10.3	10.5	1.6	19.0
Mean speed before the incident (mph)	All incident	1.9	59.5	63.4	13.2	80.3	1.0	64.1	66.3	9.5	80.7
	Normal Incident	1.9	60.1	63.6	12.6	80.3	2.3	64.3	66.3	9.3	80.7
	Primary Incident	2.0	40.1	43.8	19.0	75.8	1.0	50.0	57.9	18.1	77.0
SD of speed before the incident (mph)	All incident	0.0	2.5	1.9	2.5	32.7	0.0	2.1	1.6	2.0	34.0
	Normal Incident	0.0	2.5	1.9	2.5	32.7	0.0	2.1	1.6	2.0	31.2
	Primary Incident	0.0	3.7	2.4	3.8	29.1	0.0	3.6	1.9	4.5	34.0
Mean prevailing speed (mph)	All incident	2.0	59.3	63.0	12.4	85.5	4.4	63.9	66.2	9.1	80.9
	Normal Incident	8.0	60.0	63.2	11.5	85.5	7.5	64.1	66.2	8.7	80.9
	Primary Incident	2.0	33.5	29.5	15.3	69.9	4.4	43.3	44.6	17.0	67.6
SD of prevailing speed (mph)	All incident	0.0	4.0	2.5	4.3	37.8	0.0	3.1	1.9	3.7	30.9
	Normal Incident	0.0	3.8	2.4	4.1	37.8	0.0	3.0	1.9	3.6	30.9
	Primary Incident	0.0	10.4	10.6	5.2	28.4	0.0	9.0	8.5	6.2	26.1

Note: Min = Minimum; Med = Median; SD = Standard deviation; Max = Maximum

Table 5-7: Descriptive Statistics of Categorical Variables

Attribute	Attribute Category	HEFT					MSS				
		Secondary Crash Likelihood				Total	Secondary Crash Likelihood				Total
		No		Yes			No		Yes		
		Count	%	Count	%		Count	%	Count	%	
Incident Attributes											
Incident type	Vehicle problem	67,917	66	1,182	42	69,099	62,309	71	457	49	62,766
	Hazard	20,133	20	230	8	20,363	17,704	20	125	13	17,829
	Crash	14,629	14	1,388	50	16,017	7,392	8	349	37	7,741
Number of responding agencies	1	70,476	69	1,392	50	71,868	63,158	72	482	52	63,640
	2+	32,203	31	1,408	50	33,611	24,247	28	449	48	24,696
EMS involvement	No	100,591	98	2,562	92	103,153	85,944	98	813	87	86,757
	Yes	2,088	2	238	9	2,326	1,461	2	118	13	1,579
Towing involvement	No	99,608	97	2,496	89	102,104	85,257	98	805	86	86,062
	Yes	3,071	3	304	11	3,375	2,148	2	126	14	2,274
Lane closure	No	98,543	96	2,245	80	100,788	85,048	97	745	80	85,793
	Yes	4,136	4	555	20	4,691	2,357	3	186	20	2,543
Incident severity*	Minor	101,289	99	2,572	92	103,861	86,574	99	831	89	87,405
	Moderate/severe	1,390	1	228	8	1,618	831	1	100	11	931
Temporal Attributes											
Day of week	Weekday	75,815	74	2,499	89	78,314	65,901	75	811	87	66,712
	Weekend	26,864	26	301	11	27,165	21,504	25	120	13	21,624
Time of day	Off-peak	59,435	58	639	23	60,074	46,505	53	263	28	46,768
	Morning peak	17,735	17	894	32	18,629	17,507	20	302	32	17,809
	Evening peak	25,509	25	1,267	45	26,776	23,393	27	366	39	23,759

Note: *Incident severity refers to the extent to which the incident impacted the traffic; EMS is Emergency Medical Service; N/A = Not Applicable.

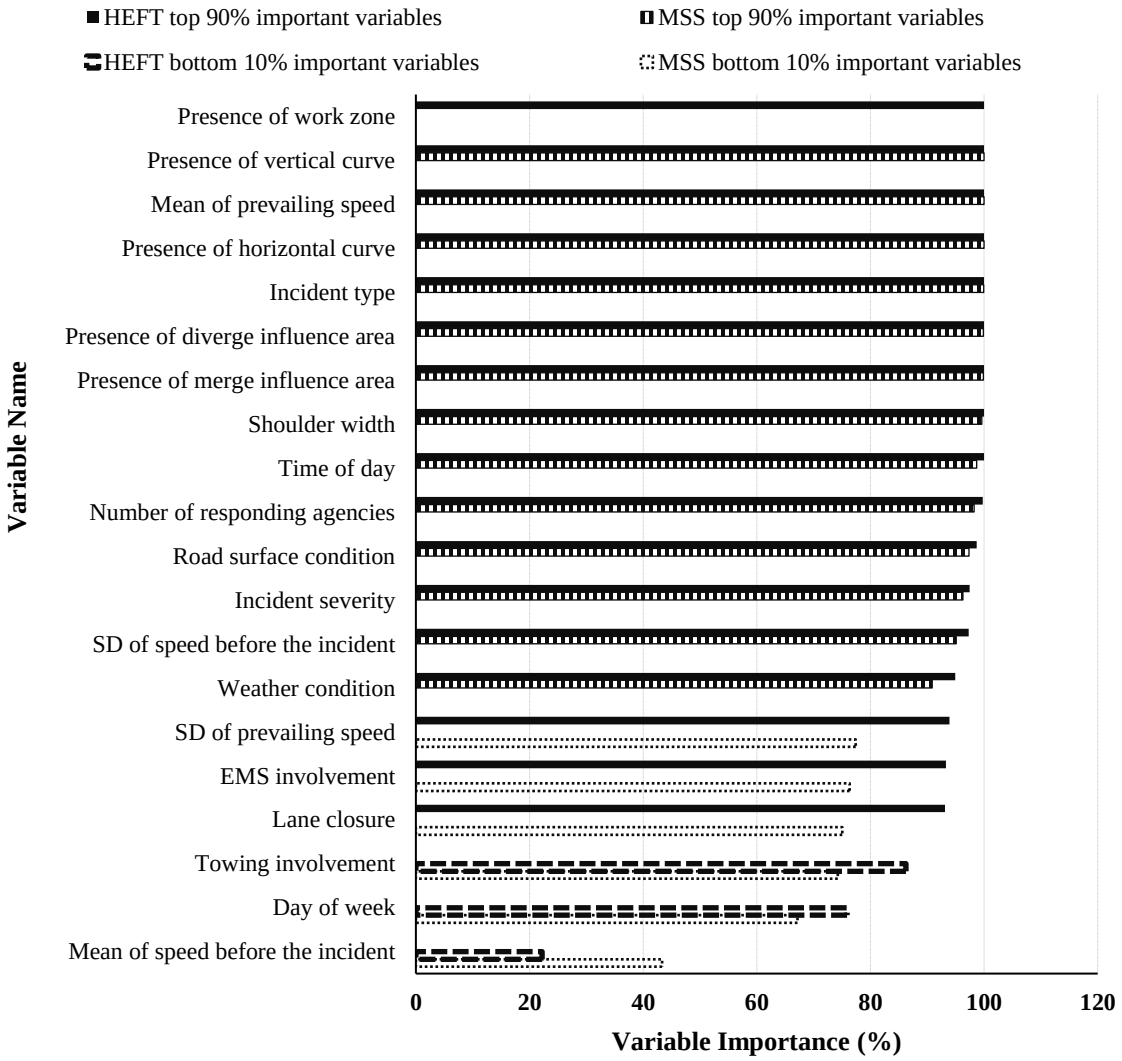
Table 5-7: Descriptive Statistics of Categorical Variables (continued)

Attribute	Attribute Category	HEFT					MSS				
		Secondary Crash Likelihood				Total	Secondary Crash Likelihood				Total
		No		Yes			No		Yes		
		Count	%	Count	%		Count	%	Count	%	
Weather Attributes											
Weather condition	Clear	100,519	98	2,439	87	102,958	86,303	99	843	91	87,146
	Adverse	2,160	2	361	13	2,521	1,102	1	88	9	1,190
Road surface condition	Dry	101,074	98	2,472	88	103,546	86,575	99	851	91	87,426
	Wet	1,605	2	328	12	1,933	830	1	80	9	910
Roadway Geometric Attributes											
Presence of horizontal curve within IIA	No	48,980	48	655	23	49,635	35,628	41	239	26	35,867
	Yes	53,699	52	2,145	77	55,844	51,777	59	692	74	52,469
Presence of vertical curve within IIA	No	67,913	66	1,675	60	69,588	43,301	50	276	30	43,577
	Yes	34,766	34	1,125	40	35,891	44,104	50	655	70	44,759
Presence of diverge influence area within IIA	No	49,102	48	616	22	49,718	46,958	54	294	32	47,252
	Yes	53,577	52	2,184	78	55,761	40,447	46	637	68	41,084
Presence of merge influence area within IIA	No	45,019	44	540	19	45,559	26,351	30	264	28	26,615
	Yes	57,660	56	2,260	81	59,920	61,054	70	667	72	61,721
Proportion of major work zone within IIA	No	63,801	62	1,085	39	64,886					
	Yes	38,878	38	1,715	61	40,593	N/A	N/A	N/A	N/A	N/A
Response variable	Secondary crash likelihood	102,679	97	2,800	3	105,479	87,405	99	931	1	88,336

Note: IIA is Incident Impact Area; N/A = Not Applicable.

5.2.2 Secondary Crash Likelihood

The penalized logistic regression was used to investigate the impact of work zones on the likelihood of secondary crashes. Variable importance, based on the percentage of times selected, is illustrated in Figure 5-10. The top 90% of selected variables when fitting the penalized logistic regression on the bootstrapped samples were considered the most important variables.



Note: EMS = Emergency Medical Service; SD = Standard Deviation.

Figure 5-10: Selection of the Important Variables for the Secondary Crash Likelihood Model

Most variables (17 of 20 variables) included in the likelihood model for the HEFT corridor were found to be important. These variables include the presence of work zone within the incident impact area, presence of diverge influence area within the incident impact area, incident type, mean prevailing speed, shoulder width, presence of horizontal curve within the incident impact area, presence of vertical curve within the incident impact area, road surface condition, lane closure, number of responding agencies, standard deviation of speed before the incident, standard deviation of prevailing speed, weather condition, time of day, incident severity, presence of merge influence area, and EMS involvement.

Of the 19 variables included in the likelihood model for the MSS, the following 13 variable were selected as the most important: presence of diverge influence area within the incident impact area, presence of merge influence area within the incident impact area, incident type, mean prevailing speed, shoulder width, presence of horizontal curve within the incident impact area, presence of vertical curve within the incident impact area, road surface conditions, lane closure, number of responding agencies, standard deviation of speed before the incident, standard deviation of prevailing speed, and weather condition.

Table 5-8 shows the penalized logistic regression results for the two study corridors and the number of times each variable was selected as an important variable in the model. The model coefficients were obtained by calculating the mean of all estimates from the bootstrap samples. The following subsections discuss the results from Table 5-8 in detail. Note that only the most important variables, significant at the 95% Bootstrap confidence interval (CI), are discussed.

Table 5-8: Results of the Penalized Logistic Regression Fitted Using Bootstrap Samples

Variable	Category	HEFT								MSS							
		Mean	OR	Med	SD	CI (%)		Count	%[b]	Mean	OR	Med	SD	CI (%)		Count	%[b]
						2.5	97.5							2.5	97.5		
Intercept	N/A	1.61	N/A	1.61	0.23	1.17	2.08	5,000	100	6.06	N/A	6.05	0.72	4.66	7.48	5,000	100
Traffic Flow Attributes																	
Mean speed before the incident (mph)	N/A	**0.01**	**1.01**	**0.01**	**0.00**	**0.01**	**0.02**	**4,874**	**97**	0.02	1.02	0.02	0.01	0.01	0.03	2,167	43
SD of speed before the incident (mph)	N/A	-0.01	0.99	-0.01	0.01	-0.04	0.01	1,112	22	**-0.07**	**0.93**	**-0.07**	**0.02**	**-0.12**	**-0.04**	**4,756**	**95**
Mean prevailing speed (mph)	N/A	**-0.11**	**0.90**	**-0.11**	**0.00**	**-0.11**	**-0.10**	**5,000**	**100**	**-0.12**	**0.89**	**-0.12**	**0.01**	**-0.15**	**-0.10**	**5,000**	**100**
SD of prevailing speed (mph)	N/A	**0.04**	**1.04**	**0.04**	**0.01**	**0.02**	**0.05**	**5,000**	**100**	0.02	1.02	0.02	0.02	-0.02	0.05	3,871	77
Incident Attributes																	
Incident type	Vehicle problem																
	Hazard	0.00	1.00	0.00	0.08	-0.13	0.17	3,124	62	**0.26**	**1.29**	**0.25**	**0.10**	**0.06**	**0.45**	**4,836**	**97**
	Crash	**0.55**	**1.74**	**0.55**	**0.08**	**0.39**	**0.71**	**5,000**	**100**	**0.53**	**1.70**	**0.53**	**0.19**	**0.17**	**0.91**	**4,996**	**100**
Lane closure	No																
	Yes	**0.13**	**1.14**	**0.12**	**0.16**	**-0.17**	**0.46**	**4,746**	**95**	-0.03	0.97	-0.03	0.34	-0.75	0.65	3,752	75
Number of responding agencies	1																
	2+	**0.12**	**1.13**	**0.12**	**0.06**	**0.01**	**0.24**	**4,864**	**97**	**0.25**	**1.28**	**0.25**	**0.10**	**0.07**	**0.44**	**4,914**	**98**
EMS involvement	No																
	Yes	0.05	1.05	0.07	0.24	-0.43	0.50	3,794	76	0.20	1.23	0.20	0.40	-0.62	1.02	3,816	76
Towing involvement	No																
	Yes	-0.17	0.84	-0.15	0.19	-0.54	0.18	4,312	86	0.00	1.00	0.01	0.35	-0.72	0.67	3,713	74
Incident severity[a]	Minor																
	Moderate/severe	**0.15**	**1.16**	**0.14**	**0.26**	**-0.34**	**0.69**	**4,657**	**93**	**0.90**	**2.46**	**0.85**	**0.56**	**0.00**	**2.11**	**4,814**	**96**

Note: [a]Incident severity refers to the extent to which the incident impacted the traffic; [b]Percent of time the variable was selected; CI = Bootstrap Confidence Interval; IIA = Incident Impact Area; Med = Median; N/A = Not Applicable; OR = Odds Ratio; SD = Standard Deviation; Variables in bold are important and significant at the 95% CI.

Table 5-8: Results of the Penalized Logistic Regression Fitted Using Bootstrap Samples (continued)

Variable	Category	HEFT								MSS							
		Mean	OR	Med	SD	CI (%)		Count	%[b]	Mean	OR	Med	SD	CI (%)		Count	%[b]
						2.5	97.5							2.5	97.5		
Intercept	N/A	1.61	N/A	1.61	0.23	1.17	2.08	5,000	100	6.06	N/A	6.05	0.72	4.66	7.48	5,000	100
Temporal Attributes																	
Day of week	Weekday																
	Weekend	**0.20**	**1.22**	**0.21**	**0.08**	**0.03**	**0.34**	**4,666**	**93**	-0.04	0.96	-0.04	0.11	-0.26	0.20	3,357	67
Time of day	Off-peak																
	Morning peak	**0.01**	**1.01**	**0.01**	**0.08**	**-0.14**	**0.17**	**4,765**	**95**	**0.24**	**1.27**	**0.23**	**0.12**	**0.01**	**0.47**	**4,936**	**99**
	Evening peak	**-0.18**	**0.83**	**-0.18**	**0.07**	**-0.30**	**-0.04**	**4,989**	**100**	-0.26	0.77	-0.26	0.13	-0.53	-0.02	4,121	82
Weather Attributes																	
Weather condition	Clear																
	Adverse	**0.34**	**1.40**	**0.34**	**0.27**	**-0.18**	**0.89**	**4,936**	**99**	**0.53**	**1.70**	**0.51**	**0.55**	**-0.45**	**1.66**	**4,545**	**91**
Road surface condition	Dry																
	Wet	**1.24**	**3.47**	**1.24**	**0.30**	**0.66**	**1.84**	**5,000**	**100**	**1.08**	**2.96**	**1.07**	**0.59**	**0.02**	**2.29**	**4,868**	**97**
Roadway Geometric Attributes																	
Shoulder width (feet)	N/A	**0.10**	**1.11**	**0.10**	**0.01**	**0.08**	**0.12**	**5,000**	**100**	**-0.07**	**0.93**	**-0.07**	**0.02**	**-0.11**	**-0.03**	**4,981**	**100**
Presence of horizontal curve within IIA	No																
	Yes	**0.57**	**1.77**	**0.57**	**0.07**	**0.45**	**0.70**	**5,000**	**100**	**0.50**	**1.66**	**0.50**	**0.10**	**0.31**	**0.69**	**5,000**	**100**
Presence of vertical curve within IIA	No																
	Yes	**0.13**	**1.14**	**0.13**	**0.06**	**0.01**	**0.26**	**4,696**	**94**	**0.82**	**2.27**	**0.82**	**0.10**	**0.63**	**1.00**	**5,000**	**100**
Presence of diverge influence area within IIA	No																
	Yes	**0.61**	**1.84**	**0.61**	**0.06**	**0.49**	**0.74**	**5,000**	**100**	**0.29**	**1.34**	**0.29**	**0.10**	**0.09**	**0.49**	**4,995**	**100**
Presence of merge influence area within IIA	No																
	Yes	**0.23**	**1.26**	**0.23**	**0.06**	**0.12**	**0.35**	**5,000**	**100**	**-0.34**	**0.71**	**-0.34**	**0.09**	**-0.52**	**-0.16**	**4,995**	**100**
Presence of major work zone within IIA	No																
	Yes	**0.33**	**1.39**	**0.33**	**0.06**	**0.21**	**0.46**	**5,000**	**100**	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note: [b]Percent of time the variable was selected; CI = Bootstrap Confidence Interval; IIA = Incident Impact Area; Med = Median; N/A = Not Applicable; OR = Odds Ratio; SD = Standard Deviation; Variables in bold are important and significant at the 95% CI.

5.2.2.1 Roadway Geometric Attributes

The following geometric variables along the HEFT section were found to be most important and significant at the 95% CI: diverge influence area, merge influence area, horizontal curve, vertical curve, shoulder width, and presence of work zone. The following five variables along the MSS were found to be most important and credible at the 95% CI: diverge influence area, merge influence area, shoulder width, horizontal curve, and vertical curve. The work zone variable is applicable only for the HEFT study corridor. The positive coefficient for the presence of work zone variable indicates that incidents with impact areas within a work zone are 36% more likely to result in a secondary crash. Work zones are associated with unexpected congestion due to a combination of factors, including daily changes in traffic patterns, narrowed rights-of-way, and complex arrangements of traffic control devices and signs (FHWA, 2007). This situation may explain the reason for the increased risk of secondary crashes in work zone areas.

Incidents with diverge influence areas within their impact areas are more likely to result in secondary crashes. Diverge influence areas are accompanied by more lane changes and high speed differentials because of drivers who are attempting to exit the freeway. This situation increases the risk of secondary crashes, and hence, may serve as a possible explanation for this observation. Conversely, the estimated parameter of the merge influence area in the MSS is negative, implying that incidents with a merge influence area within the impact area are less likely to cause secondary crashes. This observation suggests that unlike diverge influence areas, merge influence areas have a lesser impact on traffic, and hence, a lower likelihood of secondary crashes. Drivers who are merging onto the mainline usually enter at a relatively slower speed than the vehicles traveling on the

mainline. Meanwhile, drivers exiting the freeway usually make several lane changes and slow down to get off the freeway. Previous research indicated that a higher proportion of crashes occur at diverging ramps than at merging ramps, where speeding was reported as a major factor for crashes at diverging ramps (McCartt et al., 2004). Nonetheless, HEFT incidents with merge influence areas within their impact areas are 26% more likely to cause secondary crashes. While the presence of work zone may be a possible explanation for this finding, further research is needed to provide a greater insight into work zone-related crashes.

As indicated in Table 5-8, incidents whose impact area involves a horizontal curve, compared to a tangent section, are more likely to result in secondary crashes. This is expected as the queue along a curved section may not be quickly visible to the upstream drivers. This finding is consistent with previous research findings (Kitali et al., 2019b). A similar observation was found on incidents with a vertical curve within the incident impact area. That is, incidents with elevated sections within the impact area are more likely to cause secondary crashes than those on level sections. The presence of vertical curves may reduce the sight distance, a condition that makes it difficult for upstream drivers to easily recognize the queue built by the initial incident.

The estimated parameter of the MSS shoulder width is negative, implying that a unit increase in shoulder width is accompanied by a 7% decrease in the likelihood of a secondary crash. One possible explanation is that shoulders provide room for veering away from a potential collision. Furthermore, when a platoon of vehicles is suddenly forced to slow down, some drivers in the middle of the platoon who are unaware of the congestion ahead tend to use shoulders for completing the deceleration maneuver. On the other hand,

the estimated parameter of the HEFT shoulder width is positive, implying that a unit increase in shoulder width is accompanied by an 11% increase in the likelihood of a secondary crash. This finding is counterintuitive, and the presence of construction activities on the HEFT corridor may be one possible reason for this observation.

5.2.2.2 Temporal Attributes

Temporal attributes serve as a proxy for traffic flow parameters, such as volume, occupancy, speed, and vehicle mix, as well as driver attitudes and familiarity (Karlaftis et al., 1999). The results in Table 5-8 show that the time of day variable is among the most important variables in the MSS model. Compared to off-peak hours, incidents that occur during morning peak hours are 25% more likely to result in secondary crashes. This finding indicates that secondary crashes are more likely to occur during congested periods. This is because drivers have less space for moving to avoid a collision in congested traffic. Similar findings were observed by previous studies (Kitali et al., 2019b, 2018).

Both the day of the week and time of day variables are among the most important variables in the HEFT model. The results for the day of the week variable indicate that HEFT incidents that occur on weekends, rather than weekdays, have a 9% likelihood of resulting in secondary crashes. On the other hand, compared to off-peak hours, incidents that occur during evening peak hours are 17% less likely to result in secondary crashes. Both findings are inconsistent with previous research findings by Kopitch and Saphores (2011), Xu et al. (2016), and Zhan et al. (2009). The presence of work zone activities on the HEFT may serve as a possible explanation for these findings.

5.2.2.3 Traffic Flow Attributes

The following variables in the HEFT model were identified as the most important and are significant at the 95% CI: mean speed before the incident, mean prevailing speed, and standard deviation of prevailing speed. For the MSS model, the mean prevailing speed and the standard deviation of speed before the incident were among the most important variables. All of these important traffic-related variables are significant at the 95% CI.

As shown in Table 5-8, the negative parameter of the mean prevailing speed indicates that the risk of secondary crashes decreases as the average prevailing speed increases. The decreasing speed represents an increase in traffic density and queue formation. Disturbances caused by the primary incident more easily propagate these queuing traffic formations, leading to an increased risk of secondary crashes. This finding is consistent with the previous studies which reported that the risk of secondary crashes increases with the decrease in average speed (Kitali et al., 2019b; Xu et al., 2016).

The standard deviation of prevailing speed is positively associated with the occurrence of secondary crashes. This result was expected, as a high variation in speed is associated with volatile interactions among vehicles that accelerate and brake frequently (Khattak and Wali, 2017). This situation increases the risk of a secondary crash.

Interestingly, the mean speed before the incident in the HEFT model is positively associated with the likelihood of secondary crashes, meaning that the risk of a secondary crash increases with speed before the incident. The standard deviation of speed before the incident on the MSS corridor is negatively associated with the risk of a secondary crash. A high standard deviation indicates higher variability, and vice versa. This metric was included to assist in capturing the effect of rapid changes in traffic conditions (e.g.,

shockwaves and braking maneuvers) associated with pre-incident conditions. It is worth noting that high and low traffic speeds were associated with low and high variations (standard deviation) in speeds, respectively. That is, if the incident occurred during high traffic speed conditions, then more significant variability in speed is likely to occur as traffic is transitioning from the free-flow state to the congested state, a situation that increases the likelihood of secondary crashes. On the other hand, if the incident occurred during low traffic speed variation (in other words, the average speed is low) the likelihood of a secondary crash was expected to be low because traffic is already in a congested state.

5.2.2.4 Incident Attributes

The most important incident-related variables in the HEFT model include incident severity, lane closure, number of responding agencies, and incident type. Only the number of responding agencies and incident type variables are significant at the 95% CI. Three of the most important incident-related variables in the HEFT model (incident type, number of responding agencies, and lane closure) are also among the most important variables in the MSS model. Compared with vehicle problem-related incidents, hazard-related and crash incidents are more likely to result in a secondary crash. From this finding, it can be inferred that the risk of crash incidents resulting in secondary crashes is two times greater than hazard-related incidents. A similar observation was made in previous research (Kitali et al., 2019b, 2018). A possible explanation for this observation may be related to the extent of impact that different incident types may have on traffic. In general, crashes are expected to have a higher likelihood of resulting in congestion than other incident types, such as hazards and vehicle problems.

As expected, the number of responding agencies was also identified as one of the significant predictor variables that influence the risk of secondary crash occurrence on both the HEFT and MSS corridors. The number of responding agencies is an indicator of the severity of the incident because severe incidents tend to require more responding agencies than less severe incidents. Moreover, incidents attended to by multiple incident responders may require lane closures, a situation that further reduces the capacity of the roadway, resulting in more congestion, and hence, increases the likelihood of a secondary crash. This fact is proven by the positive coefficient of the lane closure variable, which indicates that incidents on the MSS that resulted in lane closure are twice as likely to result in a secondary crash, compared to incidents that did not result in lane closure. Previous research reported a similar finding (Kitali et al., 2019b, 2018).

5.2.2.5 Weather Attributes

The results in Table 5-8 show that wet road surface conditions are positively associated with the risk of secondary crashes on both study corridors, indicating that incidents that occurred on wet road surfaces are two times more likely to result in secondary crashes than those that occurred during dry surface conditions. A similar observation was found in previous research (Xu et al., 2016). This finding is intuitive, as drivers tend to drive more slowly during wet surface conditions than during dry surface conditions, a situation which reduces highway capacity, and hence, increases congestion.

5.3 Leading Causes of Cascading Crashes

5.3.1 Descriptive Statistics

To identify leading causes of cascading crashes, 95,583 incidents from the SunGuide® database and high-resolution speed data from the HERE Technologies were evaluated. A total of 1,008 secondary crashes were identified from 932 primary incidents. This means, 76 primary incidents resulted in more than one secondary crashes. As indicated in Figure 5-11, out of 1,008 incidents that were identified as secondary crashes, 70 occurred within the impact area of 66 secondary crashes and their respective primary incident impact areas. In other words, 6% of primary incidents resulted in a series of cascading crashes.

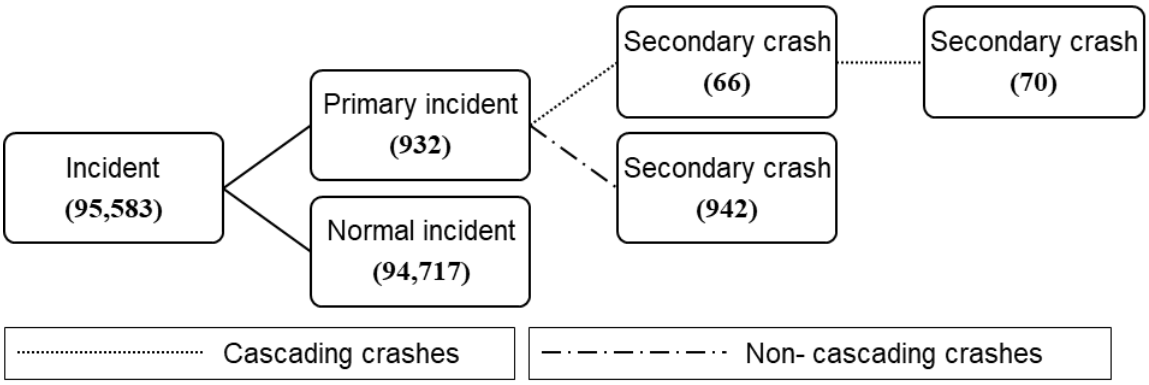


Figure 5-11: Cascading and Non-Cascading Crashes Identified in The Study

Table 5-9 summarizes the list of potential variables that may influence the occurrence of cascading crashes. The following 18 independent variables were included in the analysis:

- traffic flow attributes: mean speed before the incident, standard deviation of speed before the incident, mean prevailing speed, and standard deviation of prevailing speed;
- temporal-related variables: time of day and day of the week;
- weather-related variables: rainfall;

- incident-related attributes: incident type, number of responding agencies, EMS involvement, towing involvement, lane closure, and incident severity; and
- geometric attributes: shoulder width, presence of horizontal curve within the incident impact area, presence of vertical curve within the incident impact area, presence of diverge influence area within the incident impact area, and presence of merge influence area within the incident impact area.

Table 5-9: Descriptive Statistics of Potential Variables Influencing the Occurrence of Cascading Crashes

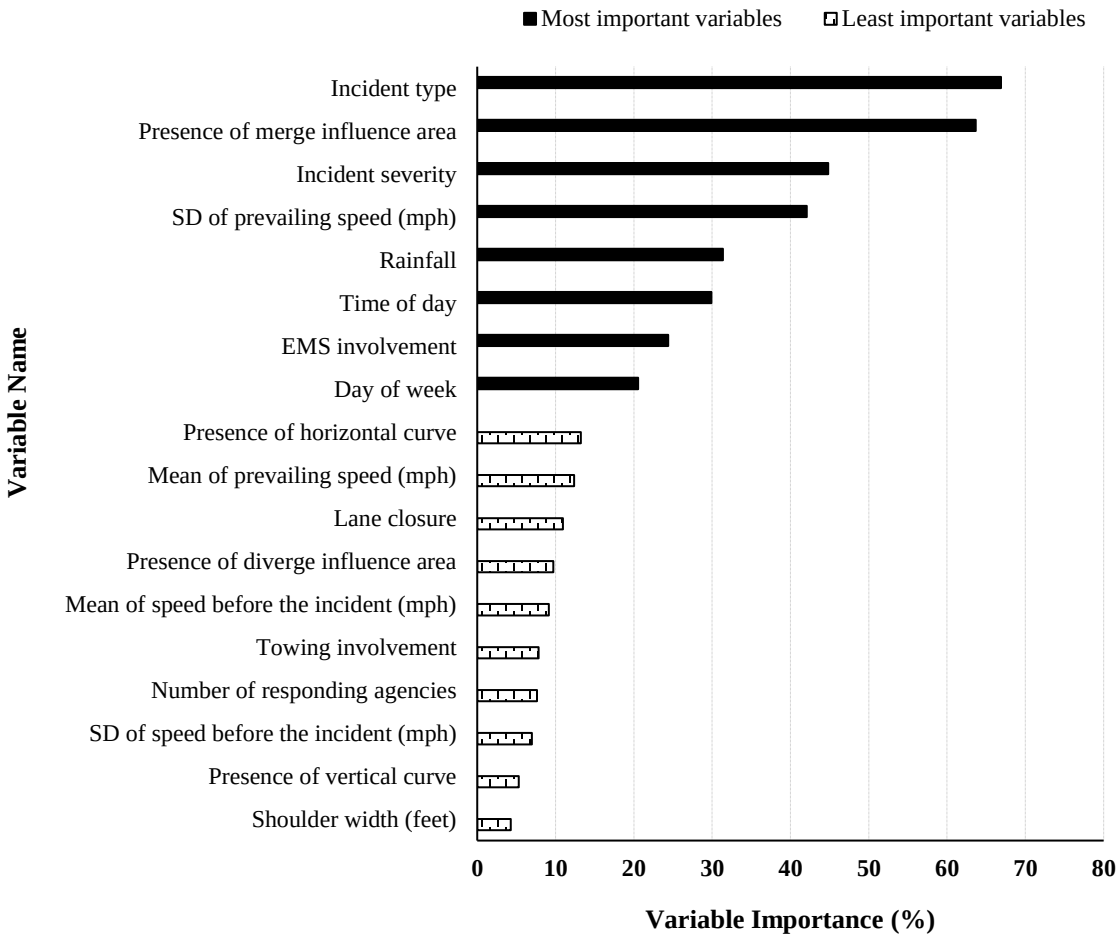
Attribute	Attribute Category	Count	Percentage (%)
Traffic Flow Attributes			
Mean speed before the incident (mph)	Low: ≤ 30	166	18
	Moderate: 30-55	245	26
	High: > 55	520	56
SD of speed before the incident (mph)	Low: ≤ 1	241	26
	Moderate: 1-4	463	50
	High: > 4	227	24
Mean prevailing speed (mph)	Low: ≤ 25	171	18
	Moderate: 25-45	300	32
	High: >45	460	49
SD of prevailing speed (mph)	Low: ≤ 7	413	44
	Moderate: 7-13	242	26
	High: > 13	276	30
Incident Attributes			
Incident type	Vehicle problem	457	49
	Hazard	125	13
	Crash	349	37
Number of responding agencies	1	482	52
	2+	449	48
EMS involvement	No	813	87
	Yes	118	13
Towing involvement	No	805	86
	Yes	126	14
Lane closure	No	745	80
	Yes	186	20
Incident severity	Minor	831	89
	Moderate/severe	100	11
Temporal Attributes			
Day of week	Weekday	811	87
	Weekend	120	13
Time of day	Off-peak	263	28
	Morning peak	302	32
	Evening peak	366	39

Table 5-9: Descriptive Statistics of Potential Variables Influencing the Occurrence of Cascading Crashes (continued)

Attribute	Attribute Category	Count	Percentage (%)
Weather Attributes			
Rainfall Intensity	No/light	843	91
	Medium/heavy	88	9
Roadway Geometric Attributes			
Shoulder width (feet)	≤ 10	315	34
	> 10	616	66
Presence of horizontal curve within the incident impact area	No	239	26
	Yes	692	74
Presence of vertical curve within the incident impact area	No	276	30
	Yes	655	70
Presence of diverge influence area within the incident impact area	No	294	32
	Yes	637	68
Presence of merge influence area within the incident impact area	No	264	28
	Yes	667	72
Likelihood of cascading crashes	No	871	94
	Yes	60	6

5.3.2 Important Variables that Influence the Likelihood of Cascading Crashes

Penalized logistic regression was used to identify the most important variables. The coefficient was obtained by calculating the mean of all estimates of the models fitted in the bootstrap samples. The odds ratio (OR), which represents how the dependent variable varies with the predictor variable, was also calculated as the exponent of the predictor coefficient. Table 5-10 shows the results of the penalized logistic regression model and the number of times the variable was selected in the model as an important variable. Figure 5-12 shows the results of the variable importance ranking based on the percentage of times a variable was selected. The bolded bars represent variables that were among the top 20% of the most important variables. These variables include incident type, presence of merge influence area within incident impact area, incident severity, standard deviation of prevailing speed, rainfall, EMS involvement, time of day, and day of the week.



Note: EMS = Emergency Medical Service; SD = Standard Deviation.

Figure 5-12: Selection of the Important Variables for Cascading Crash Likelihood Model

5.3.2.1 Traffic Flow Attributes

Results indicate that the standard deviation of the prevailing speed variable was one of the most important variables. The coefficient in Table 5-10 shows a 67% higher likelihood of a cascading crash to occur when the variation in prevailing speed is moderate rather than low. This finding was expected, as the greater the speed variance, the greater the number of interactions among vehicles that accelerate and brake frequently. This situation exacerbates the risk of an incident resulting in a series of cascading crashes.

Conversely, the risk of a cascading crash occurring when the variation in prevailing speed is higher decreases by 30% compared to when the variation is low.

Table 5-10: Results of the Penalized Logistic Regression Fitted Using Bootstrap Samples

Variable	Category	Mean	OR	Median	SD	CI (%)		Count	%*
						5.0	95.0		
Intercept	N/A	-0.02	0.98	-0.04	0.34	-0.53	0.56	5000	100.0
Traffic Flow Attributes									
Mean speed before the incident (mph)	Moderate: 30-55								
	Low: ≤ 30	0.33	1.40	0.31	0.26	0.02	0.76	458	9.2
	High: > 55	0.29	1.33	0.33	0.37	-0.34	0.83	157	3.1
SD of speed before the incident (mph)	Moderate: 1-4								
	Low: ≤ 1	0.25	1.29	0.29	0.31	-0.32	0.69	199	4.0
	High: > 4	0.29	1.33	0.28	0.29	-0.23	0.77	348	7.0
Mean prevailing speed (mph)	Moderate: 25-45								
	Low: ≤ 25	0.46	1.59	0.44	0.28	0.05	0.96	747	14.9
	High: > 45	-0.28	0.75	-0.26	0.20	-0.65	-0.02	618	12.4
SD of prevailing speed (mph)	Moderate: 7-13								
	Low: ≤ 7	**0.51**	**1.67**	**0.48**	**0.27**	**0.09**	**0.99**	**2104**	42.1
	High: > 13	-0.36	0.70	-0.34	0.28	-0.85	-0.01	397	7.9
Incident Attributes									
Incident type	Vehicle problem								
	Hazard	-0.24	0.79	-0.25	0.37	-0.77	0.46	583	11.7
	Crash	**0.54**	**1.72**	**0.51**	**0.30**	**0.10**	**1.08**	**3344**	**66.9**
Lane closure	No								
	Yes	0.44	1.55	0.39	0.36	0.03	1.10	547	10.9
Number of responding agencies	1								
	2+	-0.55	0.58	-0.54	0.28	-1.04	-0.08	381	7.6
EMS involvement	No								
	Yes	**0.53**	**1.70**	**0.48**	**0.39**	**0.04**	**1.23**	**1219**	**24.4**
Towing involvement	No								
	Yes	0.41	1.50	0.39	0.51	-0.55	1.18	391	7.8
Incident severity	Minor								
	Moderate/severe	**0.68**	**1.97**	**0.61**	**0.46**	**0.07**	**1.52**	**2241**	**44.8**

Note: *Percent of times a variable is selected as an important variable; Variables in bold are important and significant at the 90% credible interval; IIA = Incident Impact Area; OR = Odds ratio; Med = Median; CI = Credible Interval.

Table 5-10: Results of the Penalized Logistic Regression Fitted Using Bootstrap Samples (continued)

Variable	Category	Mean	OR	Median	SD	CI (%)		Count	%*
						5.0	95.0		
Temporal Attributes									
Day of week	Weekday								
	Weekend	**-0.47**	**0.63**	**-0.46**	**0.33**	**-1.01**	**-0.03**	**1027**	**20.5**
Time of day	Off-peak								
	Morning peak	**0.50**	**1.64**	**0.46**	**0.27**	**0.09**	**1.00**	**1496**	**29.9**
	Evening peak	-0.02	0.98	-0.04	0.39	-0.65	0.72	104	2.1
Weather Attributes									
Rainfall	No/light								
	Medium/heavy	**0.68**	**1.97**	**0.61**	**0.49**	**0.08**	**1.52**	**1569**	**31.4**
Roadway Geometric Attributes									
Shoulder width (feet)	≤ 10								
	> 10	-0.10	0.90	-0.17	0.41	-0.69	0.60	215	4.3
Presence of horizontal curve within IIA	No								
	Yes	-0.39	0.68	-0.35	0.31	-0.94	-0.01	662	13.2
Presence of vertical curve within IIA	No								
	Yes	-0.11	0.90	-0.17	0.40	-0.67	0.63	266	5.3
Presence of diverge influence area within IIA	No								
	Yes	-0.33	0.72	-0.31	0.27	-0.78	0.00	486	9.7
Presence of merge influence area within IIA	No								
	Yes	**-0.57**	**0.57**	**-0.55**	**0.29**	**-1.08**	**-0.13**	**3184**	**63.7**

Note: *Percent of times a variable is selected as an important variable; Variables in bold are important and significant at the 90% credible interval; IIA = Incident Impact Area; OR = Odds ratio; Med = Median; CI = Credible Interval.

A high standard deviation indicates a higher variability, and vice versa. This metric was included to assist in capturing the effect of rapid changes in traffic conditions (e.g., shockwaves and braking maneuvers) associated with pre-incident conditions.

It is worth noting that high traffic speeds are associated with low standard deviations, whereas low traffic speeds have high speed variations. That is, if an incident occurred when traffic speeds were high, higher variability in speed was likely to occur as traffic transitioned from a free-flow state to a congested state, a situation that increases the

likelihood of secondary crashes. On the other hand, if an incident occurred when the variation of the traffic speed estimates was high (in other words, the average speed was low), the likelihood of additional crashes to occur is expected to be low since traffic is already in a congested state, and a significant variation in speed is not expected.

5.3.2.2 Incident Attributes

Important incident-related variables included incident type, EMS involvement, and incident severity. Results suggest that compared to vehicle problems, crashes are 72% more likely to result in cascading crashes. A possible explanation for this observation may be related to the extent of impact different incident types may have on traffic. In general, crashes are expected to have a higher likelihood of resulting in congestion than other incident types, such as hazards and vehicle problems.

The probability of incidents with EMS involvement resulting in cascading crashes is 70% higher than those that did not have EMS as one of the responding agencies. The presence of EMS as one of the responding agencies may serve as an indicator of the severity of an incident. EMS responses often result in lane closures, further reducing the capacity of the roadway and resulting in more congestion, which increases the likelihood of cascading crashes. This fact is proven by the positive coefficient of the incident severity variable, which indicates that incidents with moderate/high severity are 97% more likely to result in cascading crashes.

5.3.2.3 Temporal Attributes

All temporal characteristics were selected as important variables. It was observed that cascading crashes are 37% less likely to occur on weekends than on weekdays. Results also indicate that cascading crashes are 64% more likely to occur during morning peak

hours than off-peak hours. This observation implies that incidents occurring during congested periods are more likely to cause cascading crashes. Congested traffic is characterized by smaller gaps between vehicles, providing drivers with less room for maneuvering to avoid a crash. While previous studies on the likelihood of secondary crashes indicated that secondary crashes are more likely to occur on weekdays and during peak hours, it can be inferred from findings from this research that, compared to secondary crashes, cascading crashes are even more likely to occur under these conditions.

5.3.2.4 Weather Attributes

Rainfall increases the likelihood of cascading crashes by 97%. This is intuitive, as drivers tend to drive more slowly when it is raining, a situation that reduces highway capacity, and hence, increases congestion. Previous research indicated that rain increases the traffic breakdown process, a situation that exacerbates the occurrence of additional crashes (Kidando et al., 2019a).

5.3.2.5 Roadway Geometric Attributes

Incidents where merge influence is within the incident impact area are 45% less likely to result in a series of cascading crashes. There was no possible explanation for this observation. Further research can assist in providing insight into this finding.

5.3.3 Discrete Bayesian Network results

Figure 5-13 illustrates the optimal Bayesian network structure that was developed from the analyses. The hybrid approach revealed that four nodes were directly associated (dependence relationship) with cascading crash likelihood. These factors are also referred to as hypothesis variables. As can be inferred from Figure 5-13, the four variables that were

found to have a direct probabilistic relationship with the likelihood of cascading crashes are the standard deviation of prevailing speed, incident severity, rainfall, and day of the week.

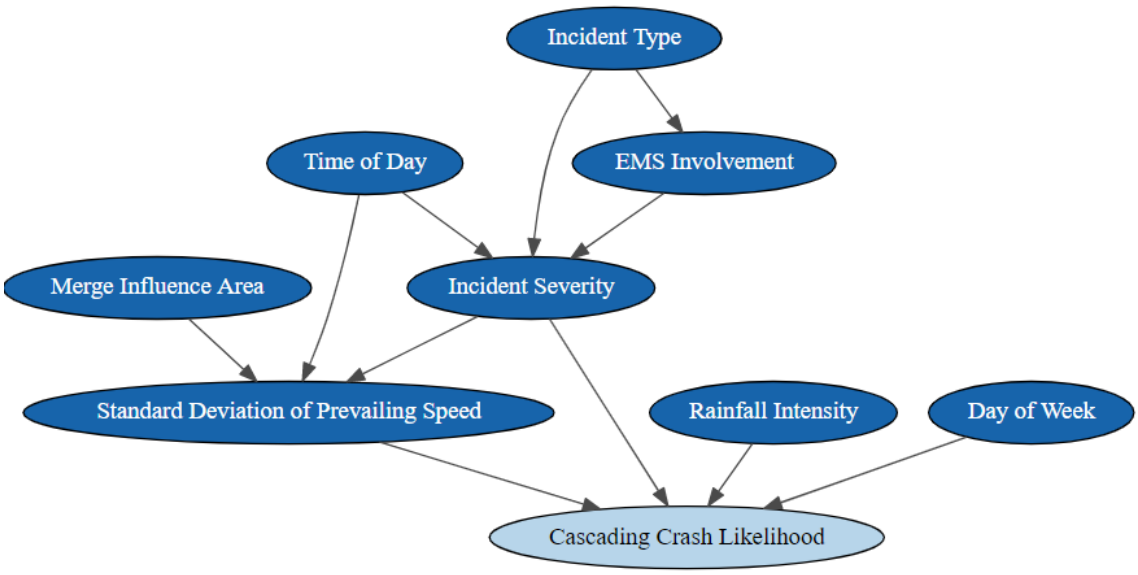


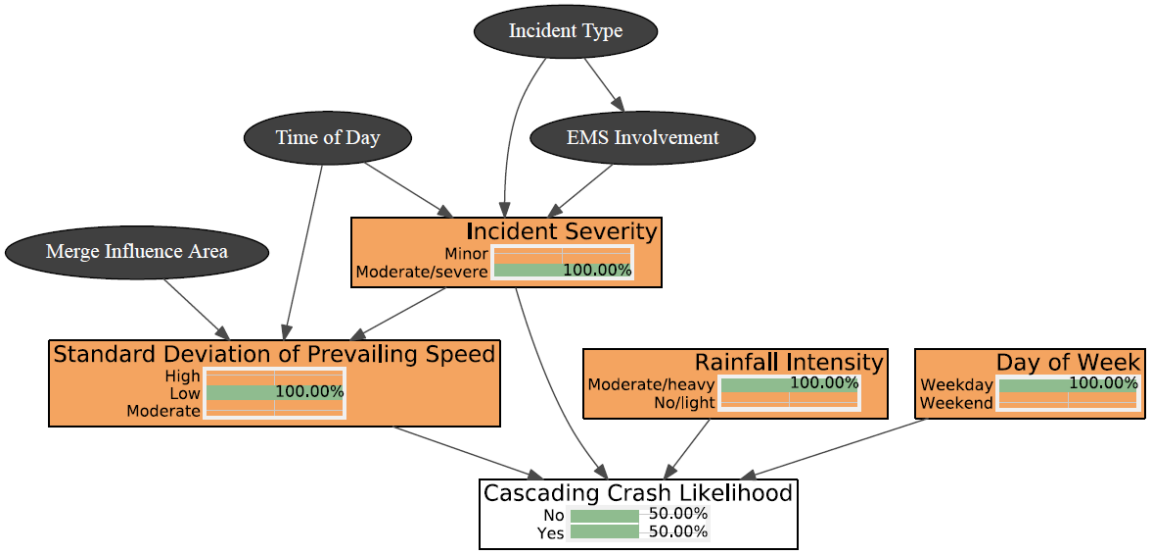
Figure 5-13: Optimal Bayesian Network Structure

Based on the optimal Bayesian network structure shown in Figure 5-13, the impact of concurrent evidence on the likelihood of cascading crashes was assessed. The analysis focused on variables that have a direct association with cascading crash likelihood. Of the 24 combinations, Table 5-11 provides the results of the top 20% combinations that had a higher predicted probability of cascading crashes than all other combinations.

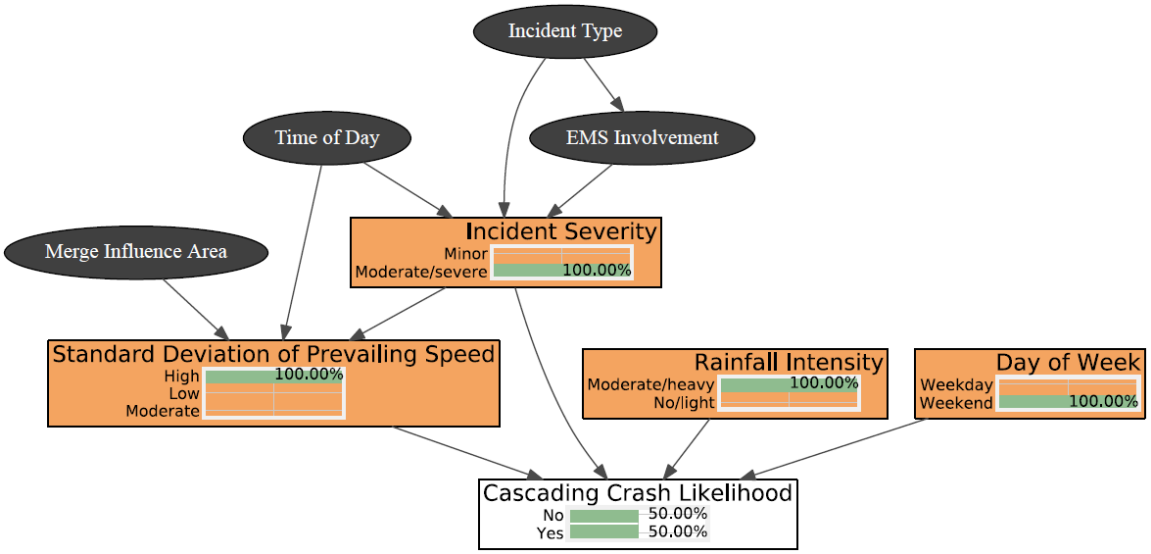
Table 5-11: Predicted Probability of Cascading Crashes

Cascading Crash Likelihood				
Predictor Variable				Predicted Probability
Incident Severity	**Rainfall Intensity**	**Day of Week**	**SD of Prevailing Speed**	
Moderate/severe	Moderate/heavy	Weekday	Low	50.00
Moderate/severe	Moderate/heavy	Weekend	High	50.00
Moderate/severe	No/light	Weekend	Low	28.57
Minor	Moderate/heavy	Weekend	Moderate	25.00
Moderate/severe	No/light	Weekend	Moderate	25.00
Moderate/severe	Moderate/heavy	Weekday	High	25.00
Moderate/severe	No/light	Weekday	Moderate	23.08

Figure 5-14 shows the Bayesian network structure with the combination of evidence that resulted in the highest likelihood of cascading crashes.



(a) First combination



(b) Second combination

Figure 5-14: Combined Evidence Sensitivity Analysis

From Figure 5-14, it can be inferred that cascading crashes are more likely to occur when the prior incident occurs on a weekday, when it is moderately or heavily raining, there is low variation in prevailing speed, and the incident resulted in a moderate/severe

impact on traffic. Similarly, cascading crashes are more likely to be caused by a moderate/severe incident that occurs when it is raining, on a weekend, and when the variation in prevailing speed is high. From these two findings, it may be concluded that cascading crashes are more likely to occur when traffic is in the transition state, i.e., when there is a platoon of vehicles traveling at high differential speeds. However, once the traffic is in a congested state, i.e., the variation in speed reduces significantly, then the likelihood of cascading crashes also decreases.

5.4 Secondary Crash Risk Prediction

5.4.1 Descriptive Statistics

A dynamic Bayesian cloglog model was developed to predict the likelihood of secondary crashes. Once an incident has occurred, traffic conditions upstream of the incident change with time, and so does the likelihood of secondary crashes. A 5-minute time interval was used from the time when the primary incident occurred to when the secondary crash occurred and from the time when the normal incident started impacting traffic to the time the traffic returned to normal. A total of 50 models were fitted.

Incident data from the MSS corridor were used to predict the likelihood of secondary crashes. About 66% of these incidents did not have an impact on traffic. For these incidents, the speed data and rainfall data for the first 10 minutes following the occurrence of the incidents were used. Thus, the models for the first two timestamps may be different from the models for the rest of the timestamps. Table 5-12 shows the distribution of the primary incidents and normal incidents used to fit the 50 models.

Table 5-12: Distribution of Primary Incident and Normal Incidents used in the Dynamic Model

Model	Time interval (minutes)	Normal incidents	Primary incidents	Total incidents	Proportion of primary incidents
m1	0-5	92,851	971	93,822	1%
m2	5-10	86,245	934	87,179	1%
m3	10-15	21,400	899	22,299	4%
m4	15-20	18,845	867	19,712	4%
m5	20-25	16,990	836	17,826	5%
m6	25-30	15,564	809	16,373	5%
m7	30-35	14,406	784	15,190	5%
m8	35-40	13,417	759	14,176	5%
m9	40-45	12,590	736	13,326	6%
m10	45-50	11,854	707	12,561	6%
m11	50-55	11,242	687	11,929	6%
m12	55-60	10,720	662	11,382	6%
m13	60-65	10,197	635	10,832	6%
m14	65-70	9,733	614	10,347	6%
m15	70-75	9,302	586	9,888	6%
m16	75-80	8,918	563	9,481	6%
m17	80-85	8,568	533	9,101	6%
m18	85-90	8,228	513	8,741	6%
m19	90-95	7,898	480	8,378	6%
m20	95-100	7,598	459	8,057	6%
m21	100-105	7,346	445	7,791	6%
m22	105-110	7,089	423	7,512	6%
m23	110-115	6,897	404	7,301	6%
m24	115-120	6,689	389	7,078	5%
m25	120-125	6,517	367	6,884	5%
m26	125-130	6,354	351	6,705	5%
m27	130-135	6,190	333	6,523	5%
m28	135-140	6,028	320	6,348	5%
m29	140-145	5,894	307	6,201	5%
m30	145-150	5,756	293	6,049	5%
m31	150-155	5,633	285	5,918	5%
m32	155-160	5,527	265	5,792	5%
m33	160-165	5,425	250	5,675	4%
m34	165-170	5,317	241	5,558	4%
m35	170-175	5,221	232	5,453	4%
m36	175-180	5,129	221	5,350	4%
m37	180-185	5,041	215	5,256	4%
m38	185-190	4,935	208	5,143	4%
m39	190-195	4,871	205	5,076	4%
m40	195-200	4,794	197	4,991	4%
m41	200-205	4,706	189	4,895	4%
m42	205-210	4,640	182	4,822	4%
m43	210-215	4,564	174	4,738	4%
m44	215-220	4,493	168	4,661	4%
m45	220-225	4,423	161	4,584	4%
m46	225-230	4,342	156	4,498	3%
m47	230-235	4,277	149	4,426	3%
m48	235-240	4,214	147	4,361	3%
m49	240-245	4,167	142	4,309	3%
m50	245-250	4,117	139	4,256	3%

Note: For the first 10 minutes, all the normal incidents that did not have an impact on traffic were included – and hence the numbers are high; m = Model.

The following 13 explanatory variables were used in the model: (1) mean speed before the incident, (2) standard deviation of speed before the incident, (3) mean prevailing speed, (4) standard deviation of prevailing speed, (5) incident type, (6) day of week, (7) time of day, (8) rainfall, (9) shoulder width, (10) presence of horizontal curve within the incident impact area, (11) presence of vertical curve within the incident impact area, (12) presence of diverge influence area within the incident impact area, and (13) presence of merge influence area within the incident impact area. Note that, other than incident type, the incident-related attributes that were identified as the most important variables in Section 5.2 (i.e., lane closure and number of responding agencies), are not included in this model since it was not clear at what time these variables were reported after the incident occurred. Since these two variables, i.e., lane closure and number of responding agencies, can be considered a surrogate measure of congestion, the temporal attributes (time of day and day of the week) were used instead.

5.4.2 Cloglog Model Results

To build the model, the first step involved defining the prior distribution. Non-informative priors were specified only in the first model since there was no previous information to generate the informative prior distributions from. For the subsequent models, the prior distributions were estimated using the posterior distributions of the immediate previous model. In this way, the coefficients of the subsequent models will be influenced by both prior information and present information.

Table 5-13 provides a posterior summary of the model. In Table 5-13, the descriptive statistics, i.e., mean, median, and standard deviation of the variable coefficients,

were derived from the 50 fitted models. The percentage of times these coefficients were significant at the 95% Bayesian Credible Interval (BCI) is also presented.

Table 5-13: Posterior Summary of Cloglog Model Results

Variable	Category	Mean	Median	SD	Percent of time it was significant
Intercept	N/A	-1.91	-1.69	0.83	100%
Traffic Flow Attributes					
Mean speed before the incident (mph)	N/A	**-0.12**	**-0.15**	**0.08**	**88%**
SD of speed before the incident (mph)	N/A	**0.14**	**0.17**	**0.05**	**82%**
Mean prevailing speed (mph)	N/A	**-0.23**	**-0.16**	**0.17**	**74%**
SD of prevailing speed (mph)	N/A	0.06	0.06	0.05	36%
Incident Attributes					
Incident type	Vehicle problem				
	Hazard	0.07	0.11	0.09	0%
	Crash	**0.53**	**0.53**	**0.08**	**94%**
Temporal Attributes					
Day of week	Weekday				
	Weekend	-0.01	0.01	0.10	0%
Time of day	Off-peak				
	Morning peak	**0.66**	**0.67**	**0.12**	**100%**
	Evening peak	0.21	0.22	0.08	18%
Weather Attributes					
Rainfall	No/light				
	Moderate/heavy	**0.71**	**0.72**	**0.15**	**76%**
Roadway Geometric Attributes					
Shoulder width (feet)	N/A	**0.34**	**0.39**	**0.16**	**98%**
Presence of horizontal curve within IIA	No				
	Yes	**1.25**	**1.20**	**0.23**	**100%**
Presence of vertical curve within IIA	No				
	Yes	-0.01	-0.03	0.21	6%
Presence of diverge influence area within IIA	No				
	Yes	**-1.24**	**-1.27**	**0.41**	**96%**
Presence of merge influence area within IIA	No				
	Yes	**-2.03**	**-2.06**	**0.23**	**100%**

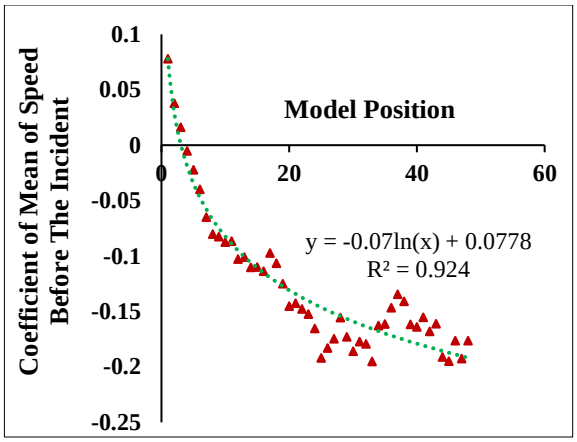
Note: Variables in bold were significant more than 90% of the times the models were fitted * Represents 95% Bayesian Credible Interval; IIA = Incident Impact Area; N/A = Not Applicable; SD = Standard Deviation.

Coefficients of the following 10 variables were found to be significant more than 70% of the time the models were fitted: mean speed before the incident, standard deviation

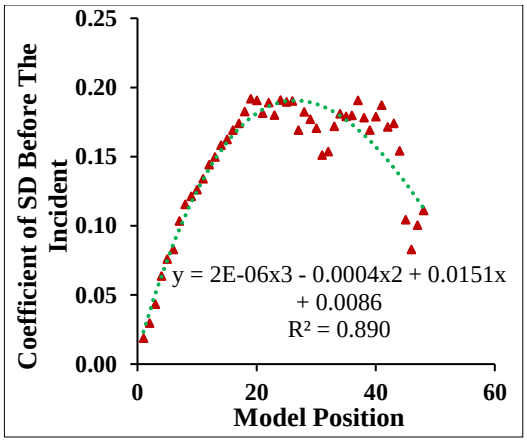
of speed before the incident, mean prevailing speed, incident type (crash), time of day (morning peak hours), rainfall, shoulder width, presence of horizontal curve within the incident impact area, presence of merge influence area within the incident impact area, and presence of diverge influence area within the incident impact area.

The signs of some of the coefficients are comparable to those presented in Section 5.2. These include the mean prevailing speed, incident type (crash), time of day (morning peak hours), rainfall (i.e., similar to weather condition and road surface condition), presence of horizontal curve within the incident impact area, and presence of merge influence area. The signs of the coefficients for the remaining four variables, i.e., mean speed before the incident, standard deviation of speed before the incident, shoulder width, and presence of diverge influence area within the incident impact area, are opposite of those presented in Section 5.2.

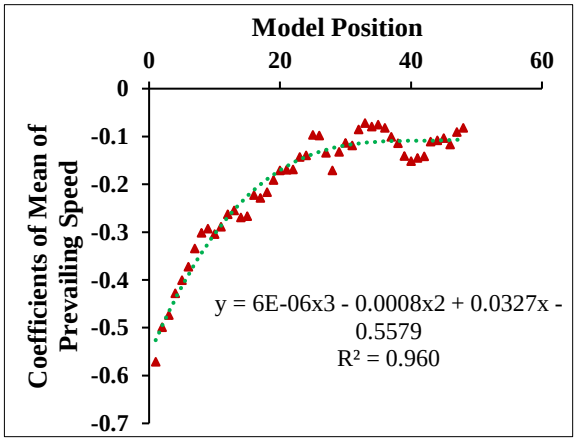
Figure 5-15 presents the plots of estimated coefficients. The best-fitted curve, along with the equation and the R-squared value of the fitted curves, are also presented in Figure 5-15. Most of the fitted curves are polynomials of different degrees, and one curve is exponential (Figure 5-15(a)). The R-squared values of the fitted curves range between 0.177 and 0.966. Note that the coefficients of the first two models (models fitted with variables collected within ten minutes since incidents started impacting traffic) were found to be distinctively different from the remaining model coefficients, and hence, excluded from the plots. This difference could be attributed to incidents without impacts, and whose prevailing traffic and rainfall intensity were collected for 10 minutes after the occurrence of the incident.



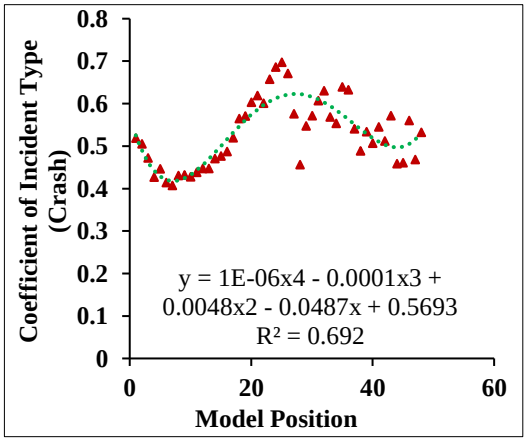
(a) Mean speed before the incident



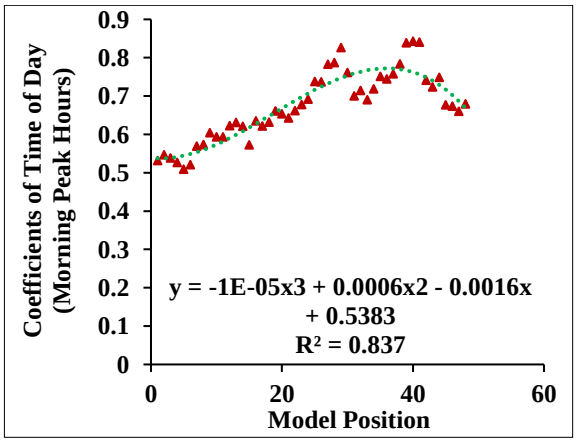
(b) SD of speed before the incident



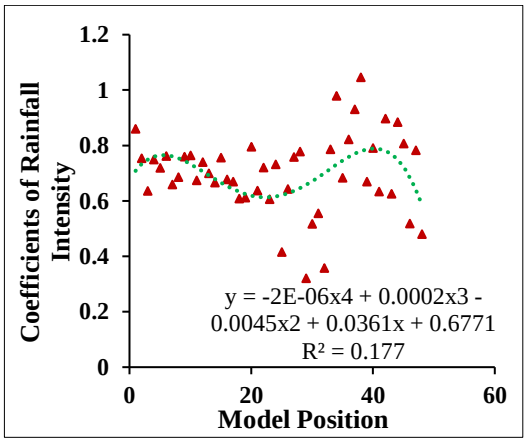
(c) Mean prevailing speed



(d) Incident type (crash)

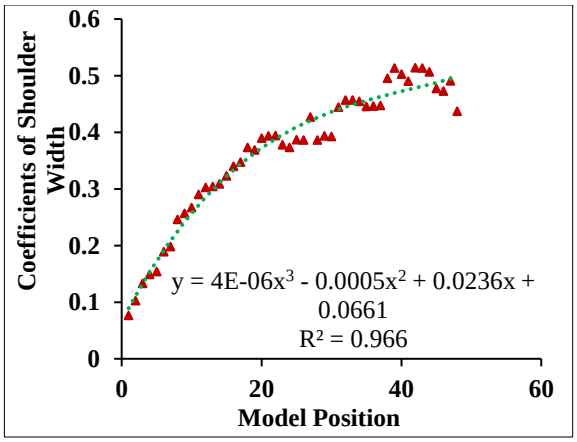


(e) Time of day (Morning peak hours)

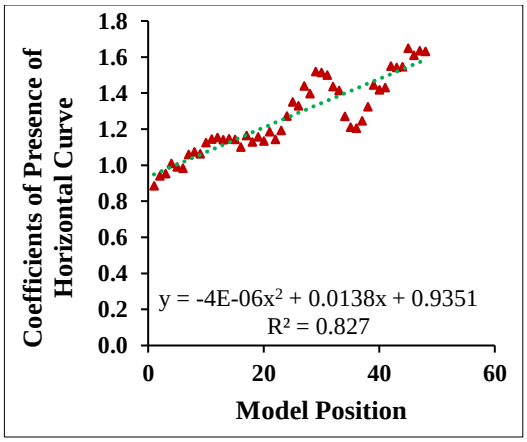


(f) Rainfall (moderate/heavy)

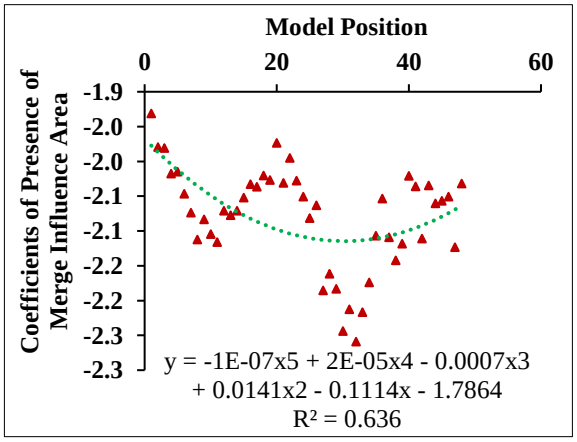
Figure 5-15: Estimated Coefficients for the Series of Fifty Cloglog Models



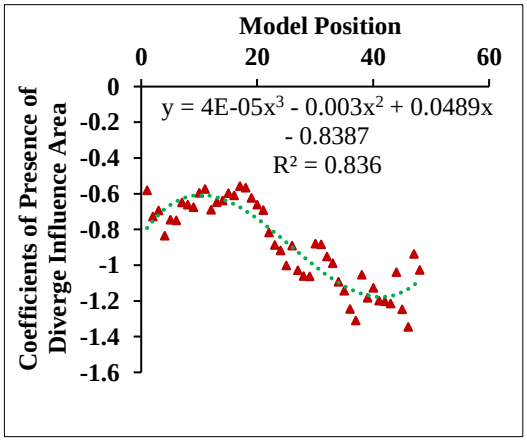
(g) Shoulder width



(h) Presence of horizontal curve



(i) Presence of merge influence area



(j) Presence of diverge influence area

Figure 5-15: Estimated Coefficients for the Series of Fifty Cloglog Models (continued)

The coefficients of the mean speed before the incident for the first five models (25 minutes from when incidents started impacting traffic) are positive, while the coefficients of the remaining subsequent models are negative. The positive coefficients indicate that secondary crashes are more likely to occur when the mean speed before the incident is high. In other words, the negative coefficients indicate that secondary crashes are more likely to occur when the mean speed before the incident is low. It is worth noting that the magnitude of the coefficients of the mean speed before the incident sharply decreases from the first five minutes the incidents started impacting traffic (see Figure 5-15(a)). The

magnitude of the negative coefficients sharply increases from the 30th minute from when the incidents started impacting traffic until 2.5 hours had passed, where the slope of the magnitude of the coefficients becomes flatter. This implies that incidents that occur in a free-flow traffic state are less likely to result in secondary crashes if they are cleared quickly. On the other hand, incidents that occur in less congested traffic and are not cleared in a timely manner are more likely to result in congestion over time, and hence, increase the likelihood of secondary crashes.

The coefficients for the standard deviation of speed before the incident is positive for all the 50 fitted models. This finding implies that secondary crashes are more likely to occur when the standard deviation of the speed before the incident is high. Overall, the magnitude of the impact of variation of speed before the incident on secondary crash likelihood is observed to increase with time. This was expected, as a high variation in speed is associated with volatile interactions among vehicles that accelerate and brake frequently (Khattak and Wali, 2017). This situation exacerbates the risk of a secondary crash. As indicated in Figure 5-15(b) the magnitude of the coefficients increased sharply within the first 75 minutes after incidents started impacting traffic. From the 75th minute (model 15), the slope becomes flatter, and eventually, the magnitude of the coefficients started decreasing from the 100th minute (model 20). This observation may be an indication of the relationship between the likelihood of secondary crashes and the evolution of the traffic flow states.

The coefficients of the mean prevailing speed were found to be negative in all of the fitted models. As mentioned earlier, a negative parameter for the mean prevailing speed indicates that the risk of secondary crashes decreases as the mean prevailing speed

increases. As shown in Figure 5-15(c), the magnitude of these coefficients decreases sharply with time until the 125th minute (model 25) from when the incidents started impacting traffic, where the slope becomes flatter. The decreasing speed represents an increase in traffic density and queue formation, and hence, may explain the pattern observed in Figure 5-15(c).

Figure 5-15(d) shows the plots of the coefficients of crash incidents. The positive sign of these coefficients indicates that, compared to vehicle problems, crash incidents are more likely to cause secondary crashes. It is interesting to observe a continuous decrease in the magnitude of the coefficients up to the 65th minute from when incidents started impacting traffic, where the magnitude of the coefficients started to increase again. The magnitude of the coefficients increased until the 130th minute and started to decrease once more. An explanation for this finding could not be determined. However, further research can assist in providing insight into this finding.

As expected, the sign of the coefficients for the time of day is positive, indicating that incidents that occur during morning peak hours are more likely to cause secondary crashes. Overall, the magnitude of these coefficients increased with time and eventually start to decrease after 200 hours from when the incidents started impacting traffic. This observation implies that incidents occurring during congested periods are more likely to induce traffic. Similar findings were also found in previous studies (Kitali et al., 2019b, 2018; Mishra et al., 2016). However, when the traffic becomes overly congested, e.g., there is little significant variation in traffic, the likelihood of secondary crashes eventually decreases.

The results in Figure 5-15(f) show that the coefficients for the rainfall variable are positive for all of the models, indicating an increased likelihood of secondary crashes during moderate/heavy rainfall. This finding was expected since rainfall tends to increase traffic breakdown and reduce roadway capacity. Specifically, when it rains, traffic slows down because of hydroplaning, a condition that occurs when a layer of water builds between the tires and the road surface leading to friction loss between the two surfaces, and reduced visibility caused by rain on the windshields and water spray from other vehicles (Kidando et al., 2019a). The increased traffic congestion caused by rainfall results in a higher likelihood of secondary crashes.

The estimated coefficients for the shoulder width are presented in Figure 5-15(g). While the sign of the coefficients for the first two models (i.e., models fitted with variables collected within ten minutes since incidents started impacting traffic) is negative, the coefficients for the remaining models are positive. Overall, the impact of shoulder width on secondary crash risk increased with time. This observation is counterintuitive to the findings presented in Section 5.2.

The signs of coefficients for the presence of horizontal curves within the incident impact area are positive. This implies that there is a higher likelihood of secondary crashes when a curved segment (rather than a straight segment) is within the incident impact area. As shown in Figure 5-15(h), this impact is observed to increase with time. This was expected as the queue along a curved section may not be quickly visible to the upstream drivers. A similar finding was observed in previous research (Kitali et al., 2019).

The signs of the coefficients for the presence of merge influence areas within the incident impact area are negative, although the magnitude of the first two models (i.e.,

models fitted with variables collected within ten minutes since incidents started impacting traffic) is exceptionally higher than the remaining models. Meanwhile, the signs of the coefficients for the presence of diverge influence area for the first two models are positive and negative for the remaining models. Overall, the magnitude of both merge and diverge influence area coefficients increases with time. This implies that the influence of these variables on secondary crash occurrence decreases with time. Both merge and diverge influence areas are accompanied by more lane changes and high speed differentials because of drivers attempting to enter and exit the freeway, respectively. However, as congestion increases, speed variation decreases simultaneously. When vehicles are moving at an approximately similar speed, the likelihood of secondary crashes decreases.

5.5 Summary

This research investigated approaches to mitigate secondary crashes on freeways. To implement this goal, approaches were proposed to identify, analyze, and predict secondary crashes in real-time. A data-driven approach was used to identify secondary crashes. To improve the accuracy of the detected secondary crashes, the proposed method considered the fact that the queue built by the primary incident grows and dissipates at a different rate along the roadway segment impacted by the incident. The analysis was based on 322,259 traffic incidents that occurred along the study corridors between January 2014 and June 2019. Overall, 4,549 secondary crashes in the upstream direction of the primary incident were identified from 3,977 primary incidents. The identified secondary crashes on the upstream direction of the primary incident accounted for 1.4% of the 322,259 incidents. This is an equivalent of 5.7 secondary crashes per mile per year.

Next, the LASSO penalized estimator was used to extract the most important explanatory variables, with minimal correlation, influencing the risk of secondary crashes. Because the proportion of primary incidents is smaller than the proportion of normal incidents, the bootstrap resampling method was used to fit the penalized logistic regression. The proposed model is considered to improve the predictive accuracy of the secondary crash risk model because it accounts for the asymmetric nature of secondary crashes, performs variable selection, and removes highly correlated variables.

The influence of potential variables that were rarely considered in previous studies, i.e., work zone, vertical curve, merge influence area, and diverge influence area, were explored. The model results indicate that the presence of work zones significantly influence the likelihood of secondary crashes. Overall, as expected, roadway geometric, temporal, traffic flow, incident, and weather attributes were found to influence the occurrence of secondary crashes.

Using the Bayesian network, the influence of concurrent factors in the likelihood of cascading crashes was investigated. The prediction inference using the optimal Bayesian network indicated the following four variables have a direct probabilistic relationship with the likelihood of cascading crashes: standard deviation of prevailing speed, incident severity, rainfall, and day of the week. Cascading crashes were found to be most likely to occur when the prior incident occurs during moderate/rainy conditions, on a weekday, under low variation in prevailing speed, and if the incident resulted in a moderate/severe impact on traffic. Also, cascading crashes were more likely to occur on a weekend, during moderate or heavy rainfall, under high variation in prevailing speed, and the primary incident resulted in a moderate/severe impact on traffic.

The identified secondary crash influential factors were used in the prediction model. The dynamic Bayesian cloglog model was used to predict the risk of secondary crashes every five minutes following the occurrence of the incident. The coefficients of the following eight variables were found to be significant more than 70% of the time the models were fitted: standard deviation of speed before the incident, mean prevailing speed, incident type (crash), time of day (morning peak hours), rainfall (moderate/heavy), shoulder width, presence of horizontal curve within the incident impact area, presence of merge influence area within the incident impact area, and presence of diverge influence area within the incident impact area.

CHAPTER 6
SUMMARY AND CONCLUSIONS

The goal of this research was to investigate approaches to mitigate secondary crashes on freeways. This goal was achieved using the following three components: (1) identify secondary crashes using a dynamic approach, (2) link the probability of secondary crashes with influential factors, and (3) develop a real-time dynamic secondary crash risk prediction model. This chapter provides a summary of this effort, research contributions, and potential future research.

6.1 Summary and Conclusions

6.1.1 Secondary Crash Identification

Accurate identification of secondary crashes is the first and the most crucial step in devising strategies to mitigate their occurrence. The primary task involved in the identification of secondary crashes focuses on defining the impact area of the primary incident. The extent of the impact area is characterized by the primary incident duration and the length of the queue initiated by the incident. This research proposed a data-driven approach to better estimate the primary incident impact area and identify secondary crashes that occurred within the impacted area. The proposed approach considered how the queue, initiated by the incident, grows and dissipates along each roadway segment upstream of the incident. This approach is able to estimate the spatial and temporal impact ranges of primary incidents while accounting for the effects of traffic flow conditions.

Traffic incidents from the SunGuide® database and high-resolution speed data from HERE Technologies were used to estimate the impact area of a primary incident. These data were collected from January 2014 to June 2019. The study area, which is located in

Florida, included a 97-mile section of Florida's Turnpike System Mainline, and the Homestead Extension of Florida Turnpike (HEFT), a 48-mile adjoining corridor. The Mainline study corridor consisted of a 69-mile Mainline Central Section (MCS) and a 28-mile Mainline South Section (MSS).

The analysis was based on 322,259 traffic incidents that occurred along the study corridors between January 2014 and June 2019. Overall, 4,549 secondary crashes in the upstream direction of the primary incident were identified from 3,977 primary incidents. The identified secondary crashes on the upstream direction of the primary incident accounted for 1.4% of the 322,259 incidents. This is an equivalent of 5.7 secondary crashes per mile per year.

Descriptive statistics of the secondary crashes indicated that 93% of the secondary crashes occurred within two hours after the occurrence of the primary incidents. Spatially, 47% of the secondary crashes occurred within two miles from the primary incident. Overall, 40% of secondary crashes occurred within two hours of the onset of a primary incident and within two miles upstream of the primary incident, the most considered spatiotemporal threshold. The following are some of the key characteristics of the primary incidents and secondary crashes:

- Only 3% of secondary crashes occurred between midnight and 5:00 AM, whereas 85% occurred during morning and evening peak hours. Specifically, 33% of secondary crashes occurred during the morning peak (i.e., 6:00 AM - 10:00 AM) while the remaining 52% occurred during the evening peak (i.e., 2:00 PM - 8:00 PM). The highest proportion of primary incidents (13%) occurred between 4:00 PM and 5:00

PM, while the highest proportion of secondary crashes (13%) occurred an hour after the primary incident, i.e., between 5:00 PM and 6:00 PM.

- The proportion of normal incidents and secondary crashes was much higher on weekdays than on weekends. Compared to other days of the week, Friday was found to experience the highest proportion of secondary crashes (20%).
- While secondary crashes were found to occur on Mondays and Fridays, normal incidents were found to occur primarily on weekdays (i.e., Monday through Friday). Only 13% of secondary crashes were found to occur on weekends.
- As expected, traffic incidents involving towing and/or EMS resulted in longer incident clearance durations, as they tend to require more time to be cleared. Approximately 94% of normal incidents were cleared within 90 minutes, while 82% of primary incidents were cleared within 90 minutes. Likewise, 94% of traffic incidents that did not involve EMS were cleared within 90 minutes, while only 64% of traffic incidents that involved EMS were cleared within 90 minutes. The longer clearance time of the primary incidents could be considered as one factor that may have contributed to the occurrence of secondary crashes.
- The severity of primary incidents was found to be one of the factors that influence the occurrence of secondary crashes. About 9% of primary incidents were moderate/severe while only 1% of normal incidents were moderate/severe. Besides the severity of primary incidents, the number of responding agencies, percentage of lanes closed, incident type, and incidents that required towing and/or EMS were also considered to be good indicators of incident severity. About 99% of normal incidents did not result in lane closure, while 21% of primary incidents resulted in a lane closure. Only 10% of

normal incidents were identified as crashes, while 47% of primary incidents were crashes. About 13% of primary incidents required towing, while only 3% of normal incidents required towing. Similarly, a higher percentage of incidents involving EMS resulted in secondary crashes (11%). While only 28% of normal incidents involved more than one responding agency, 51% of primary incidents and 55% of secondary crashes involved more than one responding agency. These statistics indicate that the severity of primary incidents influences the occurrence of secondary crashes.

- Compared to normal incidents (2%), a higher proportion of primary incidents (13%) occurred during cloudy/foggy/rainy conditions. Similarly, a higher percentage of primary incidents (11%) and secondary crashes (18%) occurred on wet surface conditions. These statistics imply that inclement weather conditions and adverse road surface conditions are among the factors that increase the probability of secondary crashes.

In practice, the proposed approach can be easily implemented considering that its algorithm does not require much computational effort except for establishing the speed profiles for normal traffic conditions. Notably, these profiles are established once and can be used for a prolonged time (up to a year). The proposed method can be used by the incident management officials while generating standard reports on a month to month, quarterly, and yearly basis. With additional programming work and the availability of access to real-time traffic and incident data, the proposed method could be utilized to accurately identify potential secondary crashes in real-time.

6.1.2 Factors Influencing the Occurrence of Secondary Crashes

This research extends the previous research on secondary crash likelihood models by proposing a method that simultaneously extracts the most important explanatory variables, with minimal correlation, influencing the risk of secondary crashes while addressing the imbalanced nature of the response variable. Specifically, the present research used the Least Absolute Shrinkage and Selection Operator (LASSO) penalized logistic regression, fitted using the bootstrap resampling approach, to identify risk factors that influence the likelihood of secondary crashes. Traffic flow, incident, temporal, weather, and roadway geometric attributes were considered as potential factors that may influence the likelihood of secondary crashes.

The influence of potential variables that were rarely utilized in previous studies, i.e., work zone, vertical curve, merge influence area, and diverge influence area, were explored. For this task, the study area included the 48-mile HEFT corridor and the 28-mile MSS corridor, both of which are a part of the Florida's Turnpike Systems in Miami, Florida.

As a first step toward achieving the research objective, potential secondary crashes were identified using high-resolution speed data and traffic incident data. The results indicated that 11.2 secondary crashes/mile/year occurred on the HEFT, while 6.5 secondary crashes/mile/year occurred on the MSS. The presence of construction activities may have contributed to the higher proportion of secondary crashes on the HEFT corridor.

Next, the LASSO penalized estimator was used to extract the most important explanatory variables, with minimal correlation, influencing the risk of secondary crashes. Because the proportion of primary incidents is smaller than the proportion of normal

incidents, the bootstrap resampling method was used to fit the penalized logistic regression. The proposed model is considered to improve the predictive accuracy of the secondary crash risk model because it accounts for the asymmetric nature of secondary crashes, performs variable selection, and removes correlated variables.

The presence of work zones was found to significantly increase the likelihood of secondary crashes. In addition, the likelihood model results indicate that roadway geometric, temporal, traffic flow, incident, and weather attributes influence the occurrence of secondary crashes. While the sign of most of these attributes is consistent with previous studies, the influence of shoulder width and day of the week on secondary crash occurrence was found to be inconsistent.

In summary, work zones were found to significantly increase the likelihood of secondary crashes, a conclusion that was derived from both the descriptive statistics and the model results. This finding warrants the inclusion of work zone presence in future secondary crash research. The results of the research will help agencies on several fronts. First, it will assist in proactively preventing secondary crashes in work zones. Second, first responders can be more vigilant and better prepared for potential secondary crashes. And finally, motorists upstream of the primary incident and the work zone could be warned about potential secondary crashes.

6.1.3 Impact of Concurrent Factors on Cascading Crash Likelihood

This research also explored the impact of concurrent factors on the probability of cascading crashes. Considering the work zone activities taking place on the HEFT during the study period, only data from the MSS were used in developing the cascading crash

model. A Bayesian network approach was used to estimate concurrent factors, i.e., related to traffic-flow, incident, temporal, weather, and roadway geometric attributes, that influence the risk of cascading crashes. Before establishing a Bayesian network, the penalized logistic regression fitted using a bootstrap resampling approach was used to select the most important variables.

About 6% of primary incidents resulted in cascading crashes. The results showed that the following attributes significantly affect the likelihood of cascading crashes: incident type, presence of merge influence area within incident impact area, incident severity, standard deviation of prevailing speed, rainfall, EMS involvement, time of day, and day of the week. The prediction inference using the optimal Bayesian network indicated the following four variables have a direct probabilistic relationship with the likelihood of cascading crashes: standard deviation of prevailing speed, incident severity, rainfall, and day of the week. Cascading crashes were found to most likely occur when the prior incident occurs during moderate/heavy rainfall condition, weekday, low variation in prevailing speed, and the incident resulted in a moderate/severe impact on traffic. Cascading crashes were also found to be more likely to occur when prior incident occurred on weekend, high variation in prevailing speed, moderate/heavy rainfall, and the incident resulted in a moderate/severe impact on traffic.

It is important to note that the Bayesian network model was utilized to give a superior comprehension of the perplexing reliance that exists among influential factors and cascading crash probability. Also, as shown in this exploration, the Bayesian network model can be utilized to assess factors that have a strong influence on cascading crash probability, and accordingly, improve the determination of fitting countermeasures.

Additionally, it is conceivable to utilize the Bayesian network method to anticipate the probability of cascading crashes after the countermeasures have been applied. This type of analysis is also referred to as intervention analysis in the Bayesian network s model.

6.1.4 Dynamic Prediction of Secondary Crashes in Real-time

The risk of secondary crashes is not static but varies with time, a situation contributed to by the changes in prevailing traffic conditions after an incident occurs. The dynamic Bayesian cloglog model was used to predict the risk of secondary crashes every five minutes following the occurrence of the incident. The coefficients of the following 10 variables were found to be significant more than 70% of the time the models were fitted: mean speed before the incident, standard deviation of speed before the incident, mean prevailing speed, incident type (crash), time of day (morning peak hours), rainfall, shoulder width, presence of horizontal curve within the incident impact area, presence of merge influence area within the incident impact area, and presence of diverge influence area within the incident impact area. The following are some of the key findings on the influence of these factors on the likelihood of secondary crashes:

- The mean speed before the incident was found to increase the risk of secondary crashes within 25 minutes from when the incidents started impacting traffic. Afterward, the magnitude of the coefficients became negative and increased sharply up to the 150^{th} minutes (model 30), where the slope of the magnitude of the coefficients became flatter.
- The standard deviation of the speed before the incident was found to increase the risk of secondary crashes. The magnitude of the impact of variation of speed before the

incident on secondary crash likelihood was found to increase with time until the 100th minute where it started to decrease.

- The coefficients of the mean prevailing speed were found to be negative indicating that the risk of secondary crashes decreases as the mean prevailing speed increases. The magnitude of these coefficients was found to decrease sharply with time until the 125th (model 25) minute, where the slope became flatter.
- In all 50 fitted models, crashes were found to be more likely to cause secondary crashes compared to hazards and vehicle problems.
- The sign of the coefficients of the time of day was found to be positive, indicating that incidents that occur during morning peak hours are more likely to cause secondary crashes. Overall, the magnitude of these coefficients was observed to increase with time.
- Moderate or heavy rainfall was associated with an increased likelihood of secondary crashes in all of the fitted models.
- Overall, the impact of shoulder width on secondary crash risk was observed to increase with time until the 200th minute, where it started to decrease.
- A higher likelihood of secondary crashes was observed when a horizontal curve (rather than the tangent) was within the incident impact area.
- The signs of the coefficients of the merge and diverge influence areas were found to be positive, and their magnitude was found to increase with time. This implies that the influence of these variables on secondary crash occurrence decreases with time.

As can be inferred from the research findings, the occurrence of secondary crashes is influenced by incident severity and how quickly the incident is cleared. Furthermore, the

likelihood of secondary crashes is closely related to the changes in the traffic flow states. That is, secondary crashes are more likely to occur when the traffic is transitioning from a free-flow state to a congested state. Once the traffic is congested and there is no more significant variation in traffic, the risk of secondary crashes also decreases.

To prevent the risk of secondary crash occurrence, traffic management strategies should be developed to accelerate the dissipation of the queue upstream of the potential primary incident. Warnings could be sent to drivers approaching a primary crash scene in real-time through various means, including dynamic message signs (DMSs), Advanced Traveler Information Systems (ATIS), such as the Waze application, and emerging technologies, such as connected vehicles, allowing them to take necessary precautions, such as detour or drive with caution, to avoid being involved in a secondary crash. Furthermore, when the conditions associated with a high likelihood of secondary crashes prevail, responding agencies, such as highway patrols, emergency medical services, towing agencies, etc., could be better prepared to respond to secondary crashes if they were to occur. These strategies will help to potentially reduce the frequency and severity of secondary crashes.

6.2 Research Contributions

Incident management agencies have been investing a substantial amount of resources to devise strategies to mitigate secondary crashes. Agencies have been struggling since identifying secondary crashes is not a straightforward process. The definition itself is subjective, and identifying secondary crashes depends on how the impact area of the primary incident is defined. The queue caused by the incident forms and dissipates at

different rates along each of the upstream segments impacted by the incident. As such, the approach employed to estimate the incident impact area has to consider this principle. Failure to properly estimate the incident impact area may lead to under- or overestimation of the impact area, and hence, the number of secondary crashes caused by the respective incident.

This research discussed the shortcomings of the existing approaches used to identify secondary crashes and proposed an improved data-driven approach. To improve the accuracy of the identified secondary crashes, the proposed method considered the fact that the queue built by the primary incident grows and dissipates at a different rate upstream of the incident.

For the first time, this research extended the previous efforts on secondary crash likelihood models by evaluating the impact of work zones on the occurrence of secondary crashes. Other potential variables that were rarely considered in previous studies, i.e., vertical curve, merge influence area, and diverge influence area, were also explored. Also for the first time, high-resolution and location-specific rainfall data were included as influential variables in modeling the risk of secondary crashes.

In addition, for the first time, this research used a Bayesian network to provide a better understanding of the complex dependence that exists among relationships between explanatory variables and cascading crash likelihood. This research also presented a binary classification approach that dynamically predicts the likelihood of secondary crashes every five minutes from when the initial incident started impacting traffic.

6.3 Future Work

Accurate estimation of the primary incident impact area depends on the availability and reliability of relevant data for traffic state estimation. High-resolution speed data extracted from the HERE Technologies was used to estimate the spatiotemporal impact area of primary incidents. However, these data are not available along all corridors. Furthermore, the use of these data is also limited by the length of the Traffic Message Channels, which are segments used by HERE to record vehicle speeds. The use of data from overly long Traffic Message Channels may not be able to precisely capture the speed changes over space.

As probe vehicle traffic data from sources, such as HERE, Bluetooth devices, Wi-Fi sensors, etc., become more prevalent, and as crowdsourced travel speed data become more readily available, future studies could incorporate virtual detectors that use data from multiple sources to obtain more disaggregated traffic data. Moreover, with the use of crowdsourced traffic data, the study locations do not have to be limited to the corridors with Traffic Message Channels. Future research could also explore the influence of Traffic Message Channel length on the accuracy of the estimated incident impact areas.

The dynamic secondary crash risk prediction model incorporates only the incident type as the most important incident-related variable. Other most important incident-related variables (lane closure and number of responding agencies) were excluded since it was not clear at what time these variables were reported after the incident occurred. An attempt could be made in the future to record the timeline of these variables, and hence, include them in the dynamic model.

REFERENCES

Algamal, Z.Y., & Lee, M.H. (2015a). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine, 67*, 136–145. doi:10.1016/j.compbiomed.2015.10.008

Algamal, Z.Y., & Lee, M.H. (2015b). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications, 42(23),* 9326–9332. doi:10.1016/j.eswa.2015.08.016

American Meteorological Society [AMS]. (n.d.). *Glossary of meteorology*. Retrieved June 17, 2019 from https://glossary.ametsoc.org/wiki/Rain

Andrew, L. (2019). *Investigating the effects of rainfall on traffic operations on Florida freeway*s. University of North Florida.

Balke, K. (2009). *Traffic incident management in construction and maintenance work zones* (FHWA-HOP-08-056). Washington, D.C.: Federal Highway Administration.

Barr, J. (2018). *New AWS public data set – real-time and archived NEXRAD weather data | Amazon Web Services. Amazon, Amazon*. Retrieved May 15, 2019 from aws.amazon.com/blogs/aws/new-aws-public-data-set-real-time-and-archived-nexrad-weather-data/.

Baykal-Gürsy, M., Xiao, W., & Ozbay, K. (2009). Modeling traffic flow interrupted by incidents. *European Journal of Operational Research*, *195*(1), 127-138.

Chang, G. L., & Rochon, S. (2011). *Performance evaluation and benefit analysis for CHART*. Hanover, Maryland: Maryland Department of Transportation.

Chatterjee, K., Hounsell, N. B., Firmin, P. E., & Bonsall, P. W. (2002). "Driver response to variable message sign information in London." *Transportation Research Part C: Emerging Technologies 10(2),* 149–69.

Chimba, D., & Kutela, B. (2014). Scanning secondary derived crashes from disabled and abandoned vehicle incidents on uninterrupted flow highways. *Journal of Safety Research*, 50, 109–116. https://doi.org/10.1016/j.jsr.2014.05.004

Chung, Y. (2013). Identifying primary and secondary crashes from spatiotemporal crash impact analysis. *Transportation Research Record: Journal of the Transportation Research Board, 2386,* 62–71.

Cong, H., Chen, C., Lin, P.-S., Zhang, G., Milton, J., & Zhi, Y. (2018). Traffic incident duration estimation based on a dual-learning Bayesian network model. *Transportation Research Record: Journal of the Transportation Research Board* 2672(45), 196–209.

Dougald, L.E., Goodall, N.J., & Venkatanarayana, R. (2016). *Traffic incident management quick clearance guidance and implications (FHWA/VTRC 16-R9)*. Virginia: Transportation Research Council.

Federal Highway Administration [FHWA]. (2007). *Utility work zone safety guidelines and training gap study and needs assessment*. Washington, D.C: Federal Highway Administration.

Florida Department of Transportation [FDOT]. (2016). *Roadway characteristic inventory (RCI): features and characteristics handbook*. Tallahassee, Florida: Florida Department of Transportation.

Florida Department of Transportation [FDOT]. (2017). *TSM&O 2017 Strategic Plan*. Tallahassee, Florida: Florida Department of Transportation.

Glotzbach, G. (2014). *The Waze connection*. Tallahassee, Florida.

Goodall, N. J. (2017). Probability of secondary crash occurrence on freeways with the use of private-sector speed data. *Transportation Research Record: Journal of the Transportation Research Board*, *2635*, 11–18.

Green, E. R., Pigman, J. G., Walton, J. R., & McCormack, S. (2012). Identification of secondary crashes and recommended countermeasures to ensure more accurate documentation. *Proceedings of the 91th Annual Meeting of the Transportation Research Board*, January 22-26, 2012. Transportation Research Board, Washington, D.C.

Harding, J., Powell, G., Yoon, R., Fikentscher, J., Doyle, C., Sade, D., Lukuc, M., Simons, J., & Wang, J. (2014). *Vehicle-to-Vehicle Communications : Readiness of V2V Technology for Application (DOT HS 812 014)*. Washington, D.C.: National Highway Traffic Safety Administration.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*, New York: Springer. doi:10.1198/jasa.2004.s339

Haule, H. J., Alluri, P., Sando, T., & Raihan, M. A. (2020). Investigating the impact of rain on crash-clearance duration. *Journal of Transportation Engineering Part A System, 146(11)*, 04020130.

Hirunyanitiwattana, W., & Mattingly, S. P. (2006). Identifying secondary crash characteristics for California highway system. *Proceedings of the 85th Annual Meeting*

of the Transportation Research Board, January 22-26, 2006. Transportation Research Board, Washington, D.C.

Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15(2008),* 1593–1623.

Imani, H. N. (2019). *The use of real-time connected vehicles and HERE data in developing an automated freeway incident detection algorithm*. University of North Florida.

Imprialou, M. I. M., Orfanou, F. P., Vlahogianni, E. I., & Karlaftis, M. G. (2014). Methods for defining spatiotemporal influence areas and secondary incident detection in freeways. *Journal of Transportation Engineering*, *140*(1), 70-80.

INRIX. (2019). *INRIX: Congestion costs each American 97 hours, $1,348 A Year*. Retrieved July 11, 2019 from http://inrix.com/press-releases/scorecard-2018-us/

Jalayer, M., Baratian-Ghorghi, F., & Zhou, H. (2015). Identifying and characterizing secondary crashes on the alabama state highway systems. *Advances in Transportation Studies*, 37, 129–140.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: springer.

Karlaftis, M. G., Latoski, S. P., Richards, N. J., & Sinha, K. C. (1999). ITS impacts on safety and traffic management: an investigation of secondary crash causes. *Journal of Intelligent Transportation Systems*, *5*(1), 39-52.

Kassambara, A. (2017). *Machine Learning Essentials: Practical Guide in R*, First Edit. ed. STHDA.

Khattak, A., Wang, X., & Zhang, H. (2009). Are incident durations and secondary incidents interdependent? *Transportation Research Record: Journal of the Transportation Research Board*, *2099*, 39-49.

Khattak, A., Wang, X., Zhang, H., X, W., & Zhang, H. (2012). Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays. *IET Intelligent Transport Systems, 6(2),* 204–214.

Khattak, A.J., & Wali, B. (2017). Analysis of volatility in driving regimes extracted from basic safety messages transmitted between connected vehicles. *Transportation Research Part C: Emerging Technologies*, *84*, 48-73.

Kidando, E., Kitali, A.E., Lyimo, S.M., Sando, T., Moses, R., Kwigizile, V., & Chimba, D. (2019a). Applying probabilistic model to quantify influence of rainy weather on

stochastic and dynamic transition of traffic conditions. *Journal of Transportation Engineering Part A System, 145(5)*, 04019017.

Kidando, E., Moses, R., Sando, T., & Ozguven, E. E. (2019b). Assessment of factors associated with travel time reliability and prediction: an empirical analysis using probabilistic reasoning approach. *Transportation Planning and Technology, 42(4),* 309–323.

Kitali, A. E., Alluri, P., Sando, T., Haule, H., Kidando, E., & Lentz, R. (2018). Likelihood estimation of secondary crashes using bayesian complementary log-log model. *Accident Analysis and Prevention* 119: 58–67. doi:10.1016/J.AAP.2018.07.003

Kitali, A. E., Kidando, E., Sando, T., Moses, R., & Ozguven, E. E. (2017). Evaluating aging pedestrian crash severity with bayesian complementary log–log model for improved prediction accuracy. *Transportation Research Record: Journal of the Transportation Research Board*, *2659*, 155-163.

Kitali, A. E., Alluri, P., Sando, T., & Lentz, R. (2019a). Impact of primary incident spatiotemporal influence thresholds on the detection of secondary crashes. *Transportation Research Record: Journal of the Transportation Research Board,* 2673(10), 271–283.

Kitali, A. E., Alluri, P., Sando, T., & Wu, W. (2019b). Identification of secondary crash risk factors using penalized logistic regression model. *Transportation Research Record: Journal of the Transportation Research Board,* 2673(11): 901-914.

Kopitch, L., & Saphores, J. D. M. (2011). Assessing effectiveness of changeable message signs on secondary crashes. *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, January 23-27, 2011. Transportation Research Board, Washington, D.C.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General,* 142(2), 573-603.

Kuhn, M. (2019). *The caret Package*.

Kutela, B., & Teng, H. (2019). Prediction of drivers and pedestrians' behaviors at signalized mid-block Danish offset crosswalks using Bayesian networks. *Journal of Safety Research, 69,* 75–83. https://doi.org/10.1016/J.JSR.2019.02.008

Latoski, S. P., Pal, R., & Sinha, K. C. (1999). Cost-effectiveness evaluation of Hoosier Helper freeway service patrol. *Journal of Transportation Engineering*, *125*(5), 429-438.

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

Lou, Y., Yin, Y., & Lawphongpanich, S. (2011). Freeway service patrol deployment planning for incident management and congestion mitigation. *Transportation Research Part C: Emerging Technologies*, *19*(2), 283–295.

McCartt, A.T., Northrup, V.S., & Retting, R.A. (2004). Types and characteristics of ramp-related motor vehicle crashes on urban interstate roadways in Northern Virginia. *Journal of Safety Research, 35(1),* 107–114.

McCormick, T.H., Raftery, A. E., Madigan, D., & Burd, R. S. (2012). Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics 68,* 23–30.

Menard, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*, 92–122.

Mishra, S., Golias, M., Sarker, A., & Naimi, A. (2016). *Effect of primary and secondary crashes: identification, visualization, and prediction research report* (CFIRE 09-05). Madison, Wisconsin: Wisconsin Department of Transportation.

Montes, Ca., Faquir, T., Hapney, TJ., & Birriel, E. (2008). *Guidelines for the use of dynamic message signs on the Florida state highway system*. Tallahassee, Florida: Florida Department of Transportation.

Moore, J. E., Giuliano, G., & Cho, S. (2004). Secondary accident rates on Los Angeles freeways. *Journal of Transportation Engineering*, *130*(3), 280-285.

Mounce, J. M., Ullman, G., Pesti, G., & Pezoldt, V. (2007). *Guidelines for the evaluation of dynamic message sign performance (FHWA/TX-07/0-4772-1).* Austin, TX: Texas Department of Transportation.

National Cooperative Highway Research Program [NCHRP]. (2014). Guidance for implementation of traffic incident management performance measurement/ performance measurement for traffic incident management programs/ Florida (webpage).
Retrieved February 12, 2018, from http://nchrptimpm.timnetwork.org/?page_id=79

National Oceanic and Atmospheric Administration [NOAA]. (n.d.). *National Doppler radar sites*. Retrieved April 29, 2019 from https://www.ncdc.noaa.gov/nexradinv/map.jsp

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. John Wiley and Sons, Inc., Hoboken, New Jersey.

O'laughlin, J., & Smith, A. (2002). Operational issues discussion paper on "incident management operations: Top five issues. *Proceedings of the National Conference on TrafficIncident Management: A Road Map to the Future*, March 11-13, 2002. American Association of State Highway and Transportation Officials, Washington, D.C.

Ou, J., Xia, J., Wang, Y., Wang, C., & Lu, Z. (2020). A data-driven approach to determining freeway incident impact areas with fuzzy and graph theory-based clustering. *Computer-Aided Civil and Infrastructure Engineering*, 35, 178–199.

Owens, N., Armstrong, A., Sullivan, P., Mitchell, C., Newton, D., Brewster, R., & Trego, T. (2010). *Traffic incident management handbook* (FHWA-HOP-10-013). Washington, D.C.: Federal Highway Administration, Office of Transportation Operations.

Park, H., & Haghani, A. (2016b). Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C: Emerging Technologies, 70*, 69-85.

Park, H., Gao, S., Haghani, A., Samuel, S., & Knodler, M. A. (2017). Sequential intepretation and prediction of secondary incident probability in real time. *Proceedings of the 96th Annual Meeting of the Transportation Research Board*, January 8-12, 2017. Transportation Research Board, Washington, D.C.

Park, H., & Haghani, A. (2016a). Use of clustering model and adjusted boxplot model for identification of secondary incidents. *Proceedings of the 95th Annual Meeting of the Transportation Research Board*, January 10-14, 2016. Transportation Research Board, Washington, D.C.

Park, H., Haghani, A., & Samuel, S. (2018). Real-time prediction and avoidance of secondary crashes under unexpected traffic congestion. *Accident Analysis and Prevention*, *112*, 39–49.

Pei, X., Sze, N.N., Wong, S.C., & Yao, D. (2016). Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong. *Accident Analysis and Prevention 95*, 512–520. doi:10.1016/j.aap.2015.06.007

Raub, R. (1997). Occurrence of secondary crashes on urban arterial roadways. *Transportation Research Record: Journal of the Transportation Research Board, 1581*, 53-58.

Sando, T., Alluri, P., Chuan, C., Haule, H., Kitali, A., Lentz, R., & Huq, A. (2018).

Evaluation of incident response improvements for statewide application: learning from the new regional traffic management center in Jacksonville, Florida. Tallahassee, Florida: Florida Department of Transportation.

Sarker, A. A., Paleti, R., Mishra, S., Golias, M. M., & Freeze, P. B. (2017). Prediction of secondary crash frequency on highway networks. *Accident Analysis and Prevention*, *98,* 108-117.

SAS Institute Inc. (2019). *SAS Visual Statistics 8*.5 Procedures. Cary, North Carolina.

Stylianou, K., & Dimitriou, L. (2018). Analysis of rear-end conflicts in urban networks using Bayesian networks. Transportation Research Record: Journal of the Transportation Research Board, 2672(38), 302–312. doi:10.1177/0361198118790843

Sun, C. C., & Chilukuri, V. (2010). Dynamic incident progression curve for classifying secondary traffic crashes. *Journal of Transportation Engineering*, *136*(12), 1153-1158.

Sun, C., & Chilukuri, V. (2006). The use of dynamic incident progression curve for classifying secondary accidents. *Proceedings of the 85th Annual Meeting of the Transportation Research Board*, January 22-26, 2006. Transportation Research Board, Washington, D.C.

Tian, Y., Chen, H., & Truong, D. (2016). A case study to identify secondary crashes on Interstate Highways in Florida by using Geographic Information Systems (GIS). *Advances in Transportation Studies, 2,* 103-112.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Transportation Research Board [TRB]. (2016). *Highway capacity manual 6th edition: a guide for multimodal mobility analysis*. Washington D.C.

Vlahogianni, E. I., Karlaftis, M. G., & Orfanou, F. P. (2012). Modeling the effects of weather and traffic on the risk of secondary incidents. *Journal of Intelligent Transportation Systems*, *16*(3), 109-117.

Vlahogianni, E. I., Karlaftis, M. G., Golias, J. C., & Halkias, B. M. (2010). Freeway operations, spatiotemporal-incident characteristics, and secondary-crash occurrence. *Transportation Research Record: Journal of the Transportation Research Board*, *2178,* 1-9.

Wang, J., Boya, L., Lanfang, Z., & Ragland, D. R. (2016). Modeling secondary accidents identified by traffic shock waves. *Accident Analysis and Prevention*, *87*, 141-147.

Wang, J., Liu, B., Fu, T., Liu, S., & Stipancic, J. (2019). Modeling when and where a secondary accident occurs". *Accident Analysis and Prevention, 130,* 160-166.

Wuillemin, P.-H. (2019). *pyAgrum documentation: release 0.15.2*.

Xie, C., &Waller, S. (201)0. Estimation and application of a Bayesian network model for discrete travel choice analysis. *Transportation Letters, 2(2),* 125–144. doi:10.3328/TL.2010.02.02.125-144

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

Xu, C., Xu, S., Wang, C., & Li, J. (2019). Investigating the factors affecting secondary crash frequency caused by one primary crash using zero-inflated ordered probit regression. *Physica A: Statistical Mechanics and Its Applications*, 524, 121–129.

Yang, H., Bartin, B., & Ozbay, K. (2014a). Mining the characteristics of secondary crashes on highways. *Journal of Transportation Engineering*, *140*(4), 04013024.

Yang, H., Ozbay, K., & Xie, K. (2014b). Assessing the risk of secondary crashes on highways. *Journal of Safety Research*, *49*, 143.e1-149.

Yang, H., Ozbay, K., Morgul, E., Bartin, B., & Xie, K. (2014c). Development of online scalable approach for identifying secondary crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2470, 24-33.

Yang, H., Wang, Z., & Xie, K. (2017). Impact of connected vehicles on mitigating secondary crash risk. *International Journal of Transportation Science and Technology*, *6*(3), 196–207.

Yang, H., Wang, Z., Xie, K., Ozbay, K., & Imprialou, M. (2018). Methodological evolution and frontiers of identifying, modeling and preventing secondary crashes on highways. *Accident Analysis and Prevention*, 117, 40–54. doi:10.1016/J.AAP.2018.04.001

Yang, K., Wang, X., & Yu, R. (2018). A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. *Transportation Research Part C: Emerging Technologies, 96,* 192–207. doi:10.1016/J.TRC.2018.09.020

Zhan, C., Gan, A., & Hadi, M. (2009). Identifying secondary crashes and their contributing factors. *Transportation Research Record: Journal of the Transportation Research Board, 2102,* 68-75.

Zhan, C., Shen, L., Hadi, M. A., & Gan, A. (2008). Understanding the characteristics of secondary crashes on freeways. *Proceedings of the 87th Annual Meeting of the*

Transportation Research Board, January 13-17, 2008. Transportation Research Board, Washington, D.C.

Zhang, H., & Khattak, A. (2010). What is the role of multiple secondary incidents in traffic operations?. *Journal of Transportation Engineering, 136*(11), 986-997.

Zhao, X., Xu, W., Ma, J., Li, H., Chen, Y., and Rong, J. (2019). Effects of connected vehicle-based variable speed limit under different foggy conditions based on simulated driving. Accident Analysis and Prevention, 128, 206–216.

Zheng, D., Chitturi, M. V., Bill, A.R., & Noyce, D. A. (2014). Secondary crash identification on a large-scale highway system. *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*, January 12-16, 2014. Transportation Research Board, Washington, D.C.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101(476)*, 1418–1429. doi:10.1198/016214506000000735

VITA

ANGELA EDES KITALI

EDUCATION

2011 – 2015 B.S., Civil Engineering
University of Dar es Salaam, Dar es Salaam, Tanzania

2015 – 2017 M.S., Civil Engineering (Transportation)
University of North Florida, Jacksonville, Florida

2017 – 2020 Graduate Research Assistant
Department of Civil and Environmental Engineering
Florida International University, Miami, Florida

2018 – 2020 Doctoral Candidate
Department of Civil and Environmental Engineering
Florida International University, Miami, Florida

PUBLICATIONS

1. Kitali, A. E., Mokhtarimousavi, S., Kadeha, C., and Alluri, P. (2020). "Severity analysis of crashes on express lane facilities using support vector machine model trained by firefly algorithm." *Traffic Injury Prevention*, 1–6.

2. Kidando, E., Karaer, A., Kutela, B., Kitali, A., Moses, R., Ozguven, E., and Sando, T. (2020). "A Novel Approach for Calibrating Freeway Highway Multi-Regimes Fundamental Diagram," *Transportation Research Record: Journal of the Transportation Research Board*, 0361198120930221.

3. Salum, J. H., Sando, T., Alluri, P., and Kitali, A. (2020). "Operational Evaluation of Freeway Service Patrols: A Case Study of Florida's Road Rangers," *Journal of Transportation Engineering, Part A*, 146 (9), 04020094.

4. Kidando, E., Kitali, A., Moses, R., and Ozguven, E. (2020). "Real-Time Visualization of Operational Performance Measures of Arterial Highway Using Open Crowdsourced Data," *Advances in Transportation Studies*, 51, 47–62.

5. Kitali, A., Kidando, E., Alluri, P., Sando, T., and Salum, J. (2020). "Modeling Severity of Motorcycle Crashes with Dirichlet Process Priors," *Journal of Transportation Safety and Security 10.1080*/19439962.2020.1738613.

6. Kitali, A., Alluri, P., Sando, T., and Wu, W. (2019). "Identification of Secondary Crash Risk Factors using Penalized Logistic Regression Model," *Transportation Research Record: Journal of the Transportation Research Board*, 2673(11), 901–914.

7. Kitali, A., Alluri, P., and Sando, T. (2019). "Impact of Primary Incident Spatiotemporal Influence Thresholds on the Detection of Secondary Crashes," *Transportation Research Record: Journal of the Transportation Research Board*, 2673(10), 271–283.

8. Salum, J., Kitali, A., Bwire, H., Sando, T., and Alluri, P. (2019) "Severity of Motorcycle Crashes in Dar es Salaam, Tanzania," *Traffic Injury Prevention*, 20(2), 189–195.

9. Kidando, E., Kitali, A., Moses, R., Lyimo, S., Kwigizile, V., Sando, T., and Chimba, D. (2019). "Applying a Probabilistic Model to Quantify the Influence of Rainy Weather on a Stochastic and Dynamic Transition of Traffic Conditions," *Journal of Transportation Engineering, Part A*, 145 (5): 04019017.

10. Haule, H., Sando, T., Kitali, A., and Richardson, R. (2018). "Investigating Proximity of Crash Locations to Aging Pedestrian Residences," *Accident Analysis and Prevention.* 122, 215–225.

11. Kitali, A., Alluri, P., Sando, T., Haule, H., Kidando, E., and Lentz, R. (2018). "Likelihood Estimation of Secondary Crashes Using Bayesian Complementary Log-Log Model," *Accident Analysis and Prevention*, 119, 58–67.

12. Kitali, A., Kidando, E., Martz, P., Alluri, P., Sando, T., Moses, R., and Lentz, R. (2018). "Evaluating Factors Influencing the Severity of Three-plus Multiple-vehicle Crashes Using Real-Time Traffic Data," *Transportation Research Record: Journal of the Transportation Research Board*, 2672(38), 128–137.

13. Kitali, A., Sando, T, Castro, A., Kobelo, D., and Mwakalonge, J. (2017). "Appraisal of Safety Effects of Pedestrian Countdown Signals to Drivers using Crash Modification Factors," *Journal of Transportation Engineering, Part A*, 144 (5): 04018011.

14. Kitali, A., Kidando, E., Sando, T., Moses, R., and Ozguven, E. (2017). "Evaluating Aging Pedestrian Crash Severity with Bayesian Complementary Log–Log Model for Improved Prediction Accuracy," *Transportation Research Record: Journal of the Transportation Research Board*, 2659, 155-163.

15. Kitali, A., and Sando, T. (2017). "A Full Bayesian Approach to Appraise the Safety Effects of Pedestrian Countdown Signals to Drivers," *Accident Analysis and Prevention*, 106, 327-335.

16. Kitali, A., Sando, T., Moses, R., and Ozguven, E. (2017). "Understanding Factors Associated with Severity of Aging Population-Involved Pedestrian Crashes in Florida," *Advances in Transportation Studies*, 42 (3), 85-98.