11-20-2023

Uncovering Temporal Dynamics in Traffic Safety: Application of Duration-Based Models for Diagnostic and Predictive Analysis

Diwas Thapa

UNCOVERING TEMPORAL DYNAMICS IN TRAFFIC SAFETY: APPLICATION OF DURATION-BASED MODELS FOR DIAGNOSTIC AND PREDICTIVE ANALYSIS

by

Diwas Thapa

A Dissertation

Submitted in Partial Fulfilment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Civil Engineering

The University of Memphis

December 2023

To my mother Kabita, my father Kedar, and my wife Kanchan who have been my pillars of strength and source of inspiration.

Acknowledgement

The past five years of my Ph.D. journey have undoubtedly been the most challenging, yet memorable days. Amidst countless obstacles, sleepless nights, and looming deadlines, I had the privilege of being surrounded by exceptionally intellectual and supportive individuals. Their assistance not only helped me endure through the struggle but excel. I am grateful for the opportunity to recognize their pivotal roles and contributions to the completion of this dissertation.

I begin by expressing my deepest gratitude to my advising professors for their unwavering support, guidance, insightful feedback, and above all, their perseverance as I transitioned from a graduate student to a researcher. First and foremost, my sincere thanks extend to my advisor, Dr. Sabyasachee Mishra. His extensive knowledge, expertise, and scholarly perspective have played a decisive role in shaping my thought process and guiding the direction of this dissertation. His constant motivation, dedication, and strong work ethic inspired and fueled my aspiration for excellence—a drive I aim to carry forward and apply in both my career and personal life. Secondly, I extend my gratitude to Dr. Mihalis Golias, whose encouraging, confident, and relaxed demeanor has greatly influenced my professional character and growth. I will forever appreciate their efforts and guidance.

I am also thankful to the members of my dissertation committee, Dr. Mohamed Osman and Dr. Claudio Ivan Meier, for graciously dedicating their time amid their busy schedules to serve as referees for my dissertation. Their unique perspectives, along with challenging and constructive critiques, greatly enhanced the quality of this work.

My appreciation further extends to my coauthors and research collaborators for their cooperation, understanding and unwavering support. I am particularly indebted to Dr. Mohamed

Osman and Dr. Rajesh Paleti for their unfailing support and willingness to aid me in every aspect of my research. The guidance from Dr. Mohamed Osman during the initial years of my Ph.D. and the support from Dr. Rajesh Paleti in the later stages played a crucial role in the successful completion of this dissertation. I would also like to express gratitude to the Tennessee Department of Transportation and its exceptional staff, whose assistance was indispensable in completing research projects and supporting my research through funding and access to data through Enhanced Tennessee Roadway Information Systems (ETRIMS) that might have otherwise been unattainable.

I also owe a debt of gratitude to my colleagues and peers, especially Ishant Sharma, Huan Ngo, Ali Riahi Samani, Avani Aravind, Suvin Padinjare Venthuruthiyil, Dimitrios Giampouranis, and Vasileios Liatsos. Their companionship, along with engaging discussions and invaluable suggestions, both in professional and personal spheres, made this journey endearing despite of its hardships and challenges. The cherished conversations and moments shared with them will forever hold a special place in my memories.

I extend my heartfelt appreciation to the University staff, the teaching faculty, and the administrative staff. Their assistance in day-to-day administrative tasks significantly eased this journey, and I acknowledge their contributions to its smooth progression.

Lastly, and most importantly, I consider myself incredibly fortunate to have a loving and supportive family. Their immeasurable contributions have played an instrumental role in shaping the person I am today. Particularly, my grandparents, parents—Kedar and Kabita—and my wife, Kanchan, have been constant sources of inspiration throughout life's hardships. To them, I dedicate my achievements with profound gratitude.

Preface

Duration-based models offer a valuable tool for improving traffic safety by uncovering the temporal aspect of traffic incidents and driver behavior. These models examine the duration between events such as accidents or instances of speeding, which are closely related to traffic safety, and establish a connection between their temporal distribution and their effects on traffic safety. This dissertation aims to advance the utility of duration models by introducing innovative approaches to their application for diagnostic and predictive analyses, thereby contributing to more effective strategies for enhancing traffic safety.

This dissertation comprises six chapters. Chapter 1 provides an in-depth introduction to the research problem and the contributions made by this document. These contributions are substantiated by four journal articles presented in the next four chapters, which are either published or under review for publication in Accident Analysis & Prevention. These articles (Chapters 2, 3, 4 and 5) are listed below. The final chapter concludes the dissertation with closing arguments and future research directions.

- Chapter 2: **Thapa, D.**, & Mishra, S. (2021). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention (156)*. https://doi.org/10.1016/j.aap.2021.106125.
- Chapter 3: **Thapa, D.**, Mishra, S., & Paleti, R. (2022). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention (169)*. https://doi.org/10.1016/j.aap.2022.106639.
- Chapter 4: **Thapa, D.**, Mishra, S., Velaga, N. R., Patil, G. R. (2023). Advancing Proactive

Crash Prediction: A Discretized Duration Approach for Predicting Crashes and Severity. *Accident Analysis & Prevention (Under review).*

- Chapter 5: **Thapa, D.**, Mishra, S., Khattak, A., Adeel, M. (2023). Assessing Driver Behavior in Work Zones: A Discretized Duration Approach to Predict Speeding. *Accident Analysis & Prevention (Under review).*

Abstract

This research consists of four studies aimed at enhancing traffic safety. These studies employ innovative approaches and model frameworks based on duration models. The first study focuses on Work Zone Intrusion Alert Systems (WZIAS) and explores their component layout and impact on work zone safety. By analyzing the duration between work zone intrusion and crashes, this study identifies critical factors contributing to work zone crashes when WZIAS are in use. The study concludes by offering recommendations to optimize WZIAS effectiveness, including potential adjustments to existing work zone guidelines.

The second study addresses the limitations of duration models in handling time-varying factors and their inability to support real-time predictive modeling. It introduces a novel duration-based model framework capable of accommodating dynamic covariates for proactive crash prediction modeling. This approach involves discretizing the time between crashes into forecasting epochs. The study explores various sampling methods to address computational challenges arising from increased data size due to forecasting epochs. A 25% epoch-level sample is found effective for computational efficiency without sacrificing accuracy.

The third study extends the duration-based crash prediction framework, to develop a novel approach for predicting crash and severity simultaneously. The study further seeks to strike a balance between model performance and estimation time. Findings suggest that a 15% sample drawn at the epoch level provides an efficient compromise. Stability analysis of predictor variables sheds light on their reliability across different samples. Variables such as *Time of day (Early afternoon), Weather condition, Lighting condition (Daytime), Illumination,* and *Volume* require larger samples for more accurate coefficient estimation. Conversely, *Time of day (Early morning, Late morning, Late afternoon), Lighting condition (Dark lighted), Terrain, Land use,*

Number of lanes, and *Speed* converge towards true estimates with small incremental increases in sample size.

The fourth study employs the duration-based proactive crash prediction framework to showcase its versatility and identify the predictors of speeding events in work zones. The model's validity is tested, with higher accuracy observed for speeding events occurring with shorter durations between consecutive occurrences. The study demonstrates how the framework can help transportation agencies identify high-risk highway segments and implement safety measures proactively.

Table of contents

List of Tables

List of Figures

1. Introduction

Traffic safety is a significant concern as it remains one of the major causes of deaths worldwide. According to the World Health Organization, approximately 1.35 million people are killed on roadways every year, and it is the leading cause of death for those between 5 and 29 years old (World Health Organization, 2018). According to the Centers for Disease Control and Prevention, the U.S. has one of the highest rates of traffic fatalities among high-income nations (Yellman & Sauber-Schatz, 2022). Despite recent advancements in vehicle safety technology and traffic safety, the U.S. witnessed a 10.5% increase in traffic fatalities in 2021, with 42,915 fatalities compared to 38,824 in 2020 (National Highway Traffic Safety Administration, 2022). Furthermore, the economic cost of traffic crashes, including medical expenses, property damage, environmental costs, and other user costs such as travel delay and lost productivity, is staggering. For instance, the cost of traffic crashes in the U.S. amounts to several billion dollars every year. In 2010, Vermont reported the lowest cost at $538 million and California registering the highest at a staggering $19.98 billion (US Department of Transportation, 2020). This underscores the urgent need for continuous efforts to improve traffic safety.

Understanding the relationship between traffic crashes and Vehicle Miles Traveled (VMT) is also important as it is intricate and not always straightforward. Firstly, as VMT increases, there is also an increase in the exposure of road users, including motorists, passengers, pedestrians, and cyclists, to potential crash scenarios. In simpler terms, more VMT generally translates to more opportunities for traffic incidents to occur. Secondly, higher VMT often corresponds to greater traffic volume, leading to congestion and stop-and-go conditions, especially during peak hours. These conditions are when crashes are more likely to happen. However, it's important to note that higher VMT, as seen during the COVID-19 pandemic, also

resulted in an increase in more severe crashes. This was due to the fact that average travel speeds increased in the absence of heavy traffic (Hughes et al., 2023). Lastly, greater VMT contributes to more wear and tear on infrastructure, which, in turn, increases the risk of crashes. Additionally, the construction and maintenance of facilities to address this wear and tear can introduce additional crash scenarios, such as those occurring in highway work zones.

Safety in highway work zones has become critically important in the United States, as these environments pose some of the highest risks to workers and motorists. A survey conducted among highway construction firms in the U.S. revealed a concerning trend of increasing work zone crashes. In 2019, an alarming 67% of construction firms reported experiencing at least one work zone crash, a significant increase compared to 2016 when only 39% of firms reported such incidents (2019 Highway Work Zone Safety Survey, 2019). This surge in work zone incidents is not limited to minor accidents; it has also led to a distressing rise in work zone injuries and fatalities. In 2010, there were approximately 37,400 injuries and 586 deaths reported in work zones. By 2018, these figures had climbed to 45,400 injuries and 756 fatalities (National Safety Council, 2020). One promising countermeasure with the potential to significantly reduce work zone crashes is the utilization of Work Zone Intrusion Alert Systems (WZIAS). These systems employ a network of sensors placed in proximity to the work zone perimeter to detect intrusions and promptly trigger alerts to both workers and motorists. The widespread adoption of WZIAS has faced challenges related to reliability and setup complexity. While some studies have investigated the effectiveness of WZIAS, these assessments have often focused solely on performance metrics like alarm accuracy, noticeability, and work zone coverage, overlooking critical other factors such as placement and worker response (Awolusi & Marks, 2019; Gambatese et al., 2017; Marks et al., 2017).

Classification of crash prediction models

Based on purpose

Crash prediction models are developed for diagnostic and predictive purposes and therefore can be divided into two broad categories which are static risk models and dynamic risk models.

Static risk models rely on aggregated traffic and crash data collected over extended periods. However, they struggle to capture the nuances of varying traffic conditions. These models are primarily employed to assess the risk associated with various factors, such as road geometry, traffic volume, historical crash data, and environmental conditions. Their primary purpose is to assess safety interventions and evaluate crash risk; therefore, they are diagnostic in nature. In essence, they play a reactive role in traffic management, utilizing static covariates that remain constant over time. On the other hand, dynamic risk models, often referred to as real-time crash prediction models, leverage disaggregated real-time traffic data to predict crashes based on observed relationships between dynamic conditions and accident occurrences. These models enable traffic engineers and planners to monitor traffic dynamics in real time, identify hazardous conditions, and proactively implement measures to prevent crashes and alleviate associated congestion. Dynamic risk models use dynamic or time-varying covariates to make crash predictions, enabling a more proactive approach to traffic safety management.

Based on type of outcome

Crash prediction models are broadly categorized into two types based on the outcomes they address: crash frequency models and crash severity models. Frequency models typically employ count data models, such as Poisson or Negative Binomial regression (Khattak et al., 2002; Ozturk et al., 2013; Qi et al., 2005; Venugopal & Tarko, 2000). For example, Safety Performance Functions (SPFs), which are often used to estimate the number of crashes in a

location, are essentially crash frequency models. On the other hand, discrete choice models like Multinomial Logit and Ordered Logit are commonly used in crash severity models (Li & Bai, 2008, 2009; Osman et al., 2016, 2019; Osman, Mishra, et al., 2018; Osman, Paleti, et al., 2018; Zhang & Hassan, 2019). Additionally, some studies have employed multivariate models capable of predicting both crash severity and frequencies (Ma et al., 2008; Ma & Kockelman, 2006a, 2006b; Song et al., 2006; Ye et al., 2013).

It is worth noting that in the recent years, the emergence of data-driven techniques, such as data mining and Machine Learning (ML), has transformed the landscape of crash prediction modeling (e.g., see (Y. Chang & Edara, 2018; Mokhtarimousavi et al., 2019; Yahaya et al., 2020; Zeng & Huang, 2014)). These techniques offer superior predictive accuracy compared to traditional econometric frameworks. However, it's important to note that they have limitations in terms of transportability and the ability to provide causal inferences. This has maintained the relevance of econometric frameworks, particularly when extracting causal relationships, and generating crash forecasts under various policy scenarios through variable coefficients and marginal effects.

Duration based crash prediction models

Duration models, also known as survival models, are econometric frameworks utilized to analyze the time elapsed between events. These models rely on survival or hazard functions and are particularly well-suited for studying the duration until a specific event occurs. Originally borrowed from epidemiology, where they are used to investigate the time between events like the onset of a disease, recovery, or fatality, duration models first found their way into the field of transportation to analyze crash data (H. L. Chang & Jovanis, 1990; Jovanis & Chang, 1989). It has since been used for several applications to model crashes on highway intersections (e.g., (Bagloee & Asadi, 2016)**)**, predict clearance time for roadway incidents (e.g., (Nam &

Mannering, 2000)), predict crashes on highway work zones (e.g., (Thapa & Mishra, 2021)**).** From a crash analysis perspective, survival models can be employed to describe the conditional probability of an event, such as a crash, occurring at a specific time, given that it has not occurred until that point. Duration models, in their simplest form, assume a constant effect of a predictor variable on the outcome over time, which is referred to as a constant hazard rate. Consequently, they are well-suited for developing reactive crash prediction models. However, it's important to note that these models have limitations when it comes to incorporating time-varying covariates for proactive crash prediction. They are not incapable of handling dynamic factors that change over time.

Research objectives

The present dissertation addresses two significant gaps as previously discussed: (i) diagnostic analysis of work zone crashes through examination of WZIAS layout and its influence on worker reactions and safety by implementing a novel duration-based approach to identify work zone crashes using field experiments, and (ii) the creation of an innovative proactive duration-based crash prediction framework capable of incorporating dynamic covariates. These are tackled as follows:

1. To gain a comprehensive understanding of WZIAS effectiveness, this dissertation goes beyond evaluating system capabilities and intrusion characteristics and considers the layout of WZIAS in the context of work zone crashes, filling a significant gap in current studies. Notably, despite the critical importance of WZIAS layout on system performance and the potential safety implications of its implementation, there are currently no formal guidelines or standards in this regard. Therefore, current study aims to contribute by examining the potential impact of WZIAS layout on work zone crashes using empirical data with the goal

of providing valuable insights and recommendations to improve work zone safety in the presence of WZIAS.

2. This dissertation developed an innovative proactive crash prediction model based on duration model that can effectively incorporate dynamic covariates. The prediction framework is tailored to handle time-varying covariates and make future crash probability predictions. This is achieved by discretizing the time intervals between crashes, allowing treatment of these discrete time intervals as outcomes in a multinomial logit model. This approach enables incorporating time-varying covariates into predictions, enhancing the model's accuracy and utility.
3. The duration-based model prediction framework is further expanded to incorporate crash severities. We enhance the discretized duration-based multinomial logit model by transforming it into a nested logit model. This modification allows for the simultaneous estimation of both crash occurrence and severity, offering a comprehensive predictive framework.
4. The versatility of the duration-based crash prediction model is further demonstrated by applying it to broader safety analyses, such as understanding speeding behaviors. Specifically, the discretized duration-based crash prediction model, enhanced with mixed effects, is implemented to better understanding driver behavior within work zones. This research places a particular emphasis on exploring the connection between speeding and a range of covariates, encompassing both time-varying and static factors. Moreover, the estimated prediction model is rigorously validated to ensure its reliability.

2. Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems

Introduction

A survey of highway construction firms has shown an increase in work zone crashes over the years with 67% of construction firms in the US experiencing at least one work zone crash in 2019 while only 39% of construction firms reported crashes in 2016 (*2019 Highway Work Zone Safety Survey*, 2019). A growing statistic can also be seen in work zone injuries and fatalities. In 2010 there were about 37,400 injuries and 586 deaths reported in work zones. The numbers rose to 45,400 injuries and 754 fatalities in 2018 (*Work Zones-Injury Facts-National Safety Council*, 2020). The rise in work zone related injuries and deaths can be attributed to an increase in Vehicle Miles Travelled (VMT) across the country. Rising VMT puts additional strain on existing highway infrastructure resulting in an increased demand for highway repair, maintenance, and construction/expansion projects. Increased interaction between workers and motorists on these projects increases the likelihood of work zone crashes unless effective countermeasures are taken. A potential countermeasure that could have notable impact on work zone crashes is the use of alert mechanisms called Work Zone Intrusion Alert Systems (WZIAS) that can detect intrusions and alerts workers. These systems pre-dominantly employ sensors placed near the work zone perimeter to detect intrusions and alarms placed close to or carried by the workers.

The first prototypes for WZIAS were developed by (Stout et al., 1993) under the Strategic Highway Research Program. The program introduced wireless and pneumatic sensor-based systems for use in maintenance work zones. Although the systems developed under the program were never adopted, systems that are currently available in the market are largely based

on ideas developed during the project. Present day WZIAS can be broadly divided into two categories based on their detection mechanism. These are; i) advanced warning systems capable of detecting potential intrusions before they occur, and ii) systems capable of detecting intrusions after vehicle enters a predefined work zone perimeter (Eseonu et al., 2018; Marks et al., 2017). Advanced alert systems typically use radar to track speed and trajectory of an incoming vehicle and alert the driver and workers when an intrusion is likely to occur. On the other hand, systems that detect intrusions after a vehicle crosses a predefined work zone perimeter employ sensors that surround a work zone perimeter. These sensors typically detect intrusions based on mechanical impact and can be mounted on traffic channelizers or laid on the ground.

Since the first prototypes were developed in 1993, numerous systems have been developed and tested for potential use, but their adoption has been limited due to unreliable performance and difficult setup. Notably, studies that have found WZIAS to be effective, have based their conclusions on their performance drawn from alarm accuracy, noticeability and work zone coverage (J. Gambatese & Lee, 2016; Marks et al., 2017; Novosel, 2014). In doing so external factors that are beyond system performance and capabilities have been ignored. For example, the speed of an intruding vehicle could have considerable impact on the occurrence and outcome of an intrusion. Furthermore, to avert crashes from high-speed intrusions, WZIA layout (separation between the system and workers) should be duly considered to guarantee the effectiveness of a system. These factors have not been accounted for by past studies. Take for example the studies undertaken by (J. Gambatese & Lee, 2016) and (Marks et al., 2017), the authors in both studies comprehensively evaluate worker response to system alerts but provide no further analysis of the results or how it could be utilized for planning layouts for WZIAS. In other words, inclusion of system capabilities, intrusion characteristics, and WZIAS layout in

investigating work zone crashes is missing in the literature. Identification of appropriate layouts for WZIAS is particularly important considering its impact on system efficacy, and the potential safety implications from its implementation. Currently, no formal guidelines or standards on WZIAS implementation exist, and we believe this is the first study investigating the potential impact of WZIAS layout on work zone crashes using experimental data.

The rest of the paper is organized as follows. In the following section we present an introduction of three systems used in field experiments followed by our review of the literature. In the methodology section, we discuss the experimental setup and the modeling approach used in the study. The data section presents a summary of the experimental data collected from our field experiments. The results from our tests and analyses are presented in the results section followed by the implications of the study and conclusion.

Overview of WZIAS

A typical work zone layout for a four-lane, two-way road with single lane closure is presented in Figure 1. Approaching vehicles first arrive at the advance warning area where regulatory and warning signs warn travelers of the work zone downstream. Traffic channelizers are used to separate the work one from adjacent lanes with active traffic. Work zones are comprised of three distinct areas, the transition area, the activity area, and the termination area. The transition area is setup using traffic channelizers laid out at about a 45-degree angle. This area provides travelers with space to adjust their speed and begin merging with the traffic on the adjacent lane. The work area within the activity area is where the actual construction work is undertaken. Buffer spaces are provided on either side of the work area to provide adequate space for workers and equipment. The termination area downstream of the activity area provides space for vehicles to shift to the adjacent lane after it has crossed the work zone. Traffic channelizers in this portion of

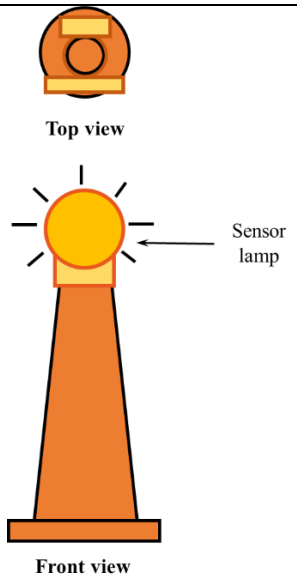
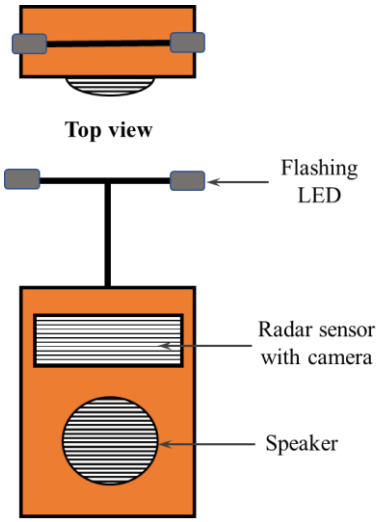
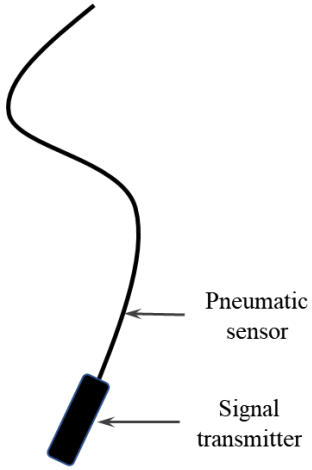
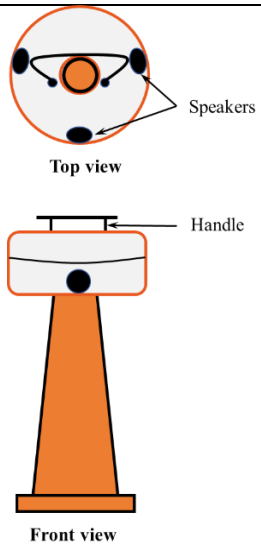
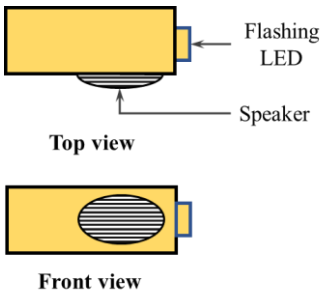
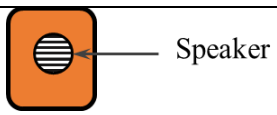
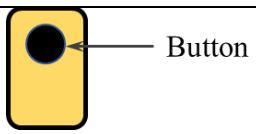
the work zone is set at a steeper angle compared to the transition taper. The Manual of Uniform Traffic Control Devices (MUTCD) provides guidelines for the length of advanced warning, transition, and termination areas based on the operating speed on the highway. However the guidelines provide no specific recommendations on buffer spaces (Federal Highway Administration, 2009a).

In the following section we provide an overview of the three systems used in this study. A schematic presenting the system components is provided in Table 1 with manufacturer recommended deployment strategy in Figure 2.

Impact Activated System (IAS)

IAS is a wireless radio-based alarm system that comprises of i) cone mountable sensor lamps, and ii) site alarms. The typical deployment strategy for the system is to mount the sensor lamps on traffic cones around the work zone perimeter with the site alarm placed in the work area close to the workers When an errant vehicle intrudes the work zone perimeter, it knocks over down the traffic cones. The sensors mounted on these cones use built in accelerometers to detect the impact and relay alert signals to nearest site alarms. When the alarms are not in range of the sensors, the alerts are relayed from one sensor to another until it reaches the nearest alarm. The site alarm on receiving the alerts produces flashing lights and a high-pitched sound alarm. The system is also capable of transmitting alerts between alarms over long distances. This is achieved by creating a zone of operation for a set of alarms which enables them to communicate over mobile networks. This feature enables the system to be extended over a long distance. For this reason, it is recommended for a site alarm be kept close to the transition taper after it has been connected to alarm(s) placed in the work area. This ensures that alert signals travel from one alarm to another even when the sensors fail to relay the signals over long distances (Figure 2).

Table 1
Schematic representation of system components

System components	Systems		
	IAS	**RAS**	**PAS**
Sensor	 Sensor lamp mounted on a traffic cone.	 The main assembly that acts as a sensor cum site alarm.	 Pneumatic sensor with the signal transmitter.
Site alarm	 Site alarm mounted on a traffic cone.		 Site alarm with inbuilt alarm and warning LEDs.
Personal alarms		 Personal alarm equipped with a speaker	 Personal alarm equipped with a button for resetting the system after it is triggered

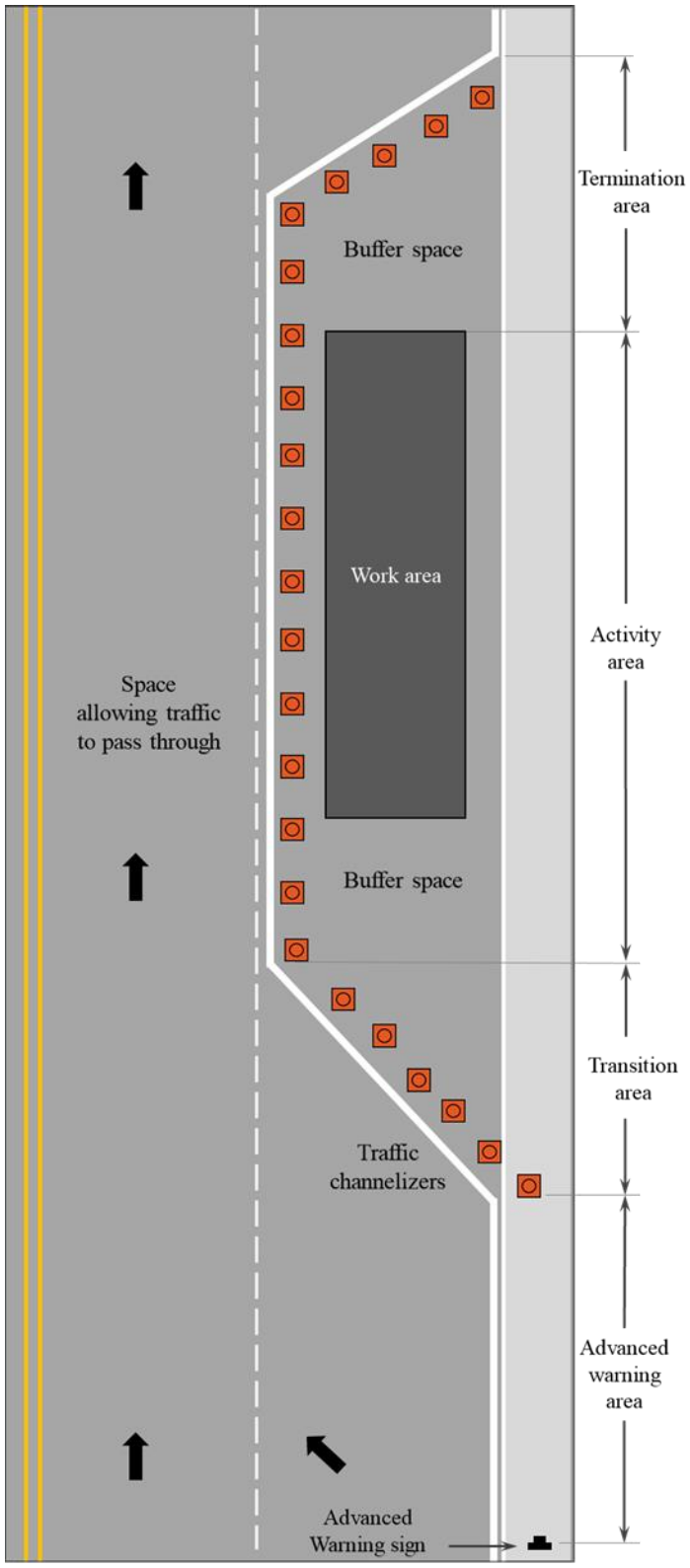


Figure 1 Typical work zone layout for a single lane closure

Summary

IAS:

- Has cone mountable sensors and a site alarm
- Sensors placed around the work zone

RAS:

- Has alarm/sensor unit and personal alarms
- Alarm/sensor unit placed facing the traffic
- Designed primarily for use by flaggers

PAS:

- Has sensor hose, site, and personal alarms
- Sensor hose laid across the lane closure

Legend

Advanced warning signs

Work area

Traffic cones

IAS

Sensors mounted on traffic cones

Site alarms

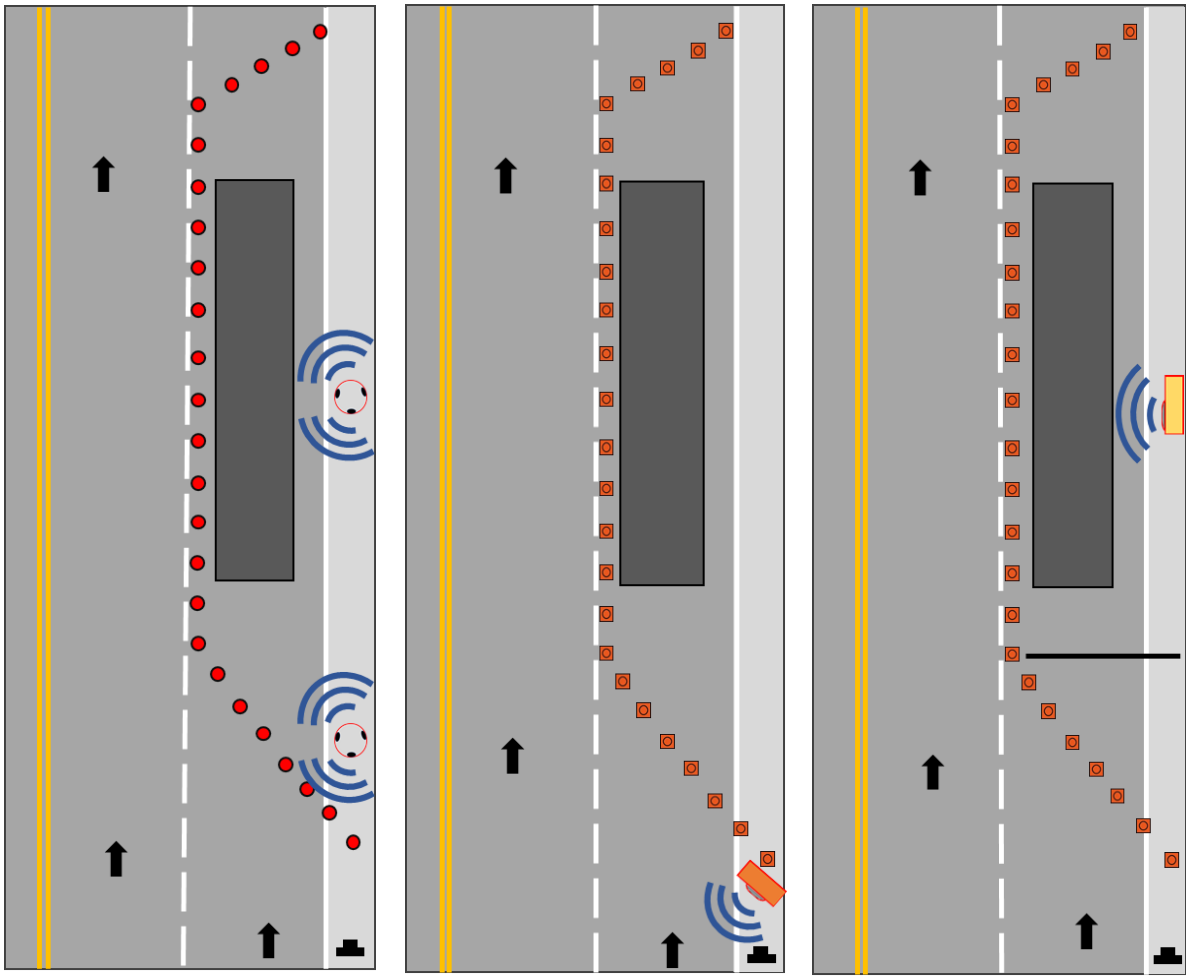
RAS

Alarm/sensor assembly

PAS

Sensor hose

Site alarm



IAS RAS PAS

Figure 2 Manufacturer recommended deployment for the three WZIA systems

Radar Activated System (RAS)

RAS is an advanced warning system capable of detecting vehicle speed and tracking its trajectory using radar. The system comprises of two components: i) a sensor/alarm unit consisting of a sensor unit and an alarm housed in a wheeled case, and ii) personal alarms for workers. The sensor/alarm unit has a built-in camera and LEDs. Personal alarms for the system are mobile sized devices that can be strapped onto a worker's arm or carried in pockets. The system is primarily intended to be used by flaggers but can also be used in advanced warning area as a standalone system to detect and warn the drivers and workers of vehicle speeding towards a work zone. As presented in Figure 2 the recommended setup for the system is to place the system in the shoulder with a flagger. Prior to its deployment, a smartphone application is needed to fully configure the system. The application configures the relative position and orientation of the system with respect to the road, and the threshold speed limit for detecting intrusions. When vehicles approach the work zone at high speed beyond the threshold speed limit, the system marks the vehicle as an intruder and activates alarms on the sensor/alarm unit and personal alarms. The personal alarms produce a high-pitched chirping sound and vibration as alerts.

Pneumatic pressure Activated System (PAS)

The PAS is comprised of three components: i) a pneumatic trip hose sensor with a signal transmitter, ii) a site alarm, and iii) personal alarms for workers. The sensor is designed to detect pressure on the hose after it has been run over by an intruding vehicle. Therefore, it is recommended that it be laid across the lane closure at the end of transition taper where the intruding vehicle is most likely to run over (Figure 2). The site alarm is housed in a hard case which is recommended to be placed in the work area close to the workers. Additionally, workers

can also use mobile sized personal alarms which can be carried on a pocket or strapped onto an arm. These personal alarms also facilitate remote reset of the system after it has been triggered. On detecting pressure, the transmitter attached to the hose sends alerts to the site alarm and personal alarms within its range. The site alarm produces sound alarm with a red blinking light and the personal alarms produce a vibratory alert. A summary of the components, and deployment strategies for the three systems is provided in Table 2 .

Literature review

Evaluation of WZIAS

Evaluation of the first WZIAS prototypes developed by (Stout et al., 1993) was carried out by the Kentucky Cabinet in 1996 (Agent & Hibbs, 1996). The study concluded that further testing on the systems was necessary before implementation on a large scale. In more recent years, several new systems have been developed and tested, however, the findings from most of these studies have cast doubt regarding effectiveness of systems. In 2010, a cone mountable tilt activated intrusion alarm employing an air horn was tested for its efficacy. The air horn used compressed CO_2 to produce high intensity alarm. The system reportedly was not efficient for use due unsatisfactory performance because of tedious setup, low durability, and frequent misfires during setup and storage. In 2012, the Minnesota Department of Transportation designed a non-intrusive advanced warning system capable of producing audio-visual alarms when vehicles crossed a certain speed limit (Hourdos, 2012). The system was called Intelligent Drum Line (IDL) and it employed a series of modified drums kept about 300 ft apart. These drums could detect the speed of approaching vehicles using radar, communicate this information to other drums and produce warning alert to the driver when certain threshold speed was passed. The warning alerts were also designed to be turned off automatically after the drivers rectified their

speed. Limited tests were conducted on the system and there is no mention of the system being used or tested afterwards. A wireless sensor network-based intrusion alert system using traffic cone mountable sensor nodes and warning devices was developed and tested by researchers for short-term work zone in 2016 (Martin et al., 2016). The system employed a barrier mountable sensor that used ultrasonic waves and a modified wristwatch to detect vehicles and alert workers, respectively. Tests carried out suggested that the system was reliable and accurate. Among the most studied systems in recent years is a radar based advanced warning system. The system uses a radar sensor to detect vehicle speed and location, and alerts workers in advance when the vehicle approaches at a high speed. The system has been subjected to several studies with promising results (Eseonu et al., 2018; Marks et al., 2017; Theiss et al., 2017; Ullman et al., 2016). The alarm siren produced by the system has been found to be particularly effective due to its resemblance to law enforcement (Ullman et al., 2016). Similarly, the other two systems that have been tested in the past are an impact activated perimeter intrusion detection system and a pneumatic trip hose sensor system (Eseonu et al., 2018; J. A. Gambatese et al., 2017; Marks et al., 2017; Novosel, 2014). The impact activated intrusion detection system uses traffic cone mountable sensors to detect impact from an intruding vehicle using built in accelerometers and relays alerts wirelessly to site alarms that produces a high-pitched alarm. Previous evaluations have suggested that the system is ideal for use in high speed highways that require long tapers although specific deployment strategies detailing layout of the system components has not been addressed (Marks et al., 2017; Novosel, 2014). The pneumatic trip sensor system used pneumatic sensors, site and personal alarms. Intruder vehicles are detected by the system only after the sensor hose has been runover. Therefore, positioning of the sensor hose is particularly important when the system is being used. When a vehicle is detected by the sensor, attached wireless

transmitter then transmits wireless alert signals to alarm units. Past findings suggest that the system is ideal for short-term maintenance work zones where larger work zone coverage is not required and frequent removal/installation of system is needed (Marks et al., 2017). However, further investigation regarding the strategic layout of the system is warranted.

To summarize in brief, although older systems have been proven to be inefficient and difficult to use, newer systems have been found to be more useful and promising. Several studies have been conducted on prospective systems over the years with the objective of evaluating their efficacy. These studies have however omitted any investigations related to practical implications of the system. More specifically answers to questions such as "*How will the layout of the system effect worker response to intrusions*?", and "*How can we deploy the system in the field to guarantee it performs with outmost efficacy*?" has not been communicated by prior studies.

Highway crash analysis

Studies investigating causal factors influencing highway crashes have heavily relied on count data models and logistic regression to model crash frequency and crash severity respectively (Lord & Mannering, 2010; Ma et al., 2008; Ma & Kockelman, 2006a; Song et al., 2006; Stipancic et al., 2019; C. Wang et al., 2011; Yang et al., 2015; Ye et al., 2013). These modeling techniques, however, only permit separate investigation of crashes (based on frequency and severity) due to the nature of the response variables. Therefore, in more recent years several multivariate modeling techniques have been employed to simultaneously model crash frequency and severity (Ma et al., 2008; Ma & Kockelman, 2006a, 2006b; Song et al., 2006; Ye et al., 2013). On similar lines, count data and logistic regression models have also been exceedingly used to study the frequency ((Khattak et al., 2002; Ozturk et al., 2013; Qi et al., 2005; Venugopal & Tarko, 2000)) and severity of work zone crashes (Y. Li & Bai, 2008, 2009; Osman et al.,

2016; Osman, Mishra, et al., 2018a; Osman, Paleti, et al., 2018a; Osman et al., 2019; K. Zhang & Hassan, 2019a), respectively. Additionally, the application of more novel techniques has gained momentum over recent years. For example, studies have explored genetic (Hashmienejad & Hasheminejad, 2017; Y. Li et al., 2018; Meng & Weng, 2011) and machine learning algorithms (Y. Chang & Edara, 2018; Mokhtarimousavi et al., 2019; Yahaya et al., 2020; Zeng & Huang, 2014) to model highway and work zone crashes. Similarly, the use of survival or hazard-based models have also gained popularity recently (Keramati et al., 2020; Wu et al., 2020). For example, (Keramati et al., 2020) used a survival model to simultaneously account for frequency and severity of crashes occurring on highway-rail grade crossings. The authors modeled crash severities as competitive outcomes with crash as the event of interest. Likewise, (Wu et al., 2020) used survival analysis to model crash counts and time interval between crashes and estimate crash modification factors for safety treatments.

Survival analysis is used to model the time until occurrence of an event using a survival or hazard functions (H. L. Chang & Jovanis, 1990; Jovanis & Chang, 1989). It is well suited for analyzing time related data where time until occurrence of an event is of interest such as the time until the onset of a disease following some medication, relapse from a disease or even the time interval between highway incidents. In transportation safety research, use of survival analysis has been mostly dominated by its application on experimental data. For example, (Sharma et al., 2011) used hazard functions for estimating dilemma zones for drivers in high-speed intersections and proposed an algorithm for reducing conflict on dilemma zones using field data. Similarly, (Choudhary & Velaga, 2020) and (M. M. Haque & Washington, 2015) used parametric hazard models to model driver stoppage during distraction using driving simulators. On similar lines, (Shangguan et al., 2020) investigated the impact of adverse environmental conditions on driver's

braking and speed reduction behavior to avoid rear end crashes using data collected from a driving simulator. Parmet et al. (2014) used survival analysis to analyze response time in driver related hazard perception concluding that hazard-based modeling approach was an appropriate approach for investigating hazard perception when using response times generated from simulations. Other safety related studies utilizing survival analysis have investigated lane keeping behavior of cyclists (Guo et al., 2013), crashes at urban intersections (Bagloee & Asadi, 2016), impact of connected vehicle environment on lane-changing behavior using data collected from a driving simulator (Ali et al., 2019), and predicting clearance time for road incidents (Chung, 2010; Nam & Mannering, 2000; Tang et al., 2020). However, its application for investigating work zone crashes and its causal factors is non-existent. Understandably, it is challenging to collect work zone crash data using field experiments and driving simulators considering the safety of the participants and the limitations imposed by simulators.

This study was in part inspired by the evident gap in the published literature concerned with the investigation of work zone crashes using survival analysis. To our knowledge no previous studies have applied survival analysis to work zone crashes. Furthermore, the goal of this study is to identify and recommend guidelines on WZIAS layout which has potentially huge implications for WZIAS implementation. In view of these gaps, we present the three main research needs addressed by this study in the following section.

Research gap and study objectives

Based on the review of literature, we identify and rid of the following gaps with this study.

i. Past studies investigating the efficacy of WZIAS have been based solely on their performance (J. A. Gambatese et al., 2017; Marks et al., 2017). Therefore, causal factors that are extrinsic to the systems have not been considered in these studies. Two of such factors

are considered in this study, i) speed of intrusion, and ii) layout of WZIAS. In doing so we recommend best practices for choosing and deploying systems in the field. The impact of high-speed intrusions on work zone crash could be partially negated by devising appropriate system deployment strategies that facilitate quicker worker response. Since the deployment strategy is unique to each system, the relative position of the system components with respect to the work zone perimeter and workers is likely to vary based on choice of the system and work zone closure. Considering this, it is imperative to identify ideal use case scenarios for each system and establish best deployment strategies for their implementation. Although a prior study has made recommendations on selection of systems (Marks et al., 2017), we go a step further and recommend ideal deployment strategies as a means to translate theoretical knowledge on system characteristics and performance into work zone standards for real world application using experimental data.

ii. Our study analyzes workers' naturalistic response to system alerts to investigate the occurrence of work zone crashes. While the analysis of naturalistic response by itself is not new to the literature, analysis of worker responses is rather novel since published research almost in its entirety has been centered around drivers (Choudhary & Velaga, 2020; Dingus et al., 2016; M. M. Haque & Washington, 2015; Shangguan et al., 2020). These studies have analyzed drivers' braking response collected using driving simulators. In contrast, our approach aims to imitate work zone crashes to collect worker response in the field for two main reasons. First, it allows us to collect the response time, i.e., the time taken by workers to perceive and react to an alarm (move out of the way to safety). The exact time taken by a worker to react cannot be collected without field experiments. Second, collecting worker responses using driving simulators is particularly challenging. Although driving simulators

are effective in studying driver behavior, they provide limited to no scope for incorporating WZIAS and recording the worker response. Furthermore, unexpected problems that are frequently exhibited by WZIAS in the real world, such as false alarms and delayed activation are best studied using field experiments.

iii. We employ non-parametric and semi-parametric survival models to analyze worker response and occurrence of crashes in presence of WZIAS using field experiments. To our knowledge, application of survival analysis to this end has not been done in the literature.

Method

As previously mentioned, this study utilized field experiments to collect and analyze workers' naturalistic response to work zone intrusion alerts produced by WZIAS. Various WZIAS layouts and intrusions speeds were used to emulate different scenarios for work zone intrusions. Worker responses to the alerts produced by WZIAS upon detection of these intrusions were then used to determine potential crashes. Determination of crash was based on worker response and alerts produced by the systems. In the following sections we discuss the experimental arrangements, procedures, and explain the methodology used to determine crashes.

Pilot testing

Field experiments for the study were conducted in two phases. In the first phase a pilot test was conducted to determine the maximum signal transmission range for the system components. This was important to ensure that the layout of the system components in our experiments was such that they were not too far apart to result in a loss of signal during transmission. The transmission range was determined as follows. The distance between the system components, sensor and alarm units, were gradually increased at 50 feet intervals. At each interval four attempts were made to activate the alarms by triggering the sensors. If all four attempts were successful, the

transmission was assumed to be complete. The maximum distance beyond which complete transmission ceased was considered as the maximum transmission distance (Novosel, 2014). This methodology was applied to find the transmission range for the following system components.

- IAS: Sensor to site alarm.
- RAS: Main assembly (sensor/alarm) to personal alarms.
- PAS: Pneumatic sensor to personal alarm.

As expected, different transmission ranges were obtained for the systems. For IAS, the transmission range from sensor to site alarm was 300 ft while for RAS the transmission range between the main assembly and personal alarms was about 400 ft. For PAS, complete transmission between sensor and site alarm was limited to 150 ft.

Table 2
Summary of systems specifications

	IAS	RAS	PAS
System components	• Cone mounted sensor lamps, and • Site alarm	• Sensor/alarm unit consisting of radar-based sensor, flashing LEDs and alarm speaker, and • personal alarms	• Pneumatic trip hose sensor, • site alarm, and • personal alarms
Alert mechanism	• Motion detection from vehicular impact on the traffic cones	• Radar based vehicle tracking	• Pressure exerted by vehicle running over the trip hose
Type of alert	• Sound and flashing lights	• Sound and flashing LED on the sensor unit, and • vibratory and sound alert on personal alarms	• Sound and flashing lights on site alarm, and • vibratory alert on personal alarms
Deployment	• Sensors mounted on traffic cones placed around the work zone perimeter, and • site alarm close to the workers	• Main unit placed on the shoulder outside the transition taper facing the oncoming traffic, and • personal alarms carried by the worker	• Pneumatic sensor laid across the closed lane in transition area, • site alarm within the work area, and • personal alarms carried by the workers

Transmission range can provide a reasonable estimate of response time needed to avert a crash. For example, when using systems with greater ranges, the sensor and alarm can be placed further apart which would provide workers with more time to react to an intrusion as the vehicle entering the perimeter will have to traverse longer distance before reaching the work area. This knowledge can aid in determining system layouts. This is particularly relevant for systems based on mechanical impact and pressure detection such as IAS and PAS. However, the same is not applicable to advanced warning systems like RAS since they are capable of alerting workers in advance. In such a case, detection range of the system can be used as a surrogate measure to estimate optimal layout of the system. Detection range can be defined as the minimum distance between intruder vehicle and the system needed to trigger an alarm.

In this study, the detection range for RAS was tested for different test speeds. In these experiments, test vehicles were driven towards the RAS main assembly at predetermined test speeds and the moment of alarm activation was recorded using video cameras. Using the recordings, the exact point at which the alarms were triggered was identified and the distance of the point from the main assembly was measured. Results suggested that the detection range was comparable to the standard Stopping Sight Distance (SSD) for the respective test speeds. Table 3 presents the results from pilot testing for transmission and detection range. The standard values of SSD for the test speeds are also provided within parenthesis.

IAS and PAS were selected for the next phase of testing wherein worker response post intrusion was collected. RAS was excluded from the second phase of tests considering advanced detection and warning.

Field testing

The field tests were conducted in a controlled facility that was closed to traffic and pedestrians. A typical lane closure identical to Figure 1 was set up using traffic channelizers to imitate a work zone. Five highway maintenance workers from TDOT were recruited as test subjects for the study. National demographic of highway construction workers suggested that only about 2.5% of the highway maintenance workers in the US were female and the average age of workers was about 44 years (*Data USA: Highway Maintenance Workers*, 2018). The workers were selected to represent this demographic. All participants in the study, drivers, and workers, were certified and experienced in highway construction and maintenance. They were also informed regarding the methodology and objective of the study before the field tests began.

Table 3
Results from pilot testing of WZIAS

Tests	**IAS**	**RAS**	**PAS**
Transmission range			
Sensor to site alarm	300 ft	NA	NA
Sensor to personal alarms	NA	400 ft	150 ft
Median detection range (n=3)	**Observed range (Standard SSD)**		
Test speed			
30 mph		175 ft (200 ft)	
45 mph		350 ft (360 ft)	
60 mph		500 ft (570 ft)	

During the experiments, the systems were setup in the lane closure following manufacturer recommendations presented in Table 2 . The workers were then positioned close to a hypothetical work area and asked to engage in an activity of their choosing in a sitting position facing away from the incoming test vehicle. To obtain naturalistic response to the intrusion, workers were not provided prior information on when an intrusion would occur. They were also instructed to react only to the alerts produced over the devices (site or personal alarms). Test vehicles were then driven into the lane closure at various speeds to imitate intrusions. Several

safety precautions were adopted to ensure the safety of the participants. Drivers of the test vehicles were instructed not to deviate from the course of their trajectory and travel on the same lane while the workers were positioned away from the trajectory of the intruding vehicles on the adjacent lane. The workers were also asked to respond by moving away from the lane closure towards the shoulder upon receiving alerts from the system being tested. To counterbalance order effects, workers were randomly chosen for experiments. A randomly chosen worker would participate in tests for a certain configuration of a system. After completion of tests on the configuration, the next worker was then chosen at random to participate in the same experimental configuration and so on. After completion of tests for a certain configuration the system being tested was switched and the tests were then carried out in a similar manner.

The experiments were varied by intrusion speeds, and relative position of the system sensors to the workers. Intrusion speeds ranging from 30-60 mph at 5 mph increments were considered for the study. The relative position between sensors and worker varied from 100-300 ft for IAS and 100-150 ft for PAS considering their transmission range as shown in Figure 3. Besides predefined speed and sensor-to-worker spacing, data was collected on i) activation of alert; ii) noticeability of alarms measured using sound intensity; and iii) worker reaction time during each experimental trial. A description of the data collected is provided in Table 4 and the various experimental configurations is summarized in Table 5. Consequently, the outcome of the intrusion, i.e., if an intrusion resulted in a crash, was decided based on activation of alert and worker reaction recorded using video cameras (see section 0 for detailed explanation).

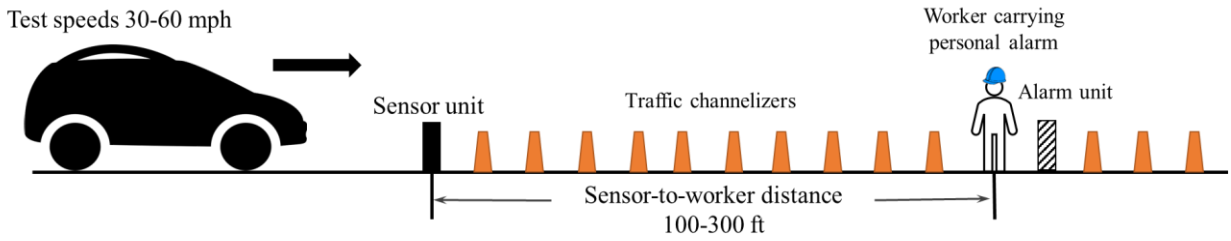


Figure 3 Schematic representation of worker relative to the system components

Crash determination and hypothesis formulation

Determination of whether an intrusion would result in a crash was based on three possible outcomes following an intrusion. These outcomes were based on worker reaction time (t_w), critical time (t_c), and activation of alarms. Worker reaction time for each experiment was determined from video recordings while the critical time was calculated based on test speed and sensor-to-worker distance. The three possible outcomes from experiments considered are as follows.

Table 4
Description of variables used in WZIAS assessment

Variables	**Description**
Speed [u, mph]	Speed of the intruding vehicle
Sensor-to-worker [D_w, ft]	Distance between the sensor and the worker for tested system
Alert (1=Yes, 0=No)	Binary variable indicating whether the alarms activated
Sound_int (dB)	Sound intensity of the site alarm at worker location used as a measure of alarm noticeability
Worker_react [t_w, s]	Time taken by a worker to perceive and react to alarms by initiating an evasive motion to move away from the work area towards the shoulder
Critical_time [t_c, s]	Measure of time taken by the test vehicle to reach the worker after it has entered the work zone perimeter, mathematically calculated as $t_c = \frac{D_w}{u * 1.47}$
Crash (1=Yes, 0=No)	Binary variable indicating if an intrusion resulted in a crash determined as follows: Alert =1 and $t_w < t_c$ then 0 Alert =1 and $t_w > t_c$ then 1 Alert = 0 then 1

- Outcome 1: Alarms activate and $t_w < t_c$ = No crash

In case the alarms activate, and a worker's response time is less than the critical time we assert that the intrusion is unlikely to result in a crash since the worker would have adequate time to get to safety.

- Outcome 2: Alarms activate and $t_w > t_c$ = Crash

When this outcome is observed, we assert a crash is imminent since the intruding vehicle would have traversed the distance between the sensor and the worker before the workers would have adequate time to react to the alarms.

- Outcome 3: Alarms fails to activate = Crash

Under this outcome we assume that workers would be unaware of the intrusion as the system fails to register any intrusion and therefore a crash would be imminent.

It is noteworthy that our approach in determining the outcome of the intrusion is based on workers' response. Drivers upon hearing alarms or striking traffic barriers, may often be able to break, stop or steer the vehicle to safety. Since our experiments were based on real world interaction between an intruding vehicle and workers this limitation could not be eliminated due to safety concerns. Four hypotheses are formulated to test the effect of the variables on work zone crashes. These hypotheses are as follows:

H1: With increase in sensor-to-worker distance, the probability of work zone crashes will decrease since the critical time increases.

H2: Greater latency in signal transmission increases the probability of work zone crashes as worker.

Since worker reaction time is dependent on the latency of signal transmission, system with shorter latency could be better able to reduce work zone crashes. Latency in signal transmission is defined as the time between intrusion detection and alerts.

Table 5
Summary of experimental configurations for WZIAS

Variables	**Experimental configurations**
Speed	Between 30-60 mph at 5 mph intervals
Sensor-to-worker	
IAS	Set at 100 ft, 200 ft and 300 ft
PAS	Set at 100 ft, and 150 ft

Note: There were a total of 7(Speed) x *(2+3)(Sens_to_alr) = 35 experimental configurations for the two systems.*

H3: With the increase in speed of the intruding vehicle, the probability of work zone crashes will increase.

As the speed of the intruding vehicle increases the critical time decreases and quicker responses from workers will be required to avoid crashes. Therefore, with higher intrusion speeds, crashes are more likely to occur.

Survival analysis

Survival analysis is popularly used in many areas of research such as epidemiology, engineering, and economics to model the time until occurrence of an event. In this study, the event is occurrence of a work zone crash. In other words, our analysis models work zone crashes considering the time until its occurrence measured since intrusion of the work zone perimeter. It is worth mentioning that this study assumes any possible contact between a worker and intruding vehicle as a crash regardless of its severity.

The survival function then gives the probability of non-crash intrusion occurring at time *T* which is longer than some specified time *t.* Assuming *f(t)* is the probability density function

and *F(t)* is the cumulative distribution function of the continuous random variable *T*, the probability that no crashes occur after time *t* is given by the survival function $S(t)$ as follows:

$$S(t) = P(T > t) = 1 - F(t) \quad (1)$$

Another concept that is related to survival function is the hazard function. Hazard function *h(t)* also called the hazard rate gives the instantaneous probability of occurrence of an event (crash) conditional on no events having occurred until the time *t*. Mathematically, it can be written as:

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(T \leq t + Dt | T > t)}{Dt} \quad (2)$$

Survival analysis collectively refers to three main survival models. These models are Kaplan-Meier (KM) estimator, Cox Proportional Hazards (Cox PH) model, and Accelerated Failure time (AFT) model which belong to non-parametric, semi-parametric, or parametric family of models, respectively.

Kaplan Meir estimator

KM estimator is a non-parametric estimator of the survival function for small time intervals. It can be written as:

$$S_{KM}(t) = \prod_{i:t_i \leq t}(1 - \frac{e_i}{n_i}) \quad (3)$$

where t_i represents time at which at least one crash is observed, e_i is the number of crashes that occurred at t_i and n_i is the number of intrusions that did not results in a crash. A notable limitation of KM estimator is its ability to incorporate variable effects. Since only the time and occurrence of crashes are included in the estimator, it cannot be used to model the effects of variables. Regardless, they can be used to compare the probability of crashes between separate groups of variables using the log rank test statistic. For example, to compare the likelihood of

crashes between two different intrusion speeds, KM estimators can be used to estimate the survival functions for each speed separately and test if they are statistically different. The log rank test statistic tests the null hypothesis that the survival functions for the two groups (in this case intrusions speeds) being compared is not statistically different. The test statistic is calculated as:

$$c^2 = \frac{\sum_{j=1}^{J}(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} V_j}} \sim N(0,1) \text{ under } H_0 \quad (4)$$

where O_j and E_j are the observed and expected number of crashes, respectively for distinct time of crashes $t_1 < t_2 < t_3 \ldots < t_j$, and V_j is the variance of observed number of crashes. Semi-parametric and fully parametric models that can address the effect of variables are often preferred over KM estimators.

Cox proportional hazard model

Due to the inability of KM estimators to include variables in estimating survival functions, use of semi-parametric Cox PH and fully parametric AFT models is often preferred. Cox PH model assumes multiplicative effect of variables on some baseline hazard to study variable effects on the time until an event. The model is based on two assumptions, i) the functional form for survival function exponential, and ii) hazard rate is constant over time. Mathematically, it can be written as follows:

$$h(t|X) = h_o(t)\exp(-bX) \quad (5)$$

where for a vector of variables X, $h(t|X)$ is the hazard function, $h_o(t)$ is the baseline hazard function and $\exp(-bX)$ is the functional form of the variables with a vector of coefficients, b. The underlying proportional hazard assumption however might not hold true for all variables in a model. Test for the assumption is particularly important when the effect of variable is of interest

(i.e., test whether the effect of variable is constant overtime or not). In case the assumption is violated, the variables violating the assumption can be controlled by stratification while simultaneously including remaining variables in the model. Such a model is referred to as stratified Cox PH model. Assuming the variable violating the proportional hazard assumption has *K* levels, the modified hazard function can be mathematically expressed using the following equation.

$$h_k(t|X) = h_{ok}(t)\exp(-bX) \quad (6)$$

Here, $h_k(t|X)$ and $h_{ok}(t)$ are the hazard and baseline hazard functions respectively for k^{th} stratum with $k = 1,2,3\ldots, K$ levels of the variable that is being stratified. Note that unlike in Eq. (5) where there is a single baseline hazard function, Eq. (6) results in a different baseline hazard function for each level of the stratified variable.

Application of Cox PH model for independent and identically distributed random variables is straightforward. However, for individuals in a study that are subjected to repeated measures (i.e., when measurements are in clusters) it is necessary to account for unobserved heterogeneities arising from different clusters that may expose individuals to different levels of hazard (M. M. Haque & Washington, 2015; J. Wang et al., 2020). Unobserved heterogeneities can be accounted for in Cox PH model by adding a frailty parameter assuming that every cluster of individuals has a different frailty, and among them the frailest would die first. The frailty parameter is essentially a random effect term that multiplicatively modifies the hazard function for each cluster. The resulting modified Cox PH model is called shared frailty Cox PH model and is of the form:

$$h_{ij}(t|u_i) = h_o(t)u_i\exp(-bX_{ij}) \quad (7)$$

where, h_{ij} represents the hazard function for i^{th} individual (worker) in the j^{th} measure (experiment); b is a vector of coefficients for the variables X_{ij} and u_i is the shared frailty with mean 1 and variance θ following a gamma distribution (for example, see (Therneau et al., 2003)).

It is worth mentioning here that a third member of the family of survival models are fully parametric AFT models. These models assume that variables have multiplicative effect on the survival time. Exponential, Weibull, log-logistic, lognormal and loglogistic are some of the commonly used parametric distributions in AFT models. There are notable limitations to AFT models. Selection of appropriate distributions for AFT models is often difficult unless the underlying distribution can be identified with certainty (Kleinbaum & Klein, 2012). Also, AFT models cannot handle zero values in the response variable (J. Zhang & Thomas, 2012). For these reasons, Cox PH and stratified Cox PH were used for statistical analyses in the study. All analyses in this study were done using R v3.5.1, and R package *survival* which utilizes penalized partial loglikelihood for model fitting (Therneau, 2020; Therneau et al., 2003).

Data

A total of 525 observations (35(experimental configurations) x 5(workers) x 3(trials) were recorded from the experiments which comprised of 315 observations for IAS and 210 observations for PAS. Descriptive statistics of variables used in our analysis is shown in **Error! R eference source not found.**. Descriptive statistics for workers are presented here to provide the reader with a summary of test subjects.

Results and discussion

KM estimators are useful in determining the change in probability of survival and testing the independence of groups in absence of variable effects. Therefore, KM estimators were used for the two systems, different test speeds and sensor-to-worker distances to test the independence of survival probability. Figure 4(a) presents the result from KM estimator for cumulative probability of work zone crashes with 95% confidence interval. A large confidence interval was observed at the end of the curve which is indicative of most crashes occurring within the first seven seconds of intrusion. Similarly, the KM estimators for different groups namely systems (Figure 4(b)), test speeds (Figure 4(c)), sensor-to-alarm distance (Figure 4(d)) are also presented. The tick marks in these plots represent censored data for which no crashes were observed. Log-rank test was conducted to test independence of groups. Results from log-rank test suggested difference in survival functions across groups (Chi-square = 72.3, p-value < 0.01 for systems; Chi-square = 97.6, p-value < 0.01 for test speeds; and Chi-square = 432, p-value < 0.01 for sensor-to-worker distances).

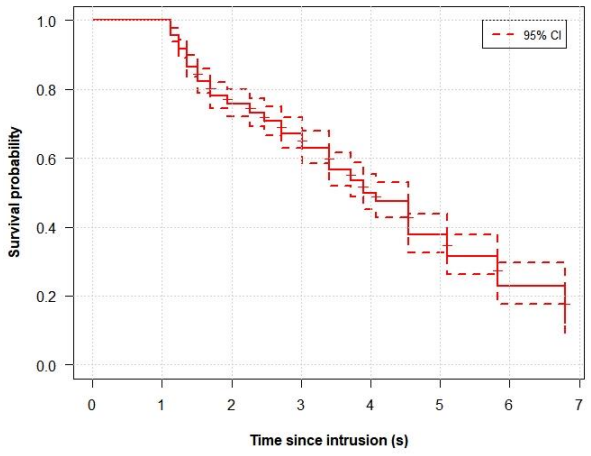
In comparing the estimators for IAS and PAS for the same time, IAS was observed to result in greater probability of survival compared to PAS after three seconds. This suggested that for longer tapers, IAS would be safer. This is because for longer tapers vehicles will have to travel for a longer duration downstream after intrusion. In such events IAS would likely result in a higher survival probability. Among the estimators for different speed groups, lower speeds displayed longer horizontal leveling. This suggested that the probability of survival remained constant for a longer period when the intruding vehicles were traveling at a lower speed. Compared to intrusions that occur at low speed, intrusions that occur at higher speeds had more noticeable impact on occurrence of work zone crashes over a shorter period. The estimators

suggest that for the same difference in time the change in survival probability for high-speed intrusions (greater than 35 mph) was higher versus low-speed intrusions. These findings hint that for high- speed intrusions, even a small increase in critical time would have measurable impact on work zone crashes. Parallel results can be drawn for the estimators on sensor-to-worker distance.

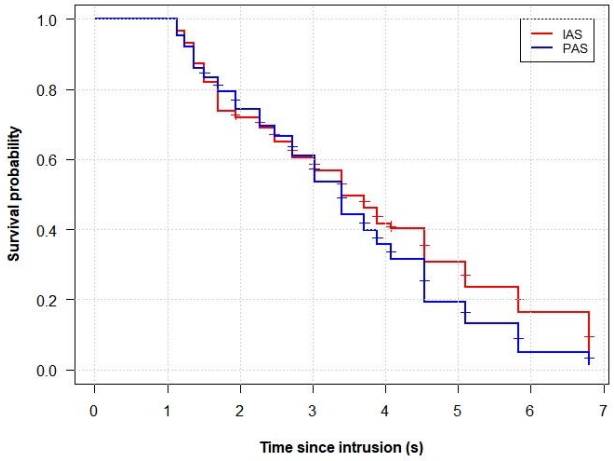
Table 6

Descriptive statistics of observations

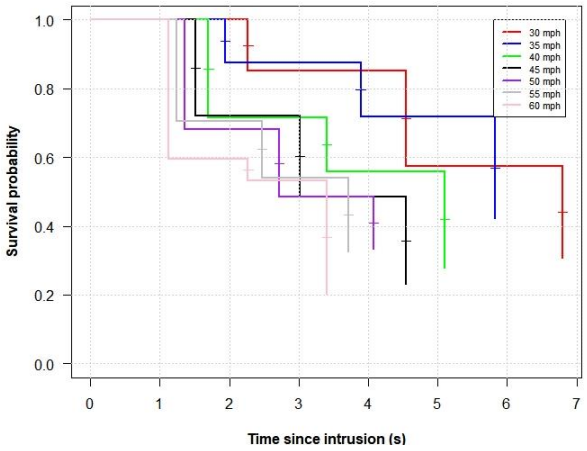
Category or Variables	**Mean**	**Std. deviation**
Worker_react		
IAS	1.98	0.38
PAS	1.96	0.41
Sound_int		
IAS		
100 ft	68.51	1.48
200 ft	57.25	1.47
300 ft	51.98	1.54
PAS		
100 ft	75.36	1.62
150 ft	69.67	1.59
System alerts and crash	**Frequency**	
IAS		
Alert (1=Yes, 0=No)	212	
Crash (1=Yes, 0=No)	120	
Total experimental trials	315	
PAS		
Alert (1=Yes, 0=No)	160	
Crash (1=Yes, 0=No)	74	
Total experimental trials	210	



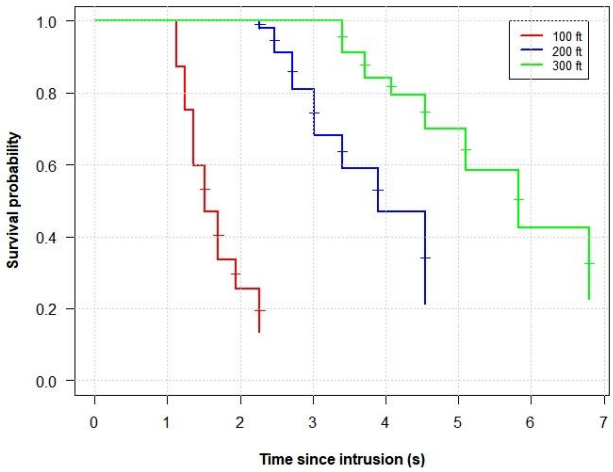
(a) Null estimator with 95% CI.



(b) Estimators for IAS and PAS.



(c) Estimators for different test speeds.



(d) Estimators for sensor-to-worker distances.

Figure 4 Kaplan Meier estimators

The vertical drop in survival probability was less frequent for great distances indicative of its positive impact on the occurrence of crashes. Therefore, for the same difference in time, the probability of survival can be expected to vary less when the separation between the sensors and workers is more. The time at which the survival probability approaches the minimum value is also noteworthy. At 100 ft, most crashes occurred within 2 seconds of intrusion while for 300 ft almost all crashes were observed between 3-7 seconds of intrusion. Based on these results it can be concluded that when workers are close to the work zone perimeter (sensor-to-worker

distance is less) even a small increment in the critical time would have measurable impact on the occurrence of crashes.

Next, three variations of the Cox PH model were fit to the experimental data obtained from both systems to investigate the effect of variables on occurrence of crashes. The first model was a Cox PH model. The second was a stratified Cox PH model that stratified variables violating the proportional odds assumption. The third model was a shared frailty Cox PH model incorporating random effects to account for heterogeneity in the data from repeated trials on the same individuals. Backward elimination approach was used to develop the models by first removing variables with high multicollinearity based on Variation Inflation Factor (VIF) followed by removal of variables that did not contribute towards model goodness of fit. Two model goodness of fit were considered while selecting variables, namely, AIC and C-statistic. Additionally, the stratified Cox PH model was developed by administering Schoenfeld test for proportional hazards assumption on the variables and then stratifying variables violating the assumption. In the shared frailty Cox PH model, a frailty term with gamma distribution (mean 1 and variance θ) was added to the Cox PH model to account for mixed effects. Summary of the three models is presented in Table 7. The shared frailty model was found to be a slightly better fit compared to the other models. Further, high values of C-statistic for all three models is indicative of their good discriminatory power (Hosmer & Lemeshow, 2000).

Table 7
Summary of the Cox models for overall survival

Model fit measure	**Cox model**	**Stratified Cox model**	**Shared frailty Cox model**
Partial loglikelihood at zero	-1379	**-1276**	-1379
Partial loglikelihood at convergence	-1163	**-986**	-1161
AIC	2335	**1978**	2234
C-statistic	0.83	**0.84**	0.83

Table 8
Result from stratified Cox model

Variables	Coefficients (SE)	Hazard ratio	VIF
Sensor-to-worker	-0.025 (0.002)***	0.97	1.30
Sound_int	0.010 (0.01)	1.01	1.33
Worker_react	0.311 (0.15)*	1.37	1.06
Number of crashes = 383			

Level of significance: ***0.001, **0.01, *0.05, # 0.1

Results from the stratified Cox model is shown in Table 8. The table presents variable coefficients with their standard errors within parentheses, and their hazard ratios and VIFs. Hazard ratios provided here can be used to quantify the change in outcome (here the probability of crash) with the change in the predictor variables. VIF for the variables in the model were close to 1 suggesting low correlation between one other (Kock & Lynn, 2012). In the initial model, the variables *Sensor-to-worker, Worker_react* and *Speed* were found to be statistically significant. Variable *Speed* was however later removed from the model due to high VIF (VIF=11). The sign of the variable coefficients gives an idea of its influence on the outcome. A negative coefficient, and hazard ratio less than 1 for a variable implies that the variable is inversely associated with the outcome. On the contrary, a positive coefficient, and a hazard ratio greater than 1 implies direct relationship between the variable and outcome. For example, a negative coefficient for *Sensor-to-worker* implies that, controlling for other factors, with an increase in sensor-to-worker distance the probability of crash decreases. More precisely the model predicts that probability of crash decreases by about 3% with every 1 ft increase in distance. The finding is intuitive since with greater separation between the worker and the sensor, intruding vehicles will need to travel further downstream after the intrusion providing additional time for the workers to react to the intrusion. This finding supports our first hypothesis H1.

Table 9
Results from Cox models for the systems

Variables	Coefficients (SE)		
	Cox model	**Stratified Cox model**	**Shared frailty model**
IAS			
Speed	0.22 (0.02)***	**0.25 (0.02)*** **	0.22 (0.02)***
Sound_int		**0.64 (0.05)*** **	0.58 (0.05)***
Alert (1=Yes, 0=No)			
Yes	-2.10 (0.29)***	-	-2.10 (0.29)***
C-statistic	0.96	**0.96**	0.97
Likelihood ratio test	342.30	**323.80**	342.7
AIC	537.6	**434.9**	537.6
Variance of gamma frailty			0.003
Number of crashes = 198			
PAS			
Speed	0.20 (0.11)***	4.08 (355.06)	**0.20 (0.02)*** **
Alert (1=Yes, 0=No)			
Yes	-1.02 (0.27)***	-	**-1.07 (0.27)*** **
Sound_int	0.56 (0.04)***	-0.01 (0.06)	**0.57 (0.04)*** **
C-statistic	0.95	0.99	**0.95**
Likelihood at convergence	-345.92	398.2	**335**
AIC	697.83	268.9	**697.2**
Variance of gamma frailty			**0.002**
Number of crashes = 185			

Level of significance: ***0.001, **0.01, *0.05, # 0.1
Note: "-" indicates the variable stratified in the model.

Similarly, a positive coefficient and hazard ratio more than 1 for *Worker_react* suggests that the variable is causally related to the work zone crashes and with unit increase in worker reaction time, probability of crash can be expected to increase by about 37%. It is obvious that work zones crashes are more likely to occur when workers fail to react timely to intrusions. Considering that the primary reason for worker's delayed response in our experiments can be attributed to greater latency in signal transmission we support hypothesis H2. Therefore, we can assert that a system's quickness in producing alert after detection is imperative towards reducing crashes. The variable *Sound_int* although statistically insignificant improved the model goodness of fit and was therefore included in the model. These results in general indicate that for any work zone regardless of the system being used, the two key factors that need consideration are

separation between the sensors and the worker and the system's ability to alert the workers in time. Among the three hypotheses, no specific findings could be reported to support or oppose H3 from the model.

The models analyzed aggregated data for both the systems. However, to study the influence of variables on each system, system specific analysis was needed. Therefore, the three variations of the Cox PH model were applied to crash data on IAS and PAS separately. The same modeling technique described in the preceding paragraphs were applied. We present the model results with parameters estimates, standard error, and model goodness of fit parameters for the models in

Table 9. Note that the variable *Alert* was stratified for the stratified Cox models for both the systems. The magnitude of coefficients for the models were comparable except for stratified Cox model for PAS. Of the three models for IAS, the stratified model was found to the superior fit. Similarly, the shared frailty Cox model was the best fit for PAS. Although model goodness of fit indicated that the stratified model was the best fit for PAS, the model was discarded due to its inconsistent estimates compared to other models. The variances of gamma frailty for IAS and PAS were found to be 0.002 and 0.003, respectively. Low magnitude of variances is indicative of small variability between the workers which can be attributed to relatively small sample size. Although accounting for mixed effects is recommended when the number of participants (workers in this case) is larger than five, interpretation of causal effects from mixed models for smaller number of participants is still considered safe (Gelman & Hill, 2007). Due to the difficulty in recruitment, this study was limited to five workers. This can be expanded further as

a potential avenue for future research. In contrast to the findings in Table 7, the influential variables for both systems were found to be *Speed, Sound_int,* and *Alert*. As expected, the coefficient for *Speed* for both the systems was positive indicating direct relationship between speed of the intruding vehicle and work zone crashes. This provided evidence to our hypothesis H3. Further, results from the frailty model for PAS resulted in a high magnitude negative coefficient for *Alert* suggesting an inverse and prominent relationship of the variable with work zone crashes.

Research implications and recommendations

The results from tests and analyses highlighted the influence of system performance and layout on work zone crashes. Results from pilot testing provided with essential information on system's transmission range and analyses of experimental data using non-parametric KM estimators and semi-parametric Cox PH models highlighted the impact of variables (i.e., *Speed, Sensor-to-alr, Sound_int, Alert, Worker_react*) on crashes. We discuss the implications of the findings in parallel with our recommendations as follows.

Speed limit

Results from our regression models in

Table 9 suggests, with unit increase in operating speed the probability of crash increases by about 22% *((exp(0.2)-1)x100%).* Since reduction in the operating speed limit could have measurable impact on crashes, we recommend reducing the speed limit near work zones whenever WZIAS is being used. Reduction in existing speed limit will reduce the probability of crash and shorten length of lane closure needed which will provide a greater opportunity for the

systems to cover the work zone (Mishra, 2013). However, reduction in speed limit should be done after careful consideration since the general practice on reduction of speed limit across the US varies with states (Bham & Mohammadi, 2011). We recommend a conservative approach that agrees with existing practices. We recommend a 5-mph and 10-mph reduction in speed limits for highways operating at 40-55 mph and 60+ mph respectively. Work zones can be set up on facilities based on their operating speed as provided in MUTCD 2009. However, appropriate guidelines and standards will need to be established for the buffer area.

Buffer space and system deployment

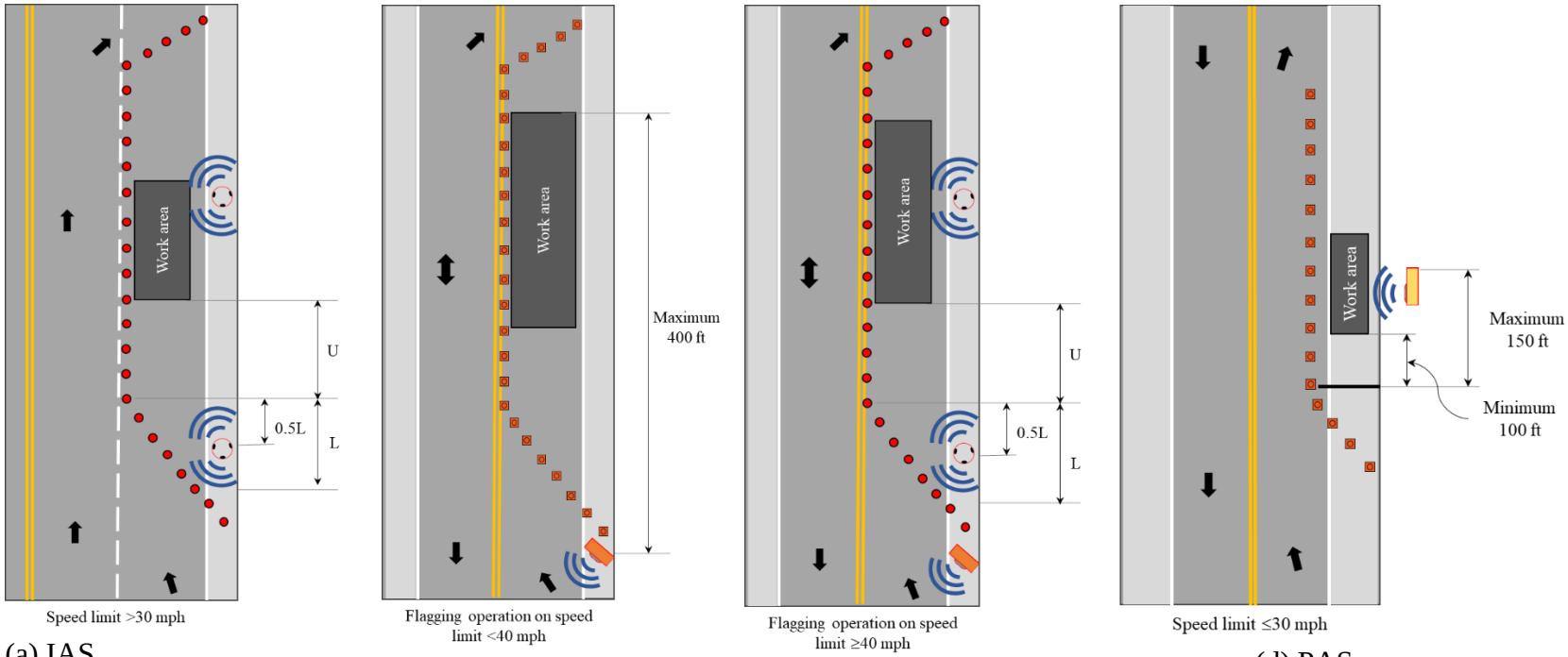
The transmission range of the system components should be given due consideration while determining the length of buffer space. We present a schematic for the recommended layout of system components based on our findings in Figure 5 In case of IAS, based on results from KM estimators (Figure 4(c)), we recommend providing minimum buffer space that in numerically equal to *revised speed limit in ft/s x 3 seconds* as most crashes above 40 mph occur within 3 seconds of intrusion. We recommend using at least two site alarms while using the system, one placed close to the transition taper and the other placed next to the work area (see Figure 5(a)). The alarm unit placed near the transition taper can be placed midway between the taper length. This configuration will ensure that intrusions detected by the sensors in the transition area is communicated to all site alarms regardless of their separation. Additionally, the spacing between the sensors in the transition taper should be based on engineering judgement such that vehicles would not be able to pass through the perimeter without striking the cones/sensors. It is noteworthy that as per MUTCD guidelines, the spacing between the traffic barriers should be limited to 40 ft on highways operating at 40 mph speed limit. Similar guidelines can be followed for the cones/sensors placed in rest of the work zones on all highways. In case of RAS, the

primary objective while using the system should be to place it within transmission range of the work area as shown in Figure 5(b). Since the system is recommended primarily for flagging, the MUTCD recommendation is to set transition taper at maximum of 150 ft for which the 400 ft transmission range of the system is adequate. When used with IAS, the layout for both the systems should be dictated by IAS and since the goal of RAS will be primarily to alert the drivers of the speed limit around a work zone. The layout for PAS should also be based on its transmission due to its comparatively limited range. We recommend buffer space for the system should be at least 100 ft with the sensor-to-alarm distance limited to 150 ft to ensure transmission and meet the MUTCD guidelines (Figure 5(d)). It is worth noting that this recommendation also satisfies our finding demonstrated in Figure 4(d) where a minimum time of at least 2 seconds is desirable for sensor-to-alarm distance of 100 ft since the system is recommended for use in facilities with operating speed less than 30 mph.

System selection

Based on the results from pilot testing and model analysis we recommend using IAS in construction work zones that require long term use of stationary traffic channelizers over long tapers. The system's transmission range allows it to be used in long tapers and therefore can used effectively in facilities where the posted speed limit is more than 30 mph. However, the time needed to set up each individual sensor makes it impractical for use in projects that require frequent repositioning. RAS is recommended for use in projects that require flagging. In our review of the literature, we could find no other systems that facilitate flagging operation and advanced intrusion detection. Further, it can be used in facilities with operating speed less than 40 mph. The 400 ft transmission range of the system makes it ideal for covering work zone perimeters with medium length tapers (Figure 5(c)). When flagging operation is needed on

facilities with speed limit greater than 40 mph, we recommend the system to be used alongside IAS to overcome the limitation imposed by its transmission range. When used with IAS, the system can be used primarily for enforcing speed limits while utilizing IAS for alerts. Finally, PAS, despite having a relative short transmission range, is easy to deploy. It is best suited for



(a) IAS (b) RAS deployed independently (c) RAS used with IAS (d) PAS

Figure 5 Recommended setup for work zone and system components

Table 10

Work zones taper and system deployment

Speed limit (mph)	Revised speed limit (mph)	Minimum taper length as per MUTCD, L (ft)	Recommended minimum buffer space, U (ft)
35	35	245	155
40	35	125	155
45	40	480	180
50	45	540	200
55	50	600	225
60	50	600	225
65	55	720	245
70	60	800	290

Table 11
Selection of WZIAS based on work zone types

System	Type of work	Taper length	Type of facility
IAS	i. Long term construction with stationary traffic channelizers	Long tapers > 150 ft	Speed limit >30 mph
RAS	i. Flagging operation ii. Short term mobile work zone requiring speed enforcement	Medium tapers < 400 ft	Used in conjunction with IAS in facilities with speed limit > 40 mph
PAS	i. Short term mobile Construction and maintenance work zones ii. Work zones with minor encroachment	Short Tapers < 150 ft	Posted speed limit < 30 mph

short term maintenance or mobile work zones and on facilities with speed limit less than 30 mph since the system's 150 ft range is adequate for work zones on facilities operating at less than 30 mph. The system is well suited for work zones on shoulders with little or no lane encroachment as shown in Figure 5(d). A summary of our recommendations is presented in Table 11.

Conclusion

This study employed non-parametric and semi-parametric survival analysis to investigate the influence of external variables associated with WZIAS on work zone crashes. The study used three WZIAS and subjected them to field tests wherein intrusions were imitated by driving test vehicles into a work zone with workers in a controlled setting. The activation of system alarms and worker reaction were then used to determine occurrence of crashes. The study contributed to the literature in the following manner.

Identification of WZIAS related external factors influencing work zone crashes

Previous studies evaluating WZIAS have focused entirely on their characteristics and performance. As per our knowledge, this is the first study that addresses the influence of external factors on the effectiveness of WZIAS. The manner of system deployment, more specifically the

layout of systems components and intrusion speed has not been accounted for by previous studies while evaluating system efficacy.

Our findings highlight the influence of intrusion speed, sensor-to-worker spacing, and system accuracy on occurrence of work zone crashes. We conclude that among all these factors intrusion speed and adequate spacing between the system sensors and workers is imperative to reducing crashes since appropriate measures pertaining to these factors can be adopted in the field. This can be achieved by reducing speed limits and standardizing the length of the buffer space to provide adequate separation.

Standardization of deployment strategies for systems

Although current literature recommends appropriate use cases for systems based on field experiments (Marks et al., 2017) specific recommendations that translate theoretical knowledge derived from field tests to standardized field practice is missing.

In this study we recommend appropriate use case scenarios for systems based on their transmission range and ease of installation. Additionally, we also present ideal deployment strategies for the system with revisions to existing MUTCD guidelines. Revisions recommended to existing guidelines include standards for buffer space and appropriate placement location of system components within a work zone.

References

2019 Highway Work Zone Safety Survey. (2019). Associated General Contractors of America. https://www.agc.org/news/2019/05/23/2019-highway-work-zone-safety-survey

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

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Abdel-Aty, M., & Pande, A. (2007). Crash data analysis: Collective vs. Individual crash level approach. *Journal of Safety Research*, *38*(5), 581–587. https://doi.org/10.1016/j.jsr.2007.04.007

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record: Journal of the Transportation Research Board*, *1897*(1), 88–95. https://doi.org/10.3141/1897-12

Afghari, A. P., Haque, M. M., & Washington, S. (2020). Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. *Accident Analysis & Prevention*, *144*, 105615. https://doi.org/10.1016/j.aap.2020.105615

Agent, K. R., & Hibbs, J. O. (1996). *Evaluation of SHRP Work Zone Safety Devices*. 24.

Ahmed, S. S., Cohen, J., & Anastasopoulos, P. Ch. (2021). A correlated random parameters with heterogeneity in means approach of deer-vehicle collisions and resulting injury-severities. *Analytic Methods in Accident Research*, *30*, 100160. https://doi.org/10.1016/j.amar.2021.100160

Al-Ghamdi, A. S. (2002). Pedestrian–vehicle crashes and analytical techniques for stratified contingency tables. *Accident Analysis & Prevention*, *34*(2), 205–214. https://doi.org/10.1016/S0001-4575(01)00015-X

Algomaiah, M., & Li, Z. (2022). Enhancing Work Zone Capacity by a Cooperative Late Merge System Using Decentralized and Centralized Control Strategies. *Journal of Transportation Engineering, Part A: Systems*, *148*(2). https://doi.org/10.1061/JTEPBS.0000632

Ali, Y., Haque, M. M., Zheng, Z., Washington, S., & Yildirimoglu, M. (2019). A hazard-based duration model to quantify the impact of connected driving environment on safety during mandatory lane-changing. *Transportation Research Part C: Emerging Technologies*, *106*(June), 113–131. https://doi.org/10.1016/j.trc.2019.07.015

Ariannezhad, A., Karimpour, A., Qin, X., Wu, Y.-J., & Salmani, Y. (2021). Handling Imbalanced Data for Real-Time Crash Prediction: Application of Boosting and Sampling Techniques. *Journal of Transportation Engineering, Part A: Systems*, *147*(3), 04020165. https://doi.org/10.1061/JTEPBS.0000499

Bagloee, S. A., & Asadi, M. (2016). Crash analysis at intersections in the CBD: A survival analysis model. *Transportation Research Part A: Policy and Practice*, *94*, 558–572. https://doi.org/10.1016/j.tra.2016.10.019

Barua, S., El-Basyouny, K., & Islam, Md. T. (2016). Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic Methods in Accident Research*, *9*, 1–15. https://doi.org/10.1016/j.amar.2015.11.002

Baruya, A. (1998). Road Safety in Europe. *9th Internation Conference: Road Safety in Europe*.

Bashir, S., & Zlatkovic, M. (2021). Assessment of Queue Warning Application on Signalized Intersections for Connected Freight Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(10), 1211–1221. https://doi.org/10.1177/03611981211015247

Benekohal, R. F., Hajbabaie, A., Medina, J. C., Wang, M.-H., & Chitturi, M. V. (2010). *SPEED PHOTO-RADAR ENFORCEMENT EVALUATION IN ILLINOIS WORK ZONES* (FHWA-ICT-10-064). Illinois Department of Transportation.

Berthaume, A. L. (2015). *Microscopic Modeling of Driver Behavior Based on Modifying Field Theory for Work Zone Application* [Doctoral Dissertation, University of Massachusetts Amherst]. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1328&context=dissertations_2

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2011). Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. *2011 World Congress on Information and Communication Technologies*, 1241–1246. https://doi.org/10.1109/WICT.2011.6141426

Bham, G. H., & Mohammadi, M. A. (2011). *Evaluation of Work Zone Speed Limits: An Objective and Subjective Analysis of Work Zones in Missouri Report*. 92.

Brownstone, D., & Small, K. A. (1989). Efficient Estimation of Nested Logit models. *Journal of Business & Economic Statistics*, *7*(1), 67–74. https://doi.org/10.1080/07350015.1989.10509714

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., & Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. *Transportation Research Part C: Emerging Technologies*, *117*, 102697. https://doi.org/10.1016/j.trc.2020.102697

Cerwick, D. M., Gkritza, K., Shaheed, M. S., & Hans, Z. (2014). A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, *3–4*, 11–27. https://doi.org/10.1016/j.amar.2014.09.002

Cestac, J., Paran, F., & Delhomme, P. (2011). Young drivers' sensation seeking, subjective norms, and perceived behavioral control and their roles in predicting speeding intention: How risk-taking motivations evolve with gender and driving experience. *Safety Science*, *49*(3), 424–432. https://doi.org/10.1016/j.ssci.2010.10.007

Chang, H. L., & Jovanis, P. P. (1990). Formulating accident occurrence as a survival process. *Accident Analysis and Prevention*, *22*(5), 407–419. https://doi.org/10.1016/0001-4575(90)90037-L

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, *43*(8), 541–557. https://doi.org/10.1016/j.ssci.2005.04.004

Chang, Y., & Edara, P. (2018). Predicting hazardous events in work zones using naturalistic driving data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-March*, 1–6. https://doi.org/10.1109/ITSC.2017.8317847

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, *90*, 128–139. https://doi.org/10.1016/j.aap.2016.02.011

Cheng, W., Gill, G. S., Dasu, R., Xie, M., Jia, X., & Zhou, J. (2017). Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis & Prevention*, *99*, 330–341. https://doi.org/10.1016/j.aap.2016.11.022

Cheng, Z., Lu, J., Zu, Z., & Li, Y. (2019). Speeding Violation Type Prediction Based on Decision Tree Method: A Case Study in Wujiang, China. *Journal of Advanced Transportation*, *2019*, 1–10. https://doi.org/10.1155/2019/8650845

Choudhary, P., & Velaga, N. R. (2020). Impact of distraction on decision making at the onset of yellow signal. *Transportation Research Part C: Emerging Technologies*, *118*(March 2019), 102741. https://doi.org/10.1016/j.trc.2020.102741

Chung, Y. (2010). Development of an accident duration prediction model on the Korean Freeway Systems. *Accident Analysis and Prevention*, *42*(1), 282–289. https://doi.org/10.1016/j.aap.2009.08.005

Data USA: Highway Maintenance Workers. (2018). https://datausa.io/profile/soc/highway-maintenance-workers

Debnath, A. K., Blackman, R., & Haworth, N. (2015). Common hazards and their mitigating measures in work zones: A qualitative study of worker perceptions. *Safety Science*, *72*, 293–301. https://doi.org/10.1016/j.ssci.2014.09.022

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113

Dissanayake, S., & Akepati, S. R. (2009). *Identification of Work Zone Crash Characteristics*. Federal Highway Administration. https://intrans.iastate.edu/app/uploads/2018/08/Dissanayake_WZCrashChar.pdf

Dissanayake, S., & Lu, J. (2002). Analysis of Severity of Young Driver Crashes: Sequential Binary Logistic Regression Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, *1784*(1), 108–114. https://doi.org/10.3141/1784-14

Dong, C., Clarke, D. B., Yan, X., Khattak, A., & Huang, B. (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate

crash frequencies at intersections. *Accident Analysis & Prevention*, *70*, 320–329. https://doi.org/10.1016/j.aap.2014.04.018

Elliott, M. A., & Thomson, J. A. (2010). The social cognitive determinants of offending drivers' speeding behaviour. *Accident Analysis & Prevention*, *42*(6), 1595–1605. https://doi.org/10.1016/j.aap.2010.03.018

Eseonu, C., Gambatese, J., & Nnaji, C. (2018). *Reducing Highway Fatalities Through Improved Adoption of Safety Technologies*.

Federal Highway Administration. (2009a). *Manual of Traffic Control Devices for Streets and Highways*.

Federal Highway Administration. (2009b). *Manual on Uniform Traffic Control Devices (MUTCD)*. https://mutcd.fhwa.dot.gov/

Federal Highway Administration. (2023). *FHWA Work Zone Facts and Statistics*. Work Zone Management Program. https://ops.fhwa.dot.gov/wz/resources/facts_stats.htm

Flannagan, C. A., Selpi, Baykas, P. B., Leslie, A., Kovaceva, J., & Thomson, R. (2019). *Analysis of SHRP2 Data to Understand Normal and Abnormal Driving Behavior in Work Zones* (FHWA-HRT-20-010). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/48835

Forward, S. E. (2009). The theory of planned behaviour: The role of descriptive norms and past behaviour in the prediction of drivers' intentions to violate. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 198–207. https://doi.org/10.1016/j.trf.2008.12.002

Fountas, G., & Anastasopoulos, P. Ch. (2017). A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. *Analytic Methods in Accident Research, 15*, 1–16. https://doi.org/10.1016/j.amar.2017.03.002

Furth, P. G. (2011). *Sampling and Estimation Techniques for Estimating Bus System Passenger-Miles*. Bureau of Transportation Statistics. https://www.bts.gov/archive/publications/journal_of_transportation_and_statistics/volume_08_number_02/paper_07/index

Furth, P. G., Killough, K. L., & Ruprecht, G. F. (1988). Cluster Sampling Techniques for Estimating Transit Patronage. *Transportation Research Record, 1165*.

Gambatese, J. A., Lee, H. W., & Nnaji, C. A. (2017). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance*. Oregon State University School of Civil and Construction Engineering.

Gambatese, J., & Lee, H. W. (2016). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance II*. (Issue 503).

Gan, H., Wei, J., & Wang, G. (2021). A generic work zone evaluation tool driven by a macroscopic traffic simulation model. *International Journal of Mobile Communications*, *19*(1), 1. https://doi.org/10.1504/IJMC.2021.111884

Garber, N. J., & Ehrhart, A. A. (2000). Effect of Speed, Flow, and Geometric Characteristics on Crash Frequency for Two-Lane Highways. *Transportation Research Record: Journal of the Transportation Research Board*, *1717*(1), 76–83. https://doi.org/10.3141/1717-10

Gelman, A., & Hill, J. (2007). When does a multilevel modeling make a difference? In *Data Analysis Using Regression and Multilevel/Hierarchical Models* (pp. 237–249). Cambridge University Press.

Golob, T. F., Recker, W. W., & Leonard, J. D. (1987). An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention*, *19*(5), 375–395. https://doi.org/10.1016/0001-4575(87)90023-6

Guo, H., Wang, W., Guo, W., & Zhao, F. (2013). Modeling lane-keeping behavior of bicyclists using survival analysis approach. *Discrete Dynamics in Nature and Society*, *2013*. https://doi.org/10.1155/2013/197518

Hamdar, S. H., Khoury, H., & Zehtabi, S. (2016). A simulator-based approach for modeling longitudinal driving behavior in construction work zones: Exploration and assessment. *SIMULATION*, *92*(6), 579–594. https://doi.org/10.1177/0037549716644515

Haque, K., Mishra, S., & Golias, M. M. (2021). Multi-period transportation network investment decision making and policy implications using econometric framework. *Research in Transportation Economics*, *89*, 101109. https://doi.org/10.1016/j.retrec.2021.101109

Haque, M. M., & Washington, S. (2015). The impact of mobile phone distraction on the braking behaviour of young drivers: A hazard-based duration model. *Transportation Research Part C: Emerging Technologies*, *50*, 13–27. https://doi.org/10.1016/j.trc.2014.07.011

Harb, R., Radwan, E., Yan, X., Pande, A., & Abdel-Aty, M. (2008). Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression. *Journal of Transportation Engineering*, *134*(5), 203–214. https://doi.org/10.1061/(ASCE)0733-947X(2008)134:5(203)

Harmon, T., Bahar, G., & Gross, F. (2018). *Crash Costs for Highway Safety Analysis*.

Hashmienejad, S. H. A., & Hasheminejad, S. M. H. (2017). Traffic accident severity prediction using a novel multi-objective genetic algorithm. *International Journal of Crashworthiness*, *22*(4), 425–440. https://doi.org/10.1080/13588265.2016.1275431

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second). John Wiley & Sons, Inc.

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019a). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019b). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

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

Hou, G., & Chen, S. (2019). An Improved Cellular Automaton Model for Work Zone Traffic Simulation Considering Realistic Driving Behavior. *Journal of the Physical Society of Japan*, *88*(8), 084001. https://doi.org/10.7566/JPSJ.88.084001

Hourdos, J. (2012). Portable, Non-Intrusive Advance Warning Devices for Work Zones with or without Flag Operators. *Minnesota Department of Transportation*, *October*.

Imprialou, M. I. M., Quddus, M., Pitfield, D. E., & Lord, D. (2016). Re-visiting crash-speed relationships: A new perspective in crash modelling. *Accident Analysis and Prevention*, *86*, 173–185. https://doi.org/10.1016/j.aap.2015.10.001

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Jonathan, A.-V., Wu, K.-F. (Ken), & Donnell, E. T. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention*, *87*, 8–16. https://doi.org/10.1016/j.aap.2015.11.006

Jovanis, P. P., & Chang, H. L. (1989). Disaggregate model of highway accident occurrence using survival theory. *Accident Analysis and Prevention*, *21*(5), 445–458. https://doi.org/10.1016/0001-4575(89)90005-5

Jovanović, D., Šraml, M., Matović, B., & Mićić, S. (2017). An examination of the construct and predictive validity of the self-reported speeding behavior model. *Accident Analysis & Prevention*, *99*, 66–76. https://doi.org/10.1016/j.aap.2016.11.015

Jung, S., Qin, X., & Noyce, D. A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis & Prevention*, *42*(1), 213–224. https://doi.org/10.1016/j.aap.2009.07.020

Kashyap, A. A., Raviraj, S., Devarakonda, A., Nayak K, S. R., K V, S., & Bhat, S. J. (2022). Traffic flow prediction models – A review of deep learning techniques. *Cogent Engineering*, *9*(1), 2010510. https://doi.org/10.1080/23311916.2021.2010510

Ke, J., Zhang, S., Yang, H., & Chen, X. (Michael). (2019). PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. *Transportmetrica A: Transport Science*, *15*(2), 872–895. https://doi.org/10.1080/23249935.2018.1542414

Keramati, A., Lu, P., Zhou, X., & Tolliver, D. (2020). A Simultaneous Safety Analysis of Crash Frequency and Severity for Highway-Rail Grade Crossings: The Competing Risks Method. *Journal of Advanced Transportation*, *2020*(1). https://doi.org/10.1155/2020/8878911

Khasnabis, S., Mishra, S., & Safi, C. (2012). Evaluation procedure for mutually exclusive highway safety alternatives under different policy objectives. *Journal of Transportation Engineering*, *138*(7), 940–948. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000397

Khattak, A. J., Khattak, A. J., & Council, F. M. (2002). Effects of work zone presence on injury and non-injury crashes. *Accident Analysis and Prevention*, *34*(1), 19–29. https://doi.org/10.1016/S0001-4575(00)00099-3

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, & W. Wong, Eds.; Thid Editi). Springer. https://doi.org/10.1007/978-1-4419-6646-9

Kloeden, C. N., McLean, J., & Glonek, G. F. V. (2002). *Reanalysis of travelling speed and the risk of crash involvement in Adelaide South Australia*. Australian Transport Safety Bureau.

Kock, N., & Lynn, G. S. (2012). Lateral Collinearity and Misleading Results in Variance-Based SEM : An Illustration and Recommendations Lateral Collinearity and Misleading Results in Variance-. *Journal of the Association for Information Systems*, *13*(7), 546–580.

Lee, C., Hellinga, B., & Saccomanno, F. (2003). Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, *1840*(1), 67–77. https://doi.org/10.3141/1840-08

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *Applied Sciences*, *10*(1), 129. https://doi.org/10.3390/app10010129

Lee, J.-T., & Fazio, J. (2005). Influential Factors in Freeway Crash Response and Clearance Times by Emergency Management Services in Peak Periods. *Traffic Injury Prevention*, *6*(4), 331–339. https://doi.org/10.1080/15389580500255773

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis & Prevention*, *135*, 105371. https://doi.org/10.1016/j.aap.2019.105371

Li, Y., & Bai, Y. (2008). Development of crash-severity-index models for the measurement of work zone risk levels. *Accident Analysis and Prevention*, *40*(5), 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012

Li, Y., & Bai, Y. (2009). Highway work zone risk factors and their impact on crash severity. *Journal of Transportation Engineering*, *135*(10), 694–701. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000055

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident Analysis and Prevention*, *111*(November 2017), 354–363. https://doi.org/10.1016/j.aap.2017.11.028

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

Ma, J., & Kockelman, K. (2006a). Crash frequency and severity modeling using clustered data from Washington state. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, October*, 1621–1626. https://doi.org/10.1109/itsc.2006.1707456

Ma, J., & Kockelman, K. M. (2006b). Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record: Journal of the Transportation Research Board*, *1950*, 24–34.

Ma, J., Kockelman, K. M., & Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*, *40*(3), 964–975. https://doi.org/10.1016/j.aap.2007.11.002

Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research*, *25*, 100113. https://doi.org/10.1016/j.amar.2020.100113

Marks, E., Vereen, S., & Awolusi, I. (2017). *Active Work Zone Safety Using Emerging Technologies 2017*. University Transportation Center for Alabama The University of Alabama.

Martin, J., Rozas, A., & Araujo, A. (2016). A WSN-Based Intrusion Alarm System to Improve Safety in Road Work Zones. *Journal of Sensors*, *2016*, 1–8. https://doi.org/10.1155/2016/7048141

Medina-Salgado, B., Sánchez-DelaCruz, E., Pozos-Parra, P., & Sierra, J. E. (2022). Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, *35*, 100739. https://doi.org/10.1016/j.suscom.2022.100739

Meng, Q., & Weng, J. (2011). A Genetic algorithm approach to assessing work zone casualty risk. *Safety Science*, *49*(8–9), 1283–1288. https://doi.org/10.1016/j.ssci.2011.05.001

Mishra, S. (2013). A Synchronized Model for Crash Prediction and Resource Allocation to Prioritize Highway Safety Improvement Projects. *Procedia - Social and Behavioral Sciences*, *104*, 992–1001. https://doi.org/10.1016/j.sbspro.2013.11.194

Mishra, S., Golias, M. M., Sharma, S., & Boyles, S. D. (2015). Optimal funding allocation strategies for safety improvements on urban intersections. *Transportation Research Part A: Policy and Practice*, *75*, 113–133. https://doi.org/10.1016/j.tra.2015.03.001

Mishra, S., Golias, M. M., & Thapa, D. (2021). *Work Zone Alert Systems*. Tennessee Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56274

Mohan, D., Bangdiwala, S. I., & Villaveces, A. (2017). Urban street structure and traffic safety. *Journal of Safety Research*, *62*, 63–71. https://doi.org/10.1016/j.jsr.2017.06.003

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis. *Transportation Research Record*, *2673*(11), 680–692. https://doi.org/10.1177/0361198119845899

Nam, D., & Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, *34*(2), 85–102. https://doi.org/10.1016/S0965-8564(98)00065-2

Novosel, C. (2014). Evaluation of Advanced Safety Perimeter Systems for Kansas Temporary Work Zones. In *Civil, Environmental, and Architectural Engineering, University of Kansas*.

Osman, M., Mishra, S., & Paleti, R. (2018a). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group

differences. *Accident Analysis and Prevention*, *118*(May), 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018b). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018c). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., Paleti, R., & Golias, M. (2019). Impacts of Work Zone Component Areas on Driver Injury Severity. *Journal of Transportation Engineering, Part A: Systems*, *145*(8), 04019032. https://doi.org/10.1061/jtepbs.0000253

Osman, M., Paleti, R., & Mishra, S. (2018a). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis and Prevention*, *111*(May 2017), 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., & Mishra, S. (2018b). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis & Prevention*, *111*, 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis and Prevention*, *97*, 261–273. https://doi.org/10.1016/j.aap.2016.10.020

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051–10072. https://doi.org/10.1007/s12652-020-02759-5

Ozturk, O., Ozbay, K., Yang, H., & Bartin, B. (2013). Crash Frequency Modeling for Highway Construction Zones. *Transportation Research Board's 92nd Annual Meeting, Washington, D.C*, 14p.

Paleti, R., Mahmud, A., Gayah, V., & Pinjari, A. (2021). When and Where does the Next Traffic Crash Occur? A Discretized Duration Based Modeling Approach. *Under Review for Publication*.

Park, E. S., & Lord, D. (2007). Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record*, *2019*, 1–6. https://doi.org/10.3141/2019-01

Pei, X., Wong, S. C., & Sze, N. N. (2011). A joint-probability approach to crash prediction models. *Accident Analysis & Prevention*, *43*(3), 1160–1166. https://doi.org/10.1016/j.aap.2010.12.026

Pham, M.-H., Bhaskar, A., Chung, E., & Dumont, A.-G. (2010). Random forest models for identifying motorway Rear-End Crash Risks using disaggregate data. *13th International IEEE Conference on Intelligent Transportation Systems*, 468–473. https://doi.org/10.1109/ITSC.2010.5625003

Provost, F., Jensen, D., & Oates, T. (2001). Progressive Sampling. In *Instance Selection and Construction for Data Mining* (pp. 151–170). Springer US. https://doi.org/10.1007/978-1-4757-3359-4_9

Qi, Y., Srinivasan, R., Teng, H., & Baker, R. F. (2005). *Frequency of Work Zone Accidents on Construction Projects*.

Quddus, M. (2013). Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, *5*(1), 27–45. https://doi.org/10.1080/19439962.2012.705232

Rahim, M. A., & Hassan, H. M. (2021). A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention*, *154*, 106090. https://doi.org/10.1016/j.aap.2021.106090

Rahman, R., Bhowmik, T., Eluru, N., & Hasan, S. (2021). Assessing the crash risks of evacuation: A matched case-control approach applied over data collected during Hurricane Irma. *Accident Analysis & Prevention*, *159*, 106260. https://doi.org/10.1016/j.aap.2021.106260

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007

Sarker, A. A., Naimi, A., Mishra, S., Golias, M. M., & Freeze, P. B. (2015). Development of a Secondary Crash Identification Algorithm and occurrence pattern determination in large scale multi-facility transportation network. *Transportation Research Part C: Emerging Technologies*, *60*, 142–160. https://doi.org/10.1016/j.trc.2015.08.011

Scott-Parker, B., Hyde, M. K., Watson, B., & King, M. J. (2013). Speeding by young novice drivers: What can personal characteristics and psychosocial theory add to our understanding? *Accident Analysis & Prevention*, *50*, 242–250. https://doi.org/10.1016/j.aap.2012.04.010

Shangguan, Q., Fu, T., & Liu, S. (2020). Investigating rear-end collision avoidance behavior under varied foggy weather conditions: A study using advanced driving simulator and survival analysis. *Accident Analysis and Prevention*, *139*(March), 105499. https://doi.org/10.1016/j.aap.2020.105499

Sharma, A., Bullock, D., & Peeta, S. (2011). Estimating dilemma zone hazard function at high speed isolated intersection. *Transportation Research Part C: Emerging Technologies*, *19*(3), 400–412. https://doi.org/10.1016/j.trc.2010.05.002

Simons-Morton, B. G., Ouimet, M. C., Chen, R., Klauer, S. G., Lee, S. E., Wang, J., & Dingus, T. A. (2012). Peer influence predicts speeding prevalence among teenage drivers. *Journal of Safety Research*, *43*(5–6), 397–403. https://doi.org/10.1016/j.jsr.2012.10.002

Song, J. J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, *97*(1), 246–273. https://doi.org/10.1016/j.jmva.2005.03.007

Stipancic, J., Miranda-Moreno, L., Saunier, N., & Labbe, A. (2019). Network screening for large urban road networks: Using GPS data and surrogate measures to model crash frequency and severity. *Accident Analysis and Prevention*, *125*(February), 290–301. https://doi.org/10.1016/j.aap.2019.02.016

Stout, D., Graham, J., Bryant-Fields, B., Migletz, J., Fish, J., & Hanscom, F. (1993). *Maintenance Work Zone Safety Devices Development and Evaluation*.

Sun, J., & Sun, J. (2016a). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Sun, J., & Sun, J. (2016b). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y., & Huang, H. (2020). Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Analytic Methods in Accident Research*, *27*, 100123. https://doi.org/10.1016/j.amar.2020.100123

Thapa, D., & Mishra, S. (2021a). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., & Mishra, S. (2021b). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*, 106125. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., Paleti, R., & Mishra, S. (2022a). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Thapa, D., Paleti, R., & Mishra, S. (2022b). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Theiss, L., Ullman, G. L., & Lindheimer, T. (2017). *Closed Course Performance Testing of the Aware Intrusion Alarm System*.

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2019). Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention*, *130*, 151–159. https://doi.org/10.1016/j.aap.2017.12.018

Therneau, T. M. (2020). *A Package for Survival Analysis in R. R package version 3.2-7*.

Therneau, T. M., Grambsch, P. M., & Panktatz, S. V. (2003). Penalized Survival Models and Frailty. *Journal of Computational and Penalized Survival Models and Frailty*, *12*(1), 156–175.

Ullman, G. L., Trout, N. D., & Theiss, L. (2016). *Driver Responses to the AWARE Intrusion Alarm System*. Texas A&M Transportation Institute.

Venugopal, S., & Tarko, A. (2000). Safety models for rural freeway work zones. *Transportation Research Record*, *1715*, 1–9. https://doi.org/10.3141/1715-01

Wang, C., Quddus, M. A., & Ison, S. G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention*, *43*(6), 1979–1990. https://doi.org/10.1016/j.aap.2011.05.016

Wang, J., Yamamoto, T., & Liu, K. (2020). Key determinants and heterogeneous frailties in passenger loyalty toward customized buses: An empirical investigation of the subscription termination hazard of users. *Transportation Research Part C: Emerging Technologies*, *115*(July 2019), 102636. https://doi.org/10.1016/j.trc.2020.102636

Wang, X., Katz, R., & Dong, X. S. (2018). *Fatal Injuries at Road Construction Sites among Construction Workers* [Quarterly]. Center for Construction Research and Training.

https://www.cpwr.com/wp-content/uploads/publications/publications_Quarter2-QDR-2018.pdf

Work Zones-Injury Facts-National Safety Council. (2020). https://injuryfacts.nsc.org/motor-vehicle/motor-vehicle-safety-issues/work-zones/

Wu, L., Meng, Y., Kong, X., & Zou, Y. (2020). Incorporating survival analysis into the safety effectiveness evaluation of treatments: Jointly modeling crash counts and time intervals between crashes. *Journal of Transportation Safety and Security*, *0*(0), 1–21. https://doi.org/10.1080/19439962.2020.1786871

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

Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., & Jiang, X. (2020). A machine-learning method for improving crash injury severity analysis: A case study of work zone crashes in Cairo, Egypt. *International Journal of Injury Control and Safety Promotion*, *27*(3), 266–275. https://doi.org/10.1080/17457300.2020.1746814

Yang, H., Ozbay, K., Ozturk, O., & Xie, K. (2015). Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. *Traffic Injury Prevention*, *16*(4), 387–396. https://doi.org/10.1080/15389588.2014.948615

Yasmin, S., & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention*, *59*, 506–521. https://doi.org/10.1016/j.aap.2013.06.040

Yasmin, S., & Eluru, N. (2018). A joint econometric framework for modeling crash counts by severity. *Transportmetrica A: Transport Science*, *14*(3), 230–255. https://doi.org/10.1080/23249935.2017.1369469

Yasmin, S., Eluru, N., Bhat, C. R., & Tay, R. (2014). A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic Methods in Accident Research*, *1*, 23–38. https://doi.org/10.1016/j.amar.2013.10.002

Yasmin, S., Eluru, N., Wang, L., & Abdel-Aty, M. A. (2018). A joint framework for static and real-time crash risk analysis. *Analytic Methods in Accident Research*, *18*, 45–56. https://doi.org/10.1016/j.amar.2018.04.001

Ye, X., Pendyala, R. M., Shankar, V., & Konduri, K. C. (2013). A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention*, *57*, 140–149. https://doi.org/10.1016/j.aap.2013.03.025

Yu, B., Chen, Y., & Bao, S. (2019). Quantifying visual road environment to establish a speeding prediction model: An examination using naturalistic driving data. *Accident Analysis & Prevention*, *129*, 289–298. https://doi.org/10.1016/j.aap.2019.05.011

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, *51*, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis and Prevention*, *73*, 351–358. https://doi.org/10.1016/j.aap.2014.09.006

Zhang, C., He, J., Wang, Y., Yan, X., Zhang, C., Chen, Y., Liu, Z., & Zhou, B. (2020). A Crash Severity Prediction Method Based on Improved Neural Network and Factor Analysis. *Discrete Dynamics in Nature and Society*. https://doi.org/10.1155/2020/4013185

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods. *IEEE Access*, *6*, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*(1), 204–215. https://doi.org/10.1016/j.ijforecast.2010.06.002

Zhang, K., & Hassan, M. (2019a). Crash severity analysis of nighttime and daytime highway work zone crashes. *PLoS ONE*, *14*(8), 1–17. https://doi.org/10.1371/journal.pone.0221128

Zhang, K., & Hassan, M. (2019b). Identifying the Factors Contributing to Injury Severity in Work Zone Rear-End Crashes. *Journal of Advanced Transportation*, *2019*, 1–9. https://doi.org/10.1155/2019/4126102

Zhao, G., Wu, C., & Qiao, C. (2013). A Mathematical Model for the Prediction of Speeding with its Validation. *IEEE Transactions on Intelligent Transportation Systems*, *14*(2), 828–836. https://doi.org/10.1109/TITS.2013.2257757

Zheng, L., & Sayed, T. (2020). A novel approach for real time crash prediction at signalized intersections. *Transportation Research Part C: Emerging Technologies*, *117*, 102683. https://doi.org/10.1016/j.trc.2020.102683

Zimmerman, K., Mzige, A. A., Kibatala, P. L., Museru, L. M., & Guerrero, A. (2012). Road traffic injury incidence and crash characteristics in Dar es Salaam: A population based

study. *Accident Analysis & Prevention*, *45*, 204–210.

https://doi.org/10.1016/j.aap.2011.06.018

3. Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches

Introduction

Traffic engineers utilize Safety Performance Functions (SPFs) to identify the causal factors of crashes and develop strategies to make highways safer. SPFs are developed using regression techniques and can be of two types: i) static risk models and ii) dynamic risk models depending on the type of data they are trained on (Yasmin et al., 2018). Static risk models regress the frequency of crashes at a particular location as a function of predictors such as traffic flow, highway geometry, environmental conditions, driver, and vehicle characteristics (e.g., age, gender, vehicle body type, etc.) using aggregated traffic and crash data collected from the location over an extended period (e.g., see (Imprialou et al., 2016; Ma et al., 2008; Ma & Kockelman, 2006a)). Although useful for safety interventions, there are two main limitations to static risk models. First, aggregated data used to train these models fail to capture variations in the predictor variable which can result in erroneous inferences of the relationship between crash and its predictor. For example, traffic flow can vary considerably across a day, month, and year. Aggregated metrics such as Average Annual Daily Traffic, when used as a precursor to crashes, cannot capture the variation in traffic flow conditions that influences crashes. Second, static risk models are only useful for reactive traffic management and lends little to no utility towards proactive traffic management. This necessitates the use of dynamic risk models.

Dynamic risk models or real-time crash prediction models use real-time disaggregated traffic data to eliminate the limitations of static risk models. Contrary to its static counterpart, dynamic risk models evaluate the probability of crash based on observed associations between traffic and crash occurrences, thus enabling unsafe traffic conditions to be used as predictors of

future crashes (Hossain et al., 2019a; C. Lee et al., 2003; Theofilatos et al., 2019). This enables traffic engineers and planners to monitor real-time traffic flow dynamics, identify unsafe conditions, make predictions, and take proactive measures in real-time to fix traffic flow conditions and avoid unfavorable outcomes such as crashes and congestion.

With the advent of advanced data-driven techniques such as data mining and Machine Learning (ML), researchers no longer need to depend on regression techniques to develop SPFs, and rightfully so. ML algorithms have been found to offer better predictive accuracy compared to traditional SPFs (Ariannezhad et al., 2021; Cai et al., 2020). However, data-driven techniques are limited by their transportability (i.e., the ability to generalize causal relationships in scenarios other than what the algorithm is trained on) and ability to provide insight into causal inferences (Mannering et al., 2020). Therefore, SPFs are still desired when the extraction of causal inferences is of interest. In particular, regression equations grant the ability to draw interpretable results through variable coefficients and marginal effects which is desired when generating crash forecasts under different policy scenarios.

Literature review

Statistical approaches in real-time crash modeling

Regression-based real-time crash prediction models in the literature are largely based on a case-control approach where crashes are matched with non-crash events and crash likelihood is modeled using the binary logit framework. For example, crashes can be assumed to have binary outcomes (crash=1, no crash=0) but crash events are seldom encountered compared to non-crash events. Therefore, every crash is matched with multiple non-crash events based on common variables such as location, time and day, real-time traffic flow, etc. which serve as controls (M. Abdel-Aty & Pande, 2007; Rahman et al., 2021; Sun & Sun, 2016a). However, an issue that

arises with this approach is the ratio of events to its controls usually is not representative of the actual phenomenon and the optimal ratio often varies from one study to another (Yasmin et al., 2018). Recently, a "sampling of alternatives" approach was proposed by researchers to overcome this shortcoming (Yasmin et al., 2018). In this approach, the analysis window (1 month) was divided into 5-minute intervals. A crash could occur on any of these time intervals, therefore allowing the time intervals to be modeled as alternatives in a Multinomial Logit (MNL) model. To avoid computational issues arising from a large number of alternatives, the researchers used a sampling approach where 30 alternatives were drawn which included the time interval where the crash was observed and 29 other random intervals. Using this approach, the authors developed a joint model by integrating monthly crash risk and real-time crash risk that could predict crashes for 5-minute time intervals in the next month. While it is appealing to forecast potential crashes far into the future, a future window of 1 month considered in the study is impractical. There can be considerable changes in traffic flow and weather conditions over longer periods which makes predictions with longer windows unrealistic and unreliable.

Another notable approach in the prediction of crash likelihood is the use of duration models, also called hazard or survival models. These models describe the conditional probability of an event (in this case a crash) occurring at a certain time *t* provided it has not occurred until then. Hazard models in their most basic form assume a constant effect of a predictor variable on the outcome over time which is referred to as constant hazard rate. The framework was first implemented to analyze crash data in 1990 (H. L. Chang & Jovanis, 1990). It has since been extended to model crashes on highway intersections (e.g., (Bagloee & Asadi, 2016)**)**, predict clearance time for roadway incidents (e.g., (Nam & Mannering, 2000)), predict crashes on highway work zones (e.g., (Thapa & Mishra, 2021a)**)**. Its application on real-time crash

prediction, however, is not straightforward because of its inability to incorporate time-varying covariates. However, if the duration between crashes is divided into equal-time intervals and modeled as outcomes in a choice model, time-varying covariates can be introduced in duration models as interval-specific variables (Paleti et al., 2021).

Table 12
Types of sampling techniques with examples

Technique	Study	Description
Random	(Zimmerman et al., 2012)*	Random selection of GPS points from a geographical cluster within a city to select individuals and households to be interviewed on traffic crashes
Interval	(Al-Ghamdi, 2002)	Investigated the determinants of pedestrian crashes by selecting every third record from a list of pedestrian crashes.
Stratified	(Theofilatos et al., 2019)	Case-control approach with samples drawn from non-crash events to estimate the effect of real-time traffic characteristics on crash occurrence
	(Harb et al., 2008)	Stratified sampling based on number of lanes, speed limit and time of day (a.m. or p.m.) to compare work zone and non-work zone crashes (necessitated by varying traffic flow conditions at these crash locations) using conditional logistic regression
	(Mohan et al., 2017)	Compared cities stratified by traffic fatality rates to investigate the effect of road type and junction density on crashes
Cluster	(Furth, 2011; Furth et al., 1988)	Sampling of trips (e.g., of trips-round trips or chain of trips on a single route) to estimate true transit patronage
	(Zimmerman et al., 2012)*	Geographical clustering of the study area within cities based on their GPS coordinates to identify clusters of individuals and households who could be interviewed on traffic crashes

*Note: (Zimmerman et al., 2012) use two sampling techniques in their study. The first is geographical clustering to divide cities into grids. Random GPS coordinates were then drawn from the grid to select individuals and households for their survey and interviews.

Sampling approaches

Developing regression models using large crash data can be computationally demanding. In this regard, probability sampling techniques can help draw representative samples that can provide accurate statistical inferences with a reduced computational load and time. Probability sampling techniques draw samples from the original data at random, therefore, assigning every observation an equal probability of being selected. Since the current study necessitates the use of sampling

techniques, we briefly describe the probability sampling techniques as they are relevant to this study.

Probability sampling can be done in one of four following ways: (i) simple random sampling by drawing samples at random; (ii) systematic or interval sampling by drawing a random sample at first and samples after a certain interval then after (every n^{th} sample); (iii) stratified sampling by drawing a sample from preidentified groups or strata in the data; and (iv) cluster sampling by identifying and selecting natural clusters within the data. The approaches in sampling have their distinct advantages and disadvantages. Simple and systemic approaches are the easiest to perform. However, for imbalanced data (where a particular observation is rare while others are abundant) simple random sampling can lead to biases. Interval sampling can also lead to biases particularly when the data being drawn from has a periodical or repetitive structure such as those observed in panel and cross-sectional data. The approach also requires a complete list of observations before being administered. Stratified and cluster sampling, although immune to biases that plague simple random and interval sampling, are particularly complicated to work with. In stratified sampling, having prior knowledge of the data properties is necessary to identify groups and strata in the data. While this can be difficult, it is particularly useful in the case of imbalanced data as samples can be drawn from groups of frequent observations to match rare observations (e.g., draw only from non-crash events to match it with crash events). Although alternatively, synthetic data generation techniques can also be used to create synthetic data points and eliminate the need for sampling (e.g., see (Ariannezhad et al., 2021; Cai et al., 2020; Ke et al., 2019). Similarly, in cluster sampling, it is challenging to determine the number of clusters required to represent large data, and this gives rise to statistical uncertainty. Table 12 describes

selected studies based on sampling techniques. The list of studies in the table is not meant to be an exhaustive one and is provided here to introduce them to the reader.

Role of traffic flow parameters

Traffic flow parameters are established precursors of highway crashes. Hence, their inclusion in crash prediction models is considered a requisite. Despite this, their specific influence on crashes has been disparate across studies. Specifically, investigations on the influence of speed on crash occurrence have reported contesting results. Although a direct relationship between speed and crash severity has been reported consistently (e.g., (Osman, Mishra, et al., 2018a; Osman, Paleti, et al., 2018a), the speed-crash frequency relationship remains questionable. The relationship between them varies across the literature from negative (Baruya, 1998), insignificant (Ma et al., 2008; Quddus, 2013), to positive (Kloeden et al., 2002). Concerning the influence of traffic volumes on the speed-crash relationship, researchers have emphasized the inclusion of real-time traffic volumes and vehicle occupancy (Aarts & van Schagen, 2006; Garber & Ehrhart, 2000) since both higher and lower traffic volumes have been reported as determinants of crashes which has been attributed to the speed variance associated with the change in traffic volumes (M. A. Abdel-Aty & Radwan, 2000; L.-Y. Chang, 2005; Garber & Ehrhart, 2000).

Research objectives

As mentioned before, the use of duration models in real-time crash prediction models is non-existent except for (Paleti et al., 2021). In the study, the authors demonstrate their conceptual framework using crash data from I-405 assuming homogenous 5-mile highway segments. Their study uses a five-hour epoch with one-hour time intervals as the future window for crash prediction. The current study leverages potential scopes for improving the original paper while also further validating the model.

The assumption of homogenous segments, that is, traffic and highway geometry do not vary across a link, is done to simplify the modeling approach. In reality, traffic and highway geometry vary significantly within links (Imprialou et al., 2016). Therefore, additional analysis with appropriate highway segmentation and multiple interstates is necessitated to further validate and showcase the framework's applicability.

The framework requires a reformulation of the crash data (i.e., discretization of inter-crash duration into small time intervals) to fit into its framework. This results in a considerable increase in data size (please refer to the *Model Framework* section for a detailed description of data reformulation and how it increases in data size). The issues of data size become even more pronounced when a smaller time discretization is used. While the authors have validated the model using a future window of five hours with 1-hour intervals, a smaller time discretization is more desirable and practical. Therefore, the issue of data size becomes inevitable and needs to be addressed.

To address these limitations, current research aims to answer the question, "*If the discretized duration-based crash modeling framework is to be used, utilization of the completely reformulated dataset becomes too exhaustive and computationally demanding. In such a case, what sampling technique(s) and sample size(s) provides the most time-efficient and consistent estimates?".* In doing so, this research further validates and improves upon the original framework by i) utilizing crash data from multiple interstates segmented by geometry (number of lanes), terrain (flat or rolling), and traffic flow (posted speed limit), ii) using a smaller and more realistic future window of 1 hour further discretized into 15-minute intervals for prediction, and iii) investigating sampling techniques and sample size required at different sampling levels (more

on this in the *Modeling Approach* section) to derive consistent estimates while also reducing the computational complexity arising from large data size.

The rest of the paper is outlined as follows. We present a detailed description of the framework and the data reformulation process from the original paper in the *Model framework* section. The study area and data used in the study are then discussed in the *Data* section. The section: *Sampling approach* expands on the approach taken in sampling crashes and estimating models. The *Results* section discusses the results from our models followed by the *Discussion* section where we examine our findings and its implications. The *Conclusion* section summarizes the findings from this study and presents avenues for future research.

Model framework

The hazard function *h(t)* in a duration-based crash model gives the probability of a crash occurring at a certain time *t* conditional upon no crashes having occurred until then. Assuming a constant hazard rate over time (*h*), the hazard function can be written as.

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{he^{-ht}}{1 - (1 - e^{-ht})} = h \tag{8}$$

In equation 8, *f(t)* is the probability distribution function and *F(t)* is the probability density function for a continuous random variable *T*. As such, *F(t)* gives the probability that a crash has occurred by *t* and it can be expressed using equation 9.

$$F(t) = Pr(T \leq t) \tag{9}$$

Assuming time to be discretized into *n* intervals each with a duration of *dt*, the probability of a crash occurring at a certain interval *n* since the beginning can be written and simplified as follows.

$$Pr(T = ndt) = Pr(T \leq ndt) - Pr(T \leq (n-1)dt) \tag{10}$$

$$= F(ndt) - F\big((n-1)dt\big)$$

$$= exp\,(-h(n-1)dt) - exp\,(-hndt)$$

$$= \frac{exp\,(-h(n-1)dt)}{1/(1 - exp\,(-hdt))}$$

The denominator in equation 10 can be written using a Taylor series, i.e., $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots \infty$ for -1 < *x* < 1 to obtain equation 4.

$$Pr(T = ndt) = \frac{exp(-h(n-1)dt)}{1 + exp(-hdt) + exp(-2hdt) + exp(-3hdt) + \cdots\ \infty} \tag{11}$$

$$= \frac{exp(U_n)}{exp(U_1) + exp(U_2) + exp(U_3) + \cdots \infty}, \text{where } U_n = -h(n-1)dt$$

$$= \frac{exp\,(U_n)}{\sum_1^\infty exp\,(U_c)}$$

After simplification of equation 11, we see that the occurrence of the crash over the n^{th} interval takes the form of a Multinomial Logit (MNL) model with infinite alternatives for *n* (i.e., $n \rightarrow \infty$). The latent propensity associated with each alternative (U_n) is associated with the hazard rate (*h*), number of alternatives (*n*), and time interval (*dt*) as $U_n = -h(n-1)dt$. This relationship allows the introduction of time-varying covariates and non-linear hazard rates into an MNL framework through the latent propensity function which is a linear combination of various degree polynomials shown in equation 12.

$$U_n = \beta_1(n-1)dt + \beta_2[(n-1)dt]^2 + \beta_2[(n-1)dt]^3 + \cdots \tag{12}$$

When higher-order polynomial terms are zero, the MNL becomes a simple exponential model. The methodological framework used to modify crash data to the discretized modeling framework is discussed in the preceding paragraphs.

For empirical application in this study, we assume that each epoch is one hour, and it is further discretized into four 15-minute intervals, i.e., *dt* =0.25 hours. We use the indexes *e* and *s*

to represent epochs and segments henceforth. Therefore, the number of intervals in each epoch is 4, which we represent using the index, *C*, i.e., *C*=4. The number of epochs is dependent on the time interval between two consecutive crashes in a segment. For instance, let us assume for roadway segment 1, a second crash is observed 3.5 hours after the first. The inter-crash duration of 3.5 hours is discretized into four epochs wherein the crash occurs at the second 15-minute interval of the fourth epoch as shown in Table 13. Additionally, the column "Beyond 1 hour of the current epoch" provides information on whether the second crash occurred in a certain epoch or the future. In the presented example, the value 1 in the column for the first three epochs communicates that the second crash occurred in the future. Conversely, its value is 0 for the fourth epoch suggesting the crash occurred in the current (fourth) epoch.

If *i* be the index representing the 15-minute intervals. Then, the time since the beginning of interval *i* in epoch *e* can be written as $t_{e,i} = (e-1)Cdt + (i-1)dt$. For the example presented in Table 13, the time for the second interval of the fourth epoch is therefore $t_{4,2} = (4-1)1 + (2-1)0.25 = 3.25$. The latent propensity function for each time interval, *i* = (1,2,3,4) of an epoch can then be written using equation 13.

Table 13
Reformulation of crash data to create forecasting epochs

Segment	Time until next crash (hours)	Epoch	First 15-min	Second 15-min	Third 15-min	Fourth 15-min	Beyond 1 hour of the current epoch
1	3.5	1	0	0	0	0	1
1	3.5	2	0	0	0	0	1
1	3.5	3	0	0	0	0	1
1	3.5	4	0	1	0	0	0

$$U_{s,e,i} = \beta_1 t_{e,i} + \beta_2 [t_{e,i}]^2 + \cdots + r' X_{s,e,i} \quad (13)$$

As shown by equation 13, each time interval can be considered an alternative in the MNL model. The first two terms on the right-hand side of the latent propensity function introduce the effect of duration dynamics. The final term adds the effect of time-varying covariates into the propensity function. In addition to the four alternatives, there is one additional C+1th alternative that indicates whether a crash occurred in the future (last column in Table 13). The propensity function for this alternative can be written as.

$$U_{s,e,C+1} = \beta_{C+1} \tag{14}$$

The conditional probability of the next crash occurring in time-interval *i* of epoch *e* for a segment *s* and random variable of time T_s, provided no crash has occurred until the previous epoch, can be written using equation 15.

$$Pr\left(T_s = t_{e,i} \middle| T_s > (e-1)Cdt\right) = \frac{exp\ (U_{s,e,i})}{\sum_{c=1}^{C} exp\ (U_{s,e,c}) + exp\ (U_{s,e,C+1})} \tag{15}$$

The unconditional probability of a crash occurring at interval *i* of epoch *e* can then be calculated using equation 16.

$$Pr(T_s = t_{e,i}) = \frac{exp\ (U_{s,e,i})}{\sum_{c=1}^{C} exp\ (U_{s,e,c}) + exp\ (U_{s,e,C+1})} \cdot \prod_{e*=1}^{e-1} \frac{exp\ (U_{s,e*,C+1})}{\sum_{c=1}^{C} exp\ (U_{s,e*,c}) + exp\ (U_{s,e*,C+1})} \tag{16}$$

The parameters in the MNL model can be estimated by maximizing the product of the unconditional likelihood function in Equation 16. This is done by defining a vector of parameters,$n = (\beta_1, \beta_2, \dots, r, \beta_{C+1})'$. The vector is obtained by vertical concatenation of parameters in the latent propensity function.

Data

One of the objectives of the study was to include multiple interstates in model estimation. Therefore, crashes from two primary interstates within Memphis city limits in Tennessee were used for the study. Crash data for the year 2019 was collected from Enhanced Tennessee Roadway Information Management System (ETRIMS) and dynamic data related to traffic flow parameters were obtained from Radar Detection System (RDS) stations. Since RDS stations in Tennessee are mostly located within city boundaries, only those crashes that occurred on primary interstates, I-40 and I-55, within Memphis were chosen. Our analysis included crashes on 21.51 miles of I-40 and 12.28 miles of I-55. Figure 6 shows the location of the city and the two interstates superimposed over the state map of Tennessee. Discretized crash data for analysis was obtained from the study area as follows.

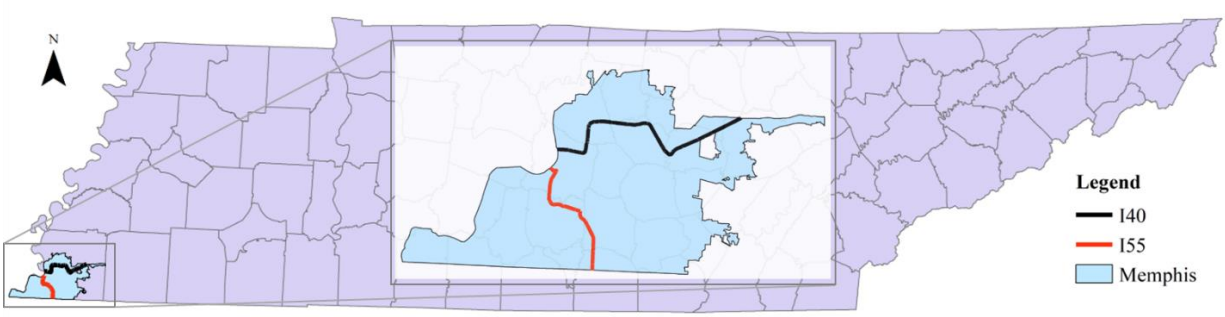


Figure 6 Study area showing I-40 and I-55

ETRIMS, Tennessee Department of Transportation's main portal for transportation-related data, was used to extract crash data. ETRIMS is an online query-based system that stores information on highway geometry, aggregated traffic flow, and highway crashes in the state. The highway inventory information in ETRIMS formed the basis for our highway segmentation. Specifically, highway segmentation was based on i) the direction of traffic, ii) the number of lanes, iii) terrain (flat or rolling), and iv) posted speed limit. This produced 146 segments for I-40, and 94 segments for I-55. Crash data for the study area also included exact coordinates which enabled mapping each crash to its respective segment based on geographical proximity and

direction of travel. For each segment, the duration between consecutive crashes was then determined using the date and time of the crash. The inter-crash duration was discretized into 1-hour epochs with 15-min intervals to create forecasting epochs as presented in Table 13. To avoid confusion, it is necessary here to distinguish between crash data retrieved from ETRIMS and crash data obtained after the creation of forecasting epochs. Therefore, we will refer to the process of creating forecasting epochs as data expansion and the data obtained from it as expanded data. The crash data from ETRIMS (without forecasting epochs) will be referred to as original crash data.

Table 14
Descriptive statistics of variables in crash dataset

Categorical variables	**Frequency**	**Percentage**
Time of day		
Early morning (6 a.m. to 9 a.m.)	208	17.72
Late morning (9 a.m. to 12 p.m.)	132	11.24
Early afternoon (12 p.m. to 3 p.m.)	175	14.91
Late afternoon (3 p.m. to 6 p.m.)	292	24.87
Evening (6 p.m. to 12 a.m.)	208	17.72
Night (12 a.m. to 6 a.m.)	159	13.54
Weather condition		
Clear	849	72.32
Cloudy, rain, fog, or snow	325	27.68
Lighting condition		
Daylight	758	64.57
Dark lighted	287	24.45
Dark, not lighted	129	10.99
Terrain		
Flat	640	54.51
Rolling	534	45.49
Land use		
Rural	448	38.16
Commercial	595	50.68
Mixed	131	11.16

Continuous variables	**Min**	**Q1**	**Median**	**Q3**	**Max**	**Mean**	**SD**
Highway geometry							
Number of lanes (one-direction)	2	3	4	5	6	3.89	1.07
Inter-crash duration (hours)	0	105	324	865.50	6,608	652	863.75

It is also worth noting here that our study excluded crashes occurring near entrance and exit ramps since traffic flow parameters vary significantly in these locations. This resulted in a total of 1,174 crashes on the two interstates. Table 14 presents the descriptive statistics of variables in the study. Figure 7 shows the distribution of inter-crash duration for all the interstate segments. The distribution of inter-crash duration suggests that about 62% of crashes occurred within 500 hours (~21 days) of the first crash. Notably, the mean and median durations were 652 hours and 324 hours respectively suggesting a right-skewed distribution. The minimum duration between crashes was 0 hours, suggesting at least two different crashes in a segment occurred at the same time. The maximum duration between crashes was 6,608 hours.

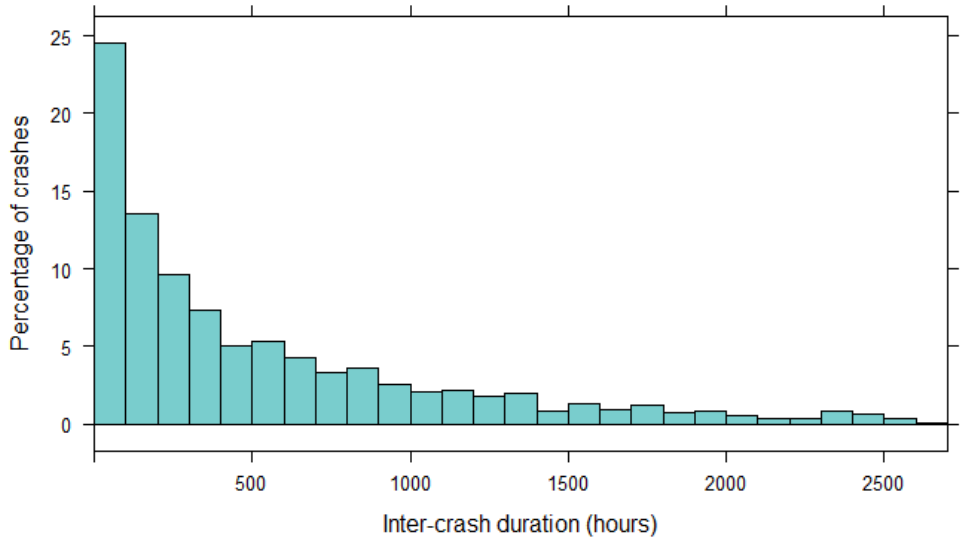


Figure 7 Distribution of inter-crash duration

ETRIMS stores aggregated data on traffic flow parameters, therefore, dynamic traffic flow for the study area was collected from a different source. Specifically, RDS stations spread out across the interstate segments were used to obtain dynamic or time-varying traffic flow parameters associated with each crash. First, RDS stations were mapped to their respective highway segments (and therefore crashes) based on geographical proximity and direction of travel. Three traffic flow parameters, viz, speed, volume, and occupancy sampled every 15-min

were obtained for each RDS station mapped to the crashes for the year 2019. The sampled flow parameters were then matched against the discretized 15-min time intervals between crashes using timestamps. The resulting database, therefore, included static and dynamic co-variates for every 15-min interval of an epoch for all crashes in the study area.

Sampling approach

As evident from our discussion of the methodology, the size of the expanded data depends on three factors, i) the size of original crash data, ii) inter-crash duration between crashes, and iii) choice of time discretization. With an increase in these factors, the expanded data enlarges which brings forth computational challenges. To overcome this, random sampling techniques with progressive sampling can be employed to identify the smallest sample that can best represent the complete dataset (Provost et al., 2001). In this method, the size of random samples used for training is gradually increased while checking whether the model accuracy increases with an increase in sample size. The smallest sample size can then be chosen for subsequent modeling. For this study, we follow a similar approach wherein different sizes of the training data are randomly sampled at different levels. Models trained on these samples are then compared against models trained on the complete training data and tested on training data. It is worth pointing out that this is practical for our dataset due to the relatively small region and time frame chosen for analysis. Our findings can however be generalized and incorporated in future studies that apply the current modeling framework to analyze crashes for a larger region and time frame. Next, we illustrate the approach used sampling data for training and testing the duration-based MNL model.

To compare model performance for various samples, it is necessary to train the MNL on different samples but test on the same. Therefore, while models were trained on samples derived

from various sampling approaches, they were all tested on the same testing data. Specifically, 90% of original crash data was selected for training and the remaining 10% was held out for testing. All subsequent samples for model training were derived from the training data. Similarly, the expanded test data was used to test all trained models. Based on different sampling approaches, ten models were trained and tested on the expanded testing data. The samples used to train these models were drawn at three levels, crash, epoch, and segments as summarized in Table 15.

Crash level sampling

At the crash level, a specific percentage of crashes were drawn from the original training set and then expanded. Specifically, we drew 5%, 25%, 50%, and 100% of original crashes and expanded them to form the training data for the crash level sampling. This was straightforward considering each crash had a unique identifier. In other words, we employed a simple random sampling approach here. These samples drawn at this level are denoted by the prefix *Cr_* in Table 15 and henceforth. For example, *Cr_100%* represents the complete expanded dataset containing all original crashes from the training set.

Epoch level sampling

For sampling at the epoch level, the original testing crash data was first expanded. Then a specific percentage of observations was drawn at random from the expanded data. It is worth mentioning here that observations in the expanded dataset correspond to various epochs of crashes and not unique crashes. We utilized unique crash identifiers and corresponding epoch numbers to draw unique samples. The number of samples drawn was 5%, 25%, and 50% of the expanded dataset. We identify these samples using the prefix *Ep_*.

Table 15
Description of training and testing data

Data	Sampling level	Model name	Description
Training	Crash	Cr_100%	100% of the expanded dataset with 100% of original training crash data
		Cr_5%	100% of the expanded dataset with 5% of original training crash data
		Cr_25%	100% of the expanded dataset with 25% of original training crash data
		Cr_50%	100% of the expanded dataset with 50% of original training crash data
Training	Epoch	Ep_5%	5% of the expanded dataset with 100% of original training crash data
		Ep_25%	25% of the expanded dataset with 100% of original training crash data
		Ep_50%	50% of the expanded dataset with 100% of original training crash data
Training	Segment	Sg_5%	100% of the expanded dataset with 5% of original training crashes drawn from segment strata
		Sg_25%	100% of the expanded dataset with 25% of original training crashes drawn from segment strata
		Sg_50%	100% of the expanded dataset with 50% of original training crashes drawn from segment strata
Testing	Crash	-	100% of the expanded dataset with 100% of original testing crash data

Note: Original training crash data = 90% of original crash data, original testing crash data = 10% of original crash data.

Segment level sampling

To draw samples at the segment level, each roadway segment was assumed to be a stratum. Crashes were then sampled from each stratum. This sampling was carried out using the original training crash data and then expanded. For example, when drawing 5% of crashes at the segment level, 5% of total crashes were drawn such that crashes sampled were from different segments. As with sampling at crash and epoch levels, 5%, 25%, and 50% crashes were sampled based on highway segments. The sampled crash data was then expanded. The resulting expanded dataset is denoted by the prefix *Sg_*.

Results

An initial analysis of variables was conducted to avoid multicollinearity issues in the predictor variables. Results showed a high correlation between occupancy and speed derived from RDS stations. This is expected since speed and occupancy are interrelated. Therefore, occupancy was removed from all our subsequent models. Analysis of dynamic traffic volumes obtained from RDS stations showed large variation in data. The volumes were therefore scaled between 0 and 1. No more data, cleaning was needed after this and the MNL was run using complete training data, to obtain the *Cr_100%* model. Table 16 presents the results from the model trained on *Cr_100%*. In our model estimation, the last alternative (whether the crash occurred beyond the current epoch) was set as the base alternative for which only a constant was estimated irrespective of the segment and epoch. Based on the results, the propensity equation for the base alternative can therefore be written as:

$$U_{s,e,Beyond\ 1\ hour\ of\ current\ epoch}(base) = 5.336$$

Relative to the base case, the propensity equation for the four intervals (i) of each epoch (e) and segment (s) can be written as:

$$U_{s,e,i} = -0.254 \times t_{e,i} + 0.327 \times \left(t_{e,i}\right)^2 - 0.823 \times Early\ morning - 0.723 \times Late\ morning \ldots - 0.029 \times Average\ speed - 0.920 \times Volume$$

A few notable observations can be made from the results in the table. First, both times since the last crash and its quadratic polynomial is statistically significant. This suggests a non-linear influence of time on the occurrence of crashes. Second, compared to night, crashes are the least likely to occur in late afternoons of the epoch compared to future epochs. When the weather condition is unfavorable, there is an increased likelihood of crashes in the current epoch. The same is true for unfavorable lighting conditions. Compared to rolling terrain, in flat terrains

crashes are more likely to occur in the future epochs. When looking at the effect of land use, we observed a higher and lower likelihood of crashes in commercial and rural areas respectively. With an increase in vehicle speed and traffic flow, crashes are more likely to occur in the future epochs. This suggests that during better flow conditions, crashes are less likely to occur frequently (i.e., inter-crash duration is higher). However, roadway segments with a larger number of lanes are more likely to encounter crashes during a certain epoch compared to the future.

Table 16
Result for the model trained on *Cr_100%* data

Variable groups	Variables	Parameter	t-stat
Intercept	Intercept	5.336	70.233
Duration dynamics	Time since last crash	-0.254	-44.350
	Square of time since last crash	0.327	41.951
Time of day (base = Night (12 a.m. to 6 a.m.))	Early morning (6 a.m. to 9 a.m.)	-0.823	-63.138
	Late morning (9 a.m. to 12 p.m.)	-0.289	-58.829
	Early afternoon (12 p.m. to 3 p.m.)	-1.005	-73.728
	Late afternoon (3 p.m. to 6 p.m.)	-2.026	-74.878
	Evening (6 p.m. to 12 a.m.)	-0.656	-62.907
Weather condition (base = Clear)	Cloud rain fog or snow	0.024	14.337
Lighting condition (base = Daylight)	Dark lighted	0.147	55.970
	Dark not lighted	0.024	18.659
Terrain (base = Rolling)	Flat	-0.407	-80.277
Land use (base = Mixed)	Rural	-0.376	-89.028
	Commercial	0.097	34.930
Traffic stream characteristics	Average speed	-0.029	-19.336
	15-minute traffic flow	-0.920	-64.593
Highway geometry	Number of lanes	0.060	2.554
Goodness of fit			
Average initial LL		-0.0831	
Average LL at convergence		-0.0139	
Average predicted LL		-0.0161	
Number of observations in the training sample		695,374	
Number of observations in the testing sample		68,544	

Effect of sampling approach and sample sizes on model estimates

Next, we ran the MNL for the remaining nine samples and tested the estimated model using the testing set. The average predicted log-likelihood for the testing set is used to compare the trained models. Table 17 presents the results from our model estimations. We see inconsistent results for models trained on samples drawn at crash and segment levels. Of all the models, models estimated using 25% and 50% of samples drawn at epoch level, *Ep_25%,* and *Ep_50%*, respectively, provided estimates that were closest to *Cr_100%*. The average predicted log-likelihood for the *Ep_25%* and *Ep_50%* were also the same. Table 18 better highlights the differences in estimates across different models compared to the *Cr_100%* model. The average of differences in parameters of a model is also provided to give an idea of the difference in the estimated value of parameters across the models. This is calculated as the sum of the % differences of all parameters divided by the number of parameters for each model. The values of average percentage difference also suggest that models trained on samples drawn at the epoch level performed considerably well than those trained on samples drawn at crash and segment levels. Specifically, for *Ep_25%* and *Ep_50%*, the average percentage difference in parameters estimates was about 31% and 34% respectively. These values are considerably smaller than the rest of the models. There was also a considerable improvement in model estimation times for the *Ep_25%* and *Ep_50%* models compared against the *Cr_100%* model. The estimation times were as follows: 159 seconds for *Ep_25%,* 334 seconds for *Ep_50%,* and 647 seconds for *Cr_100%*.

Table 17
Parameter estimates and models fit for samples drawn at various levels

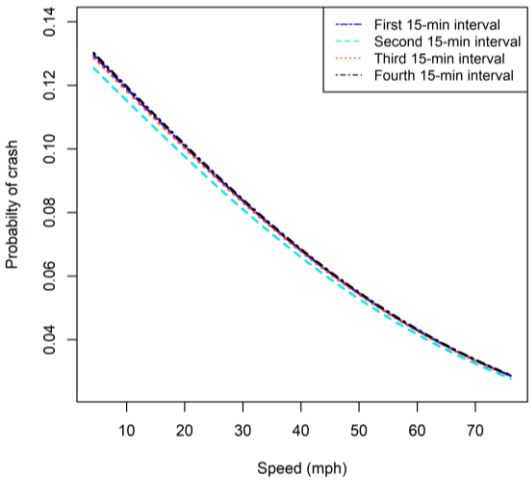
Variable groups	Variables	Crash level sampling				Epoch level sampling			Segment level sampling		
		Cr_100 %	Cr_5%	Cr_25 %	Cr_50 %	Ep_5%	Ep_25 %	Ep_50 %	Sg_5%	Sg_25%	Sg_50%
Intercept	Intercept	5.336	7.203	4.586	5.016	5.903	5.619	5.505	7.273	6.321	5.770
Duration dynamics	Time since last crash	-0.254	0.294	0.052	-0.047	-0.261	-0.318	-0.284	0.271	-0.323	-0.039
	Square of time since last crash	0.327	2.550	0.710	0.562	0.493	0.257	0.267	2.538	0.864	0.630
Time of day (base = Night)	Early morning	-0.823	-2.650	-1.078	-0.742	-0.791	-0.824	-0.838	-2.728	-0.521	-1.097
	Late morning	-0.289	-0.605	-0.010	-0.373	-0.601	-0.431	-0.373	-0.637	-0.321	-0.596
	Early afternoon	-1.005	0.221	-0.606	-1.025	-1.208	-1.111	-1.077	0.205	-1.891	-1.206
	Late afternoon	-2.026	-0.894	-2.647	-2.182	-2.465	-2.059	-2.054	-0.892	-2.857	-2.154
	Evening	-0.656	-1.075	-0.065	-0.480	-0.172	-0.574	-0.642	-1.070	-0.307	-0.298
Weather condition (base = Clear)	Cloud rain fog or snow	0.024	-1.470	-0.262	-0.009	0.198	0.082	0.087	-1.536	0.131	-0.194
Lighting condition (base = Daylight)	Dark lighted	0.147	-1.076	0.275	0.251	0.612	0.262	0.224	-1.014	0.251	0.118
	Dark not lighted	0.024	0.179	0.090	-0.002	-0.007*	0.029	0.038	0.178	0.229	-0.065
Terrain (base = Rolling)	Flat	-0.407	-0.482	0.094	-0.154	-0.538	-0.475	-0.378	-0.454	-0.679	-0.040
Land use (base = Mixed)	Rural	-0.376	-0.351	0.222	-0.091	-0.465	-0.372	-0.320	-0.329	-0.907	-0.006*
	Commercial	0.097	0.430	0.052	0.110	0.265	0.109	0.072	0.397	0.448	0.012
Traffic stream characteristics	Average speed	-0.029	-0.023	-0.052	-0.039	-0.020	-0.025	-0.025	-0.019	-0.038	-0.028
	Volume	-0.920	-1.180	-1.366	-1.162	-1.167	-0.901	-0.925	-1.208	-0.845	-1.292
Highway geometry	Number of lanes	0.060	0.405	0.031*	0.055	-0.048*	0.068*	0.022*	0.377	0.419	0.032*
Goodness of fit											
Average initial LL		-0.0831	-0.1169	-0.0776	-0.0812	-0.0798	-0.0791	-0.0853	-0.1179	-0.0558	-0.0710
Average LL at convergence		-0.0139	-0.0163	-0.0132	-0.0137	-0.0126	-0.0131	-0.0134	-0.0166	-0.0093	-0.0122
Average predicted LL		-0.0160	-0.0162	-0.0166	-0.0161	-0.0165	-0.0162	-0.0162	-0.0163	-0.0162	-0.0161
Number of observations in training sample		695,374	24,501	187,330	354,264	34,843	173,843	347,687	24,575	272,146	428,823
Number of observations in the testing sample		68,544									

Note: All parameters are statistically significant at a 5% level of significance except for those accompanied by an asterisk (*).

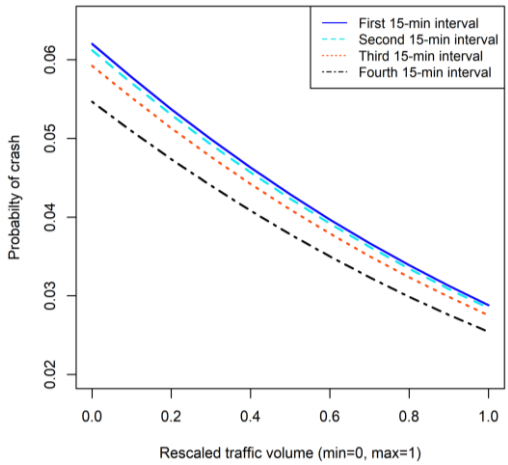
Table 18
Percentage difference compared to the model trained on *Cr_100%* data

Variable groups	Variables	Crash level sampling			Epoch level sampling			Segment level sampling		
		Cr_5%	**Cr_25 %**	**Cr_50 %**	**Ep_5%**	**Ep_25 %**	**Ep_50 %**	**Sg_5%**	**Sg_25%**	**Sg_50%**
Intercept	Intercept	-34.99	14.06	6.00	-10.63	-5.30	-3.17	-36.30	-18.46	-8.13
Duration dynamics	Time since last crash	215.75	120.47	81.50	-2.76	-25.20	-11.81	206.69	-27.17	84.65
	Square of time since last crash	-679.82	-117.13	-71.87	-50.76	21.41	18.35	-676.15	-164.22	-92.66
Time of day (base = Night)	Early morning	-221.99	-30.98	9.84	3.89	-0.12	-1.82	-231.47	36.70	-33.29
	Late morning	-109.34	96.54	-29.07	-107.96	-49.13	-29.07	-120.42	-11.07	-106.23
	Early afternoon	121.99	39.70	-1.99	-20.20	-10.55	-7.16	120.40	-88.16	-20.00
	Late afternoon	55.87	-30.65	-7.70	-21.67	-1.63	-1.38	55.97	-41.02	-6.32
	Evening	-63.87	90.09	26.83	73.78	12.50	2.13	-63.11	53.20	54.57
Weather condition (base = Clear)	Cloud rain fog or snow	6225.00	1191.67	137.50	-725.00	-241.67	-262.50	6500.00	-445.83	908.33
Lighting condition (base = Daylight)	Dark lighted	831.97	-87.07	-70.75	-316.33	-78.23	-52.38	789.80	-70.75	19.73
	Dark not lighted	-645.83	-275.00	108.33	129.17	-20.83	-58.33	-641.67	-854.17	370.83
Terrain (base = Rolling)	Flat	-18.43	123.10	62.16	-32.19	-16.71	7.13	-11.55	-66.83	90.17
Land use (base = Mixed)	Rural	6.65	159.04	75.80	-23.67	1.06	14.89	12.50	-141.22	98.40
	Commercial	-343.30	46.39	-13.40	-173.20	-12.37	25.77	-309.28	-361.86	87.63
Traffic stream characteristics	Average speed	20.69	-79.31	-34.48	31.03	13.79	13.79	34.48	-31.03	3.45
	Volume	-28.26	-48.48	-26.30	-26.85	2.07	-0.54	-31.30	8.15	-40.43
Highway geometry	Number of lanes	-575.00	48.33	8.33	180.00	-13.33	63.33	-528.33	-598.33	46.67
Goodness of fit										
Average percentage change in parameters		599.93	152.82	45.40	113.47	30.94	33.74	609.97	177.54	121.85
Percentage difference in average initial LL		-40.67	6.62	2.29	3.97	4.81	-2.65	-41.88	32.85	14.56
Percentage difference in average LL at convergence		-17.27	5.04	1.44	9.35	5.76	3.60	-19.42	33.09	12.23
Percentage difference in average predicted LL		-1.25	-3.75	-0.62	-3.13	-1.25	-1.25	-1.87	-1.25	-0.62

Note: A negative percentage change in average LL indicates a lower average LL for the model versus the Cr_100% model.



a. Change in probability of crash with speed



b. Change in probability of crash with traffic flow

Figure 8 Change in probability of crashes across different time intervals

In summary, the findings from model runs indicate that complete training data is not a necessity to obtain consistent estimates when using the framework. This finding is particularly useful considering that smaller discretization of inter-crash duration is desired, and the size of expanded data increases considerably for smaller time resolutions. Therefore, when using this framework, modelers can focus their effort and resources on using finer time resolutions or increasing study area than on using larger samples for improving predictive accuracy. Based on findings, we suggest if the data size becomes an issue, crashes should be drawn at the epoch level to obtain consistent estimates. This will reduce computational time considerably with negligible loss in predictive accuracy.

Effect Of Speed and Volume on Crash Occurrence

As we mentioned earlier, various effects of speed and traffic flow on the occurrence of crashes have been reported in the literature. It is, therefore, necessary to explore the influence of traffic flow parameters on crash occurrence using the current framework. Based on the model estimates, we present the change in the probability of crashes for the range of speed and traffic volumes in

Figure 8. Figure 8(a) suggests that for the same speed, the probability of crashes is lower in the second 15-minute interval of an epoch. Similarly, from Figure 8(b) it is evident that for the same traffic volume, crashes are less likely to occur in later intervals of an epoch.

Discussion and future implications

This study considered two interstates in the city of Memphis, TN for investigating sampling techniques and sample size requirements for a recently proposed duration model-based crash prediction framework. Despite the relatively small study area, the expanded data obtained after reformulation (creation of forecasting epochs) was considerably large. For comparison, the original training data consisted of only 1,174 unique crashes. Upon expansion using a discretization of 1-hour epochs with 15-minute intervals, the number of observations increased to 695,374. Notably, this increase in data size is mainly due to our choice of finer discretization. Nevertheless, the same issues can be expected when applying the framework. The choice of time discretization is crucial in real-time crash prediction modeling since a finer resolution can capture variations in traffic flow conditions more precisely in turn providing meaningful insights on their effects. For example, for a model formulation where the epochs are considerably longer, say, an hour or multiple hours, the effect of speed can be positive implying crashes are less likely to occur in the future. Such a finding would be a considerable departure from those considering shorter epochs such as ours where we report a contradictory result. However, as mentioned earlier, smaller discretization will result in considerable data size and challenge in model estimation. Nevertheless, sampling techniques can be used to reduce the computational complexity and still derive accurate estimates when applying the framework as demonstrated in this study.

Despite challenges with data size, the potential implications of using smaller discretization in the current framework are attractive since real-time data collected at small intervals can better capture the variations in dynamic traffic flow and predict crashes in a segment more accurately. Beyond, crash prediction, applications of this framework can extend to address questions such as "*How will crashes be temporally distributed in a segment during congestions?"* and "*How likely are crashes to occur during a particular time of a day?*". Answer to these questions can in turn be used to predict secondary crashes, queue lengths and travel delay (Sarker et al., 2015), and adopt optimal strategies for safety improvements (e.g., see (K. Haque et al., 2021; Khasnabis et al., 2012; Mishra, 2013; Mishra et al., 2015)). Understandably, such applications will require strategic organization of real-time data collection and its processing (data cleaning, expansion, and sampling in this case). This will bring forth considerable challenges in terms of data storage and computational power needed to obtain crash probabilities in real-time. To this end, traffic planners can consider scalable distributed systems where each component can be assigned specific tasks to reduce the computational load and therefore the overall computational time. These systems can be installed and maintained in-house in local or regional Transportation Management Centers where the traffic data for a geographical area is collected and stored. Additionally, data latency (the time taken to transmit the collected data) and bandwidth (amount of data that can be transferred at a time) should also be given due consideration to ensure the data is collected and analyzed accurately and in real-time. These requirements are inherent to any real-time crash prediction model that includes the current framework.

Conclusions

Duration models, which have been widely applied to static risk models for crash prediction, have not been employed on dynamic or real-time cash prediction models due to their inability to handle time-varying covariates. A recent study has introduced a duration-based MNL framework that overcomes this issue by dividing the duration between crashes into equal-time intervals and modeling them as alternatives in an MNL wherein the time-varying covariates can be introduced as interval-specific variables. This approach however comes at the expense of a large data size resulting from the discretization of inter-crash duration into equal time intervals to create forecasting epochs. Even for a relatively smaller study area, the expanded data can be expected to be considerably large. To alleviate the computational complexity arising from large data, this study investigated the effect of sampling techniques and sample sizes on the computational time and parameter estimates. In doing so, the assumption of homogenous highway segments and a longer forecasting window adopted in the original study were also addressed. The assumption of homogeneous segments was addressed by segmenting I-40 and I-55 in Memphis, Tennessee based on geometric attribute (number of lanes), terrain (flat or rolling), and flow (speed limit). A realistic future window of 1-hour discretized into 15-minute time intervals was used in this study to discretize the duration between crashes. Training the MNL framework on samples drawn using various techniques and its evaluation using testing data revealed that models trained on smaller samples can provide reasonably accurate estimates provided the sampling is done appropriately. Among samples drawn at the crash, epoch, and segment levels, epoch level samples provide the most consistent estimates and better prediction. Results indicate that as little as 25% of samples drawn at the epoch level can reduce training time to 25% of the time needed to train the model using complete data with minimal error in parameter estimates.

Although rigorous and established methods have been adopted in current research, it is not without limitations. Future studies can address three notable limitations in this study. First, the current study ignores unobserved heterogeneity between segments. If unobserved heterogeneity exists, the parameter estimates derived might not be consistent. Second, the current study overlooks traffic movements when segmenting highways. Variation in traffic movements across highway segments can influence crashes. Third, the focus of this study was to explore sampling techniques that make the framework easier to train. Therefore, no comparisons were made with other prediction models. However, a comparison between the current framework and other established real-time crash prediction models in terms of their predictive accuracy is necessary. In particular, ML algorithms have proven to be more accurate than regression techniques. Future studies can compare the predictive abilities of the current framework with machine learning algorithms explored in the literature.

References

2019 Highway Work Zone Safety Survey. (2019). Associated General Contractors of America. https://www.agc.org/news/2019/05/23/2019-highway-work-zone-safety-survey

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

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Abdel-Aty, M., & Pande, A. (2007). Crash data analysis: Collective vs. Individual crash level approach. *Journal of Safety Research*, *38*(5), 581–587. https://doi.org/10.1016/j.jsr.2007.04.007

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record: Journal of the Transportation Research Board*, *1897*(1), 88–95. https://doi.org/10.3141/1897-12

Afghari, A. P., Haque, M. M., & Washington, S. (2020). Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. *Accident Analysis & Prevention*, *144*, 105615. https://doi.org/10.1016/j.aap.2020.105615

Agent, K. R., & Hibbs, J. O. (1996). *Evaluation of SHRP Work Zone Safety Devices*. 24.

Ahmed, S. S., Cohen, J., & Anastasopoulos, P. Ch. (2021). A correlated random parameters with heterogeneity in means approach of deer-vehicle collisions and resulting injury-severities. *Analytic Methods in Accident Research*, *30*, 100160. https://doi.org/10.1016/j.amar.2021.100160

Al-Ghamdi, A. S. (2002). Pedestrian–vehicle crashes and analytical techniques for stratified contingency tables. *Accident Analysis & Prevention*, *34*(2), 205–214. https://doi.org/10.1016/S0001-4575(01)00015-X

Algomaiah, M., & Li, Z. (2022). Enhancing Work Zone Capacity by a Cooperative Late Merge System Using Decentralized and Centralized Control Strategies. *Journal of Transportation Engineering, Part A: Systems*, *148*(2). https://doi.org/10.1061/JTEPBS.0000632

Ali, Y., Haque, M. M., Zheng, Z., Washington, S., & Yildirimoglu, M. (2019). A hazard-based duration model to quantify the impact of connected driving environment on safety during mandatory lane-changing. *Transportation Research Part C: Emerging Technologies*, *106*(June), 113–131. https://doi.org/10.1016/j.trc.2019.07.015

Ariannezhad, A., Karimpour, A., Qin, X., Wu, Y.-J., & Salmani, Y. (2021). Handling Imbalanced Data for Real-Time Crash Prediction: Application of Boosting and Sampling Techniques. *Journal of Transportation Engineering, Part A: Systems*, *147*(3), 04020165. https://doi.org/10.1061/JTEPBS.0000499

Bagloee, S. A., & Asadi, M. (2016). Crash analysis at intersections in the CBD: A survival analysis model. *Transportation Research Part A: Policy and Practice*, *94*, 558–572. https://doi.org/10.1016/j.tra.2016.10.019

Barua, S., El-Basyouny, K., & Islam, Md. T. (2016). Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic Methods in Accident Research*, *9*, 1–15. https://doi.org/10.1016/j.amar.2015.11.002

Baruya, A. (1998). Road Safety in Europe. *9th Internation Conference: Road Safety in Europe*.

Bashir, S., & Zlatkovic, M. (2021). Assessment of Queue Warning Application on Signalized Intersections for Connected Freight Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(10), 1211–1221. https://doi.org/10.1177/03611981211015247

Benekohal, R. F., Hajbabaie, A., Medina, J. C., Wang, M.-H., & Chitturi, M. V. (2010). *SPEED PHOTO-RADAR ENFORCEMENT EVALUATION IN ILLINOIS WORK ZONES* (FHWA-ICT-10-064). Illinois Department of Transportation.

Berthaume, A. L. (2015). *Microscopic Modeling of Driver Behavior Based on Modifying Field Theory for Work Zone Application* [Doctoral Dissertation, University of Massachusetts Amherst]. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1328&context=dissertations_2

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2011). Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. *2011 World Congress on Information and Communication Technologies*, 1241–1246. https://doi.org/10.1109/WICT.2011.6141426

Bham, G. H., & Mohammadi, M. A. (2011). *Evaluation of Work Zone Speed Limits: An Objective and Subjective Analysis of Work Zones in Missouri Report*. 92.

Brownstone, D., & Small, K. A. (1989). Efficient Estimation of Nested Logit models. *Journal of Business & Economic Statistics*, *7*(1), 67–74. https://doi.org/10.1080/07350015.1989.10509714

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., & Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. *Transportation Research Part C: Emerging Technologies*, *117*, 102697. https://doi.org/10.1016/j.trc.2020.102697

Cerwick, D. M., Gkritza, K., Shaheed, M. S., & Hans, Z. (2014). A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, *3–4*, 11–27. https://doi.org/10.1016/j.amar.2014.09.002

Cestac, J., Paran, F., & Delhomme, P. (2011). Young drivers' sensation seeking, subjective norms, and perceived behavioral control and their roles in predicting speeding intention:

How risk-taking motivations evolve with gender and driving experience. *Safety Science*, *49*(3), 424–432. https://doi.org/10.1016/j.ssci.2010.10.007

Chang, H. L., & Jovanis, P. P. (1990). Formulating accident occurrence as a survival process. *Accident Analysis and Prevention*, *22*(5), 407–419. https://doi.org/10.1016/0001-4575(90)90037-L

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, *43*(8), 541–557. https://doi.org/10.1016/j.ssci.2005.04.004

Chang, Y., & Edara, P. (2018). Predicting hazardous events in work zones using naturalistic driving data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-March*, 1–6. https://doi.org/10.1109/ITSC.2017.8317847

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, *90*, 128–139. https://doi.org/10.1016/j.aap.2016.02.011

Cheng, W., Gill, G. S., Dasu, R., Xie, M., Jia, X., & Zhou, J. (2017). Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis & Prevention*, *99*, 330–341. https://doi.org/10.1016/j.aap.2016.11.022

Cheng, Z., Lu, J., Zu, Z., & Li, Y. (2019). Speeding Violation Type Prediction Based on Decision Tree Method: A Case Study in Wujiang, China. *Journal of Advanced Transportation*, *2019*, 1–10. https://doi.org/10.1155/2019/8650845

Choudhary, P., & Velaga, N. R. (2020). Impact of distraction on decision making at the onset of yellow signal. *Transportation Research Part C: Emerging Technologies*, *118*(March 2019), 102741. https://doi.org/10.1016/j.trc.2020.102741

Chung, Y. (2010). Development of an accident duration prediction model on the Korean Freeway Systems. *Accident Analysis and Prevention*, *42*(1), 282–289. https://doi.org/10.1016/j.aap.2009.08.005

Data USA: Highway Maintenance Workers. (2018). https://datausa.io/profile/soc/highway-maintenance-workers

Debnath, A. K., Blackman, R., & Haworth, N. (2015). Common hazards and their mitigating measures in work zones: A qualitative study of worker perceptions. *Safety Science*, *72*, 293–301. https://doi.org/10.1016/j.ssci.2014.09.022

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113

Dissanayake, S., & Akepati, S. R. (2009). *Identification of Work Zone Crash Characteristics*. Federal Highway Administration. https://intrans.iastate.edu/app/uploads/2018/08/Dissanayake_WZCrashChar.pdf

Dissanayake, S., & Lu, J. (2002). Analysis of Severity of Young Driver Crashes: Sequential Binary Logistic Regression Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, *1784*(1), 108–114. https://doi.org/10.3141/1784-14

Dong, C., Clarke, D. B., Yan, X., Khattak, A., & Huang, B. (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate

crash frequencies at intersections. *Accident Analysis & Prevention*, *70*, 320–329. https://doi.org/10.1016/j.aap.2014.04.018

Elliott, M. A., & Thomson, J. A. (2010). The social cognitive determinants of offending drivers' speeding behaviour. *Accident Analysis & Prevention*, *42*(6), 1595–1605. https://doi.org/10.1016/j.aap.2010.03.018

Eseonu, C., Gambatese, J., & Nnaji, C. (2018). *Reducing Highway Fatalities Through Improved Adoption of Safety Technologies*.

Federal Highway Administration. (2009a). *Manual of Traffic Control Devices for Streets and Highways*.

Federal Highway Administration. (2009b). *Manual on Uniform Traffic Control Devices (MUTCD)*. https://mutcd.fhwa.dot.gov/

Federal Highway Administration. (2023). *FHWA Work Zone Facts and Statistics*. Work Zone Management Program. https://ops.fhwa.dot.gov/wz/resources/facts_stats.htm

Flannagan, C. A., Selpi, Baykas, P. B., Leslie, A., Kovaceva, J., & Thomson, R. (2019). *Analysis of SHRP2 Data to Understand Normal and Abnormal Driving Behavior in Work Zones* (FHWA-HRT-20-010). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/48835

Forward, S. E. (2009). The theory of planned behaviour: The role of descriptive norms and past behaviour in the prediction of drivers' intentions to violate. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 198–207. https://doi.org/10.1016/j.trf.2008.12.002

Fountas, G., & Anastasopoulos, P. Ch. (2017). A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. *Analytic Methods in Accident Research, 15*, 1–16. https://doi.org/10.1016/j.amar.2017.03.002

Furth, P. G. (2011). *Sampling and Estimation Techniques for Estimating Bus System Passenger-Miles*. Bureau of Transportation Statistics. https://www.bts.gov/archive/publications/journal_of_transportation_and_statistics/volume_08_number_02/paper_07/index

Furth, P. G., Killough, K. L., & Ruprecht, G. F. (1988). Cluster Sampling Techniques for Estimating Transit Patronage. *Transportation Research Record, 1165*.

Gambatese, J. A., Lee, H. W., & Nnaji, C. A. (2017). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance*. Oregon State University School of Civil and Construction Engineering.

Gambatese, J., & Lee, H. W. (2016). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance II*. (Issue 503).

Gan, H., Wei, J., & Wang, G. (2021). A generic work zone evaluation tool driven by a macroscopic traffic simulation model. *International Journal of Mobile Communications*, *19*(1), 1. https://doi.org/10.1504/IJMC.2021.111884

Garber, N. J., & Ehrhart, A. A. (2000). Effect of Speed, Flow, and Geometric Characteristics on Crash Frequency for Two-Lane Highways. *Transportation Research Record: Journal of the Transportation Research Board*, *1717*(1), 76–83. https://doi.org/10.3141/1717-10

Gelman, A., & Hill, J. (2007). When does a multilevel modeling make a difference? In *Data Analysis Using Regression and Multilevel/Hierarchical Models* (pp. 237–249). Cambridge University Press.

Golob, T. F., Recker, W. W., & Leonard, J. D. (1987). An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention*, *19*(5), 375–395. https://doi.org/10.1016/0001-4575(87)90023-6

Guo, H., Wang, W., Guo, W., & Zhao, F. (2013). Modeling lane-keeping behavior of bicyclists using survival analysis approach. *Discrete Dynamics in Nature and Society*, *2013*. https://doi.org/10.1155/2013/197518

Hamdar, S. H., Khoury, H., & Zehtabi, S. (2016). A simulator-based approach for modeling longitudinal driving behavior in construction work zones: Exploration and assessment. *SIMULATION*, *92*(6), 579–594. https://doi.org/10.1177/0037549716644515

Haque, K., Mishra, S., & Golias, M. M. (2021). Multi-period transportation network investment decision making and policy implications using econometric framework. *Research in Transportation Economics*, *89*, 101109. https://doi.org/10.1016/j.retrec.2021.101109

Haque, M. M., & Washington, S. (2015). The impact of mobile phone distraction on the braking behaviour of young drivers: A hazard-based duration model. *Transportation Research Part C: Emerging Technologies*, *50*, 13–27. https://doi.org/10.1016/j.trc.2014.07.011

Harb, R., Radwan, E., Yan, X., Pande, A., & Abdel-Aty, M. (2008). Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression. *Journal of Transportation Engineering*, *134*(5), 203–214. https://doi.org/10.1061/(ASCE)0733-947X(2008)134:5(203)

Harmon, T., Bahar, G., & Gross, F. (2018). *Crash Costs for Highway Safety Analysis*.

Hashmienejad, S. H. A., & Hasheminejad, S. M. H. (2017). Traffic accident severity prediction using a novel multi-objective genetic algorithm. *International Journal of Crashworthiness*, *22*(4), 425–440. https://doi.org/10.1080/13588265.2016.1275431

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second). John Wiley & Sons, Inc.

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019a). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019b). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

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

Hou, G., & Chen, S. (2019). An Improved Cellular Automaton Model for Work Zone Traffic Simulation Considering Realistic Driving Behavior. *Journal of the Physical Society of Japan*, *88*(8), 084001. https://doi.org/10.7566/JPSJ.88.084001

Hourdos, J. (2012). Portable, Non-Intrusive Advance Warning Devices for Work Zones with or without Flag Operators. *Minnesota Department of Transportation*, *October*.

Imprialou, M. I. M., Quddus, M., Pitfield, D. E., & Lord, D. (2016). Re-visiting crash-speed relationships: A new perspective in crash modelling. *Accident Analysis and Prevention*, *86*, 173–185. https://doi.org/10.1016/j.aap.2015.10.001

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Jonathan, A.-V., Wu, K.-F. (Ken), & Donnell, E. T. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention*, *87*, 8–16. https://doi.org/10.1016/j.aap.2015.11.006

Jovanis, P. P., & Chang, H. L. (1989). Disaggregate model of highway accident occurrence using survival theory. *Accident Analysis and Prevention*, *21*(5), 445–458. https://doi.org/10.1016/0001-4575(89)90005-5

Jovanović, D., Šraml, M., Matović, B., & Mićić, S. (2017). An examination of the construct and predictive validity of the self-reported speeding behavior model. *Accident Analysis & Prevention*, *99*, 66–76. https://doi.org/10.1016/j.aap.2016.11.015

Jung, S., Qin, X., & Noyce, D. A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis & Prevention*, *42*(1), 213–224. https://doi.org/10.1016/j.aap.2009.07.020

Kashyap, A. A., Raviraj, S., Devarakonda, A., Nayak K, S. R., K V, S., & Bhat, S. J. (2022). Traffic flow prediction models – A review of deep learning techniques. *Cogent Engineering*, *9*(1), 2010510. https://doi.org/10.1080/23311916.2021.2010510

Ke, J., Zhang, S., Yang, H., & Chen, X. (Michael). (2019). PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. *Transportmetrica A: Transport Science*, *15*(2), 872–895. https://doi.org/10.1080/23249935.2018.1542414

Keramati, A., Lu, P., Zhou, X., & Tolliver, D. (2020). A Simultaneous Safety Analysis of Crash Frequency and Severity for Highway-Rail Grade Crossings: The Competing Risks Method. *Journal of Advanced Transportation*, *2020*(1). https://doi.org/10.1155/2020/8878911

Khasnabis, S., Mishra, S., & Safi, C. (2012). Evaluation procedure for mutually exclusive highway safety alternatives under different policy objectives. *Journal of Transportation Engineering*, *138*(7), 940–948. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000397

Khattak, A. J., Khattak, A. J., & Council, F. M. (2002). Effects of work zone presence on injury and non-injury crashes. *Accident Analysis and Prevention*, *34*(1), 19–29. https://doi.org/10.1016/S0001-4575(00)00099-3

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, & W. Wong, Eds.; Thid Editi). Springer. https://doi.org/10.1007/978-1-4419-6646-9

Kloeden, C. N., McLean, J., & Glonek, G. F. V. (2002). *Reanalysis of travelling speed and the risk of crash involvement in Adelaide South Australia*. Australian Transport Safety Bureau.

Kock, N., & Lynn, G. S. (2012). Lateral Collinearity and Misleading Results in Variance-Based SEM : An Illustration and Recommendations Lateral Collinearity and Misleading Results in Variance-. *Journal of the Association for Information Systems*, *13*(7), 546–580.

Lee, C., Hellinga, B., & Saccomanno, F. (2003). Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, *1840*(1), 67–77. https://doi.org/10.3141/1840-08

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *Applied Sciences*, *10*(1), 129. https://doi.org/10.3390/app10010129

Lee, J.-T., & Fazio, J. (2005). Influential Factors in Freeway Crash Response and Clearance Times by Emergency Management Services in Peak Periods. *Traffic Injury Prevention*, *6*(4), 331–339. https://doi.org/10.1080/15389580500255773

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis & Prevention*, *135*, 105371. https://doi.org/10.1016/j.aap.2019.105371

Li, Y., & Bai, Y. (2008). Development of crash-severity-index models for the measurement of work zone risk levels. *Accident Analysis and Prevention*, *40*(5), 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012

Li, Y., & Bai, Y. (2009). Highway work zone risk factors and their impact on crash severity. *Journal of Transportation Engineering*, *135*(10), 694–701. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000055

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident Analysis and Prevention*, *111*(November 2017), 354–363. https://doi.org/10.1016/j.aap.2017.11.028

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

Ma, J., & Kockelman, K. (2006a). Crash frequency and severity modeling using clustered data from Washington state. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, October*, 1621–1626. https://doi.org/10.1109/itsc.2006.1707456

Ma, J., & Kockelman, K. M. (2006b). Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record: Journal of the Transportation Research Board*, *1950*, 24–34.

Ma, J., Kockelman, K. M., & Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*, *40*(3), 964–975. https://doi.org/10.1016/j.aap.2007.11.002

Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research*, *25*, 100113. https://doi.org/10.1016/j.amar.2020.100113

Marks, E., Vereen, S., & Awolusi, I. (2017). *Active Work Zone Safety Using Emerging Technologies 2017*. University Transportation Center for Alabama The University of Alabama.

Martin, J., Rozas, A., & Araujo, A. (2016). A WSN-Based Intrusion Alarm System to Improve Safety in Road Work Zones. *Journal of Sensors*, *2016*, 1–8. https://doi.org/10.1155/2016/7048141

Medina-Salgado, B., Sánchez-DelaCruz, E., Pozos-Parra, P., & Sierra, J. E. (2022). Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, *35*, 100739. https://doi.org/10.1016/j.suscom.2022.100739

Meng, Q., & Weng, J. (2011). A Genetic algorithm approach to assessing work zone casualty risk. *Safety Science*, *49*(8–9), 1283–1288. https://doi.org/10.1016/j.ssci.2011.05.001

Mishra, S. (2013). A Synchronized Model for Crash Prediction and Resource Allocation to Prioritize Highway Safety Improvement Projects. *Procedia - Social and Behavioral Sciences*, *104*, 992–1001. https://doi.org/10.1016/j.sbspro.2013.11.194

Mishra, S., Golias, M. M., Sharma, S., & Boyles, S. D. (2015). Optimal funding allocation strategies for safety improvements on urban intersections. *Transportation Research Part A: Policy and Practice*, *75*, 113–133. https://doi.org/10.1016/j.tra.2015.03.001

Mishra, S., Golias, M. M., & Thapa, D. (2021). *Work Zone Alert Systems*. Tennessee Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56274

Mohan, D., Bangdiwala, S. I., & Villaveces, A. (2017). Urban street structure and traffic safety. *Journal of Safety Research*, *62*, 63–71. https://doi.org/10.1016/j.jsr.2017.06.003

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis. *Transportation Research Record*, *2673*(11), 680–692. https://doi.org/10.1177/0361198119845899

Nam, D., & Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, *34*(2), 85–102. https://doi.org/10.1016/S0965-8564(98)00065-2

Novosel, C. (2014). Evaluation of Advanced Safety Perimeter Systems for Kansas Temporary Work Zones. In *Civil, Environmental, and Architectural Engineering, University of Kansas*.

Osman, M., Mishra, S., & Paleti, R. (2018a). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group

differences. *Accident Analysis and Prevention*, *118*(May), 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018b). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018c). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., Paleti, R., & Golias, M. (2019). Impacts of Work Zone Component Areas on Driver Injury Severity. *Journal of Transportation Engineering, Part A: Systems*, *145*(8), 04019032. https://doi.org/10.1061/jtepbs.0000253

Osman, M., Paleti, R., & Mishra, S. (2018a). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis and Prevention*, *111*(May 2017), 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., & Mishra, S. (2018b). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis & Prevention*, *111*, 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis and Prevention*, *97*, 261–273. https://doi.org/10.1016/j.aap.2016.10.020

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051–10072. https://doi.org/10.1007/s12652-020-02759-5

Ozturk, O., Ozbay, K., Yang, H., & Bartin, B. (2013). Crash Frequency Modeling for Highway Construction Zones. *Transportation Research Board's 92nd Annual Meeting, Washington, D.C*, 14p.

Paleti, R., Mahmud, A., Gayah, V., & Pinjari, A. (2021). When and Where does the Next Traffic Crash Occur? A Discretized Duration Based Modeling Approach. *Under Review for Publication*.

Park, E. S., & Lord, D. (2007). Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record*, *2019*, 1–6. https://doi.org/10.3141/2019-01

Pei, X., Wong, S. C., & Sze, N. N. (2011). A joint-probability approach to crash prediction models. *Accident Analysis & Prevention*, *43*(3), 1160–1166. https://doi.org/10.1016/j.aap.2010.12.026

Pham, M.-H., Bhaskar, A., Chung, E., & Dumont, A.-G. (2010). Random forest models for identifying motorway Rear-End Crash Risks using disaggregate data. *13th International IEEE Conference on Intelligent Transportation Systems*, 468–473. https://doi.org/10.1109/ITSC.2010.5625003

Provost, F., Jensen, D., & Oates, T. (2001). Progressive Sampling. In *Instance Selection and Construction for Data Mining* (pp. 151–170). Springer US. https://doi.org/10.1007/978-1-4757-3359-4_9

Qi, Y., Srinivasan, R., Teng, H., & Baker, R. F. (2005). *Frequency of Work Zone Accidents on Construction Projects*.

Quddus, M. (2013). Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, *5*(1), 27–45. https://doi.org/10.1080/19439962.2012.705232

Rahim, M. A., & Hassan, H. M. (2021). A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention*, *154*, 106090. https://doi.org/10.1016/j.aap.2021.106090

Rahman, R., Bhowmik, T., Eluru, N., & Hasan, S. (2021). Assessing the crash risks of evacuation: A matched case-control approach applied over data collected during Hurricane Irma. *Accident Analysis & Prevention*, *159*, 106260. https://doi.org/10.1016/j.aap.2021.106260

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007

Sarker, A. A., Naimi, A., Mishra, S., Golias, M. M., & Freeze, P. B. (2015). Development of a Secondary Crash Identification Algorithm and occurrence pattern determination in large scale multi-facility transportation network. *Transportation Research Part C: Emerging Technologies*, *60*, 142–160. https://doi.org/10.1016/j.trc.2015.08.011

Scott-Parker, B., Hyde, M. K., Watson, B., & King, M. J. (2013). Speeding by young novice drivers: What can personal characteristics and psychosocial theory add to our understanding? *Accident Analysis & Prevention*, *50*, 242–250. https://doi.org/10.1016/j.aap.2012.04.010

Shangguan, Q., Fu, T., & Liu, S. (2020). Investigating rear-end collision avoidance behavior under varied foggy weather conditions: A study using advanced driving simulator and survival analysis. *Accident Analysis and Prevention*, *139*(March), 105499. https://doi.org/10.1016/j.aap.2020.105499

Sharma, A., Bullock, D., & Peeta, S. (2011). Estimating dilemma zone hazard function at high speed isolated intersection. *Transportation Research Part C: Emerging Technologies*, *19*(3), 400–412. https://doi.org/10.1016/j.trc.2010.05.002

Simons-Morton, B. G., Ouimet, M. C., Chen, R., Klauer, S. G., Lee, S. E., Wang, J., & Dingus, T. A. (2012). Peer influence predicts speeding prevalence among teenage drivers. *Journal of Safety Research*, *43*(5–6), 397–403. https://doi.org/10.1016/j.jsr.2012.10.002

Song, J. J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, *97*(1), 246–273. https://doi.org/10.1016/j.jmva.2005.03.007

Stipancic, J., Miranda-Moreno, L., Saunier, N., & Labbe, A. (2019). Network screening for large urban road networks: Using GPS data and surrogate measures to model crash frequency and severity. *Accident Analysis and Prevention*, *125*(February), 290–301. https://doi.org/10.1016/j.aap.2019.02.016

Stout, D., Graham, J., Bryant-Fields, B., Migletz, J., Fish, J., & Hanscom, F. (1993). *Maintenance Work Zone Safety Devices Development and Evaluation*.

Sun, J., & Sun, J. (2016a). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Sun, J., & Sun, J. (2016b). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y., & Huang, H. (2020). Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Analytic Methods in Accident Research*, *27*, 100123. https://doi.org/10.1016/j.amar.2020.100123

Thapa, D., & Mishra, S. (2021a). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., & Mishra, S. (2021b). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*, 106125. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., Paleti, R., & Mishra, S. (2022a). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Thapa, D., Paleti, R., & Mishra, S. (2022b). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Theiss, L., Ullman, G. L., & Lindheimer, T. (2017). *Closed Course Performance Testing of the Aware Intrusion Alarm System*.

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2019). Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention*, *130*, 151–159. https://doi.org/10.1016/j.aap.2017.12.018

Therneau, T. M. (2020). *A Package for Survival Analysis in R. R package version 3.2-7*.

Therneau, T. M., Grambsch, P. M., & Panktatz, S. V. (2003). Penalized Survival Models and Frailty. *Journal of Computational and Penalized Survival Models and Frailty*, *12*(1), 156–175.

Ullman, G. L., Trout, N. D., & Theiss, L. (2016). *Driver Responses to the AWARE Intrusion Alarm System*. Texas A&M Transportation Institute.

Venugopal, S., & Tarko, A. (2000). Safety models for rural freeway work zones. *Transportation Research Record*, *1715*, 1–9. https://doi.org/10.3141/1715-01

Wang, C., Quddus, M. A., & Ison, S. G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention*, *43*(6), 1979–1990. https://doi.org/10.1016/j.aap.2011.05.016

Wang, J., Yamamoto, T., & Liu, K. (2020). Key determinants and heterogeneous frailties in passenger loyalty toward customized buses: An empirical investigation of the subscription termination hazard of users. *Transportation Research Part C: Emerging Technologies*, *115*(July 2019), 102636. https://doi.org/10.1016/j.trc.2020.102636

Wang, X., Katz, R., & Dong, X. S. (2018). *Fatal Injuries at Road Construction Sites among Construction Workers* [Quarterly]. Center for Construction Research and Training.

https://www.cpwr.com/wp-content/uploads/publications/publications_Quarter2-QDR-2018.pdf

Work Zones-Injury Facts-National Safety Council. (2020). https://injuryfacts.nsc.org/motor-vehicle/motor-vehicle-safety-issues/work-zones/

Wu, L., Meng, Y., Kong, X., & Zou, Y. (2020). Incorporating survival analysis into the safety effectiveness evaluation of treatments: Jointly modeling crash counts and time intervals between crashes. *Journal of Transportation Safety and Security*, *0*(0), 1–21. https://doi.org/10.1080/19439962.2020.1786871

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

Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., & Jiang, X. (2020). A machine-learning method for improving crash injury severity analysis: A case study of work zone crashes in Cairo, Egypt. *International Journal of Injury Control and Safety Promotion*, *27*(3), 266–275. https://doi.org/10.1080/17457300.2020.1746814

Yang, H., Ozbay, K., Ozturk, O., & Xie, K. (2015). Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. *Traffic Injury Prevention*, *16*(4), 387–396. https://doi.org/10.1080/15389588.2014.948615

Yasmin, S., & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention*, *59*, 506–521. https://doi.org/10.1016/j.aap.2013.06.040

Yasmin, S., & Eluru, N. (2018). A joint econometric framework for modeling crash counts by severity. *Transportmetrica A: Transport Science*, *14*(3), 230–255. https://doi.org/10.1080/23249935.2017.1369469

Yasmin, S., Eluru, N., Bhat, C. R., & Tay, R. (2014). A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic Methods in Accident Research*, *1*, 23–38. https://doi.org/10.1016/j.amar.2013.10.002

Yasmin, S., Eluru, N., Wang, L., & Abdel-Aty, M. A. (2018). A joint framework for static and real-time crash risk analysis. *Analytic Methods in Accident Research*, *18*, 45–56. https://doi.org/10.1016/j.amar.2018.04.001

Ye, X., Pendyala, R. M., Shankar, V., & Konduri, K. C. (2013). A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention*, *57*, 140–149. https://doi.org/10.1016/j.aap.2013.03.025

Yu, B., Chen, Y., & Bao, S. (2019). Quantifying visual road environment to establish a speeding prediction model: An examination using naturalistic driving data. *Accident Analysis & Prevention*, *129*, 289–298. https://doi.org/10.1016/j.aap.2019.05.011

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, *51*, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis and Prevention*, *73*, 351–358. https://doi.org/10.1016/j.aap.2014.09.006

Zhang, C., He, J., Wang, Y., Yan, X., Zhang, C., Chen, Y., Liu, Z., & Zhou, B. (2020). A Crash Severity Prediction Method Based on Improved Neural Network and Factor Analysis. *Discrete Dynamics in Nature and Society*. https://doi.org/10.1155/2020/4013185

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods. *IEEE Access*, *6*, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*(1), 204–215. https://doi.org/10.1016/j.ijforecast.2010.06.002

Zhang, K., & Hassan, M. (2019a). Crash severity analysis of nighttime and daytime highway work zone crashes. *PLoS ONE*, *14*(8), 1–17. https://doi.org/10.1371/journal.pone.0221128

Zhang, K., & Hassan, M. (2019b). Identifying the Factors Contributing to Injury Severity in Work Zone Rear-End Crashes. *Journal of Advanced Transportation*, *2019*, 1–9. https://doi.org/10.1155/2019/4126102

Zhao, G., Wu, C., & Qiao, C. (2013). A Mathematical Model for the Prediction of Speeding with its Validation. *IEEE Transactions on Intelligent Transportation Systems*, *14*(2), 828–836. https://doi.org/10.1109/TITS.2013.2257757

Zheng, L., & Sayed, T. (2020). A novel approach for real time crash prediction at signalized intersections. *Transportation Research Part C: Emerging Technologies*, *117*, 102683. https://doi.org/10.1016/j.trc.2020.102683

Zimmerman, K., Mzige, A. A., Kibatala, P. L., Museru, L. M., & Guerrero, A. (2012). Road traffic injury incidence and crash characteristics in Dar es Salaam: A population based

study. *Accident Analysis & Prevention*, *45*, 204–210.

https://doi.org/10.1016/j.aap.2011.06.018

4. Advancing Proactive Crash Prediction: A Discretized Duration Approach for Predicting Crashes and Severity

Introduction

Crash prediction models can be categorized into two main types: diagnostic crash prediction models, also known as reactive crash prediction models, and proactive or real-time crash prediction models. These two types of prediction models differ in their application and the variables they incorporate. Reactive crash prediction models rely on historical crash data, as well as static covariates (variables that do not change over time) and dynamic covariates (variables that do change over time), aggregated over a specific period. Examples of such dynamic covariates include Average Annual Daily Traffic and average speed. These models are valuable for developing safety performance functions, which help identify the precursors of crashes and evaluate the impact of safety interventions and policies on highway safety (Yasmin et al., 2018). On the other hand, proactive crash prediction models refer to real-time crash prediction models that utilize historical crash data and static covariates, such as roadway condition and roadway geometry, along with disaggregated dynamic covariates that vary with time. These dynamic covariates can include traffic volume, speed, and weather conditions collected in near real-time. By incorporating dynamic predictors, these models can account for changing traffic and weather conditions, allowing for the forecasting of the likelihood of future crashes in real time. This, in turn, enables the implementation of crash mitigation strategies.

Proactive crash prediction models have garnered significant attention from researchers in recent years due to their potential to forecast and prevent future crashes. The availability of granular traffic flow data, such as near real-time traffic flow data collected at small time intervals, from Intelligent Transportation System infrastructure, coupled with the computational

performance of modern computers, has played a crucial role in increasing the popularity of these models. Computationally effective data-driven methods, such as Machine Learning (ML), have also aided in their popularity as they replace traditional statistical models which are often slower to calibrate (Mannering et al., 2020). Additionally, data-driven methods have also demonstrated superior data fit and predictive capabilities as they are not constrained by assumptions inherent to traditional econometric frameworks, such as statistical distribution and variable correlation. However, data-driven methods have their own limitations too. They struggle with problems related to model transferability, generalization, and the inability to quantify variable effects. In this context, statistical econometric frameworks, through variable coefficients and elasticities, can quantify variable effects and provide model transferability and generalization. In these respects, statistical models can be considered superior to data-driven methods.

Due to the benefits offered by statistical econometric frameworks, there are ongoing efforts to enhance and refine traditional statistical approaches to address their limitations and apply them to proactive modeling. For instance, researchers have extended standard econometric frameworks by incorporating flexible structures to develop mixed and generalized models. These models can account for unobserved heterogeneity and hierarchical structures for variable correlations and dependencies. More recently, researchers developed and implemented a new crash prediction framework (Thapa et al., 2022a). In their study, researchers developed a duration-based crash prediction model that combines elements of the survival model and Multinomial Logit model (MNL). In this modeling approach, the time duration between crashes is divided into 1-hour epochs, which are further subdivided into 4 15-minute time intervals. Each epoch between two consecutive crashes is treated as a separate observation, with the time intervals serving as choice alternatives. By adopting this approach, the framework can forecast

the likelihood of future crashes by considering two types of covariates. Firstly, static covariates associated with crashes, such as highway geometry and environmental conditions, are repeated over each epoch. Secondly, dynamic covariates, such as traffic flow and speed, change across epochs and within the 15-minute time intervals. The authors of the study discovered that the duration-based model could generate reasonably accurate estimates even when dealing with small sample sizes.

The current study builds upon the duration-based model by incorporating crash severities. While prediction of crash occurrence has already been addressed in previous research, forecasting likelihood of different crash severities is crucial from multiple perspectives, including safety, economic, and planning considerations. The costs associated with crashes vary significantly depending on their severity. For instance, the comprehensive unit cost of a Property Damage Only (PDO) crash in the US was estimated to be around $12,000 in 2016, whereas a fatal crash was estimated to exceed $11 million (Harmon et al., 2018). Additionally, crash severities are linked to road user costs. Studies have indicated that more severe crashes require more time to clear, resulting in higher road user costs (Golob et al., 1987; J.-T. Lee & Fazio, 2005). Therefore, prioritizing the identification and addressing of factors contributing to more severe crashes is crucial from both safety and economic perspectives. Furthermore, from a planning standpoint, the ability to forecast crash severities provides transportation agencies with valuable insights. Agencies are often constrained with limited resources and personnel, making it necessary to identify critical segments in advance and proactively address adverse traffic flow conditions. By forecasting crash severities, agencies can prioritize the allocation and deployment of resources and personnel to prevent severe crashes and mitigate their impacts, contributing to more efficient and effective traffic operations and planning.

Literature review

Research in crash prediction has focused on forecasting both crash occurrences and severities. In the following sections, we provide a literature review of prediction models based on the specific outcomes they forecast. While we will discuss both proactive and reactive crash prediction models, this review will place greater emphasis on proactive crash prediction models, as they align with the scope of our study.

Crash prediction models

The first group of studies focuses on real-time forecasting of future crashes, employing both data-driven and statistical methods. Researchers have utilized various approaches to develop these models. Data-driven methods have gained popularity in the literature, with several notable examples including Support Vector Machines (Sun & Sun, 2016b; R. Yu & Abdel-Aty, 2013), decision trees and random forests (Beshah et al., 2011; Pham et al., 2010), neural networks (P. Li et al., 2020), and Bayesian statistics (Hossain & Muromachi, 2012; Zheng & Sayed, 2020). These data-driven methods have proven effective in capturing complex relationships and patterns in crash data, allowing for real-time forecasting of future crash occurrences.

On the statistical side, the case-control design approach has been the most popular method for developing proactive crash prediction models (Hossain et al., 2019b). In this approach, crashes are matched with non-crash events based on specific variables such as location and time of the crash (M. Abdel-Aty et al., 2004). The resulting dataset, with binary outcomes indicating crash or non-crash events, is well-suited for binary logistic regression. However, researchers have also explored the use of data-driven methods and Bayesian statistics to enhance the modeling capabilities of this approach (Hossain et al., 2019b). In addition to the traditional case-control approach, alternative methodologies have been proposed. For example, (Yasmin et

al., 2018) developed a MNL that considered 5-minute intervals for the next 30 days as choice alternatives, representing the occurrence of crashes in future time intervals. Given the substantial number of choice alternatives, the authors employed sampling techniques (selecting 29 randomly sampled time intervals and 1 interval with a crash) from the 30-day period.

More recently, researchers implemented a real-time crash prediction model by combining survival model with the MNL model. Survival models or duration models have been employed to model traffic crashes using static data (e.g., (Jovanis & Chang, 1989; Thapa & Mishra, 2021a), however, they are incapable of incorporating time-varying covariates. The researchers developed a new method to restructure the crash data by creating forecasting epochs and time-intervals that can be associated with the dynamic covariates (Thapa et al., 2022a).

Crash severity prediction models

The second group of studies focuses on predicting crash severity. Data-driven methods have been used more often to forecast crash severities, with various approaches utilized in different studies. Deep learning methods have been applied in crash severity prediction (Rahim & Hassan, 2021), while Support Vector Machines have been utilized in studies by (Chen et al., 2016; Iranitalab & Khattak, 2017). Random forests have also been used as a predictive technique for crash severity forecasting (Iranitalab & Khattak, 2017). Other methods such as neural networks and decision trees have been explored in some studies (J. Lee et al., 2019; Ospina-Mateus et al., 2021; C. Zhang et al., 2020). In recent years, a significant focus has been placed on comparing the performance of these algorithms in crash severity prediction (Santos et al., 2022). It is important to note that most prediction models within this group are reactive in nature, aiming to predict crash severity based on historical data and established patterns.

The most common statistical approach for developing crash severity prediction models is applying discrete choice models, specifically multinomial and ordered response logit/probit models. However, more advanced statistical models such as random parameter mixed models have gained popularity among researchers in recent years, as they offer solutions to the fixed parameter restriction imposed by choice models. Uncorrelated random parameter models (Fountas & Anastasopoulos, 2017) correlated random parameter models (Ahmed et al., 2021; Fountas & Anastasopoulos, 2017), and generalized ordered response models (Osman, Mishra, et al., 2018b; Osman, Paleti, et al., 2018a; Osman et al., 2019; Yasmin et al., 2014) are some of the examples of these advanced statistical models. These models enable researchers to account for parameter variations across different observations, providing more flexibility in capturing the complexity of crash severity prediction. Another approach for crash severity prediction involves the use of sequential models that can account for the dependency between various levels of crash severities. Studies have explored the application of sequential models in crash severity prediction, allowing for the consideration of dependencies between crash severities (Dissanayake & Lu, 2002; Jung et al., 2010).

With the advent of advanced models, researchers have conducted studies to examine and compare their predictive performance. For instance, (Yasmin & Eluru, 2013) compared different generalized and mixed models within the frameworks of ordered and unordered choice modeling. Their findings indicated that mixed generalized ordered logit and mixed MNL models showed promise in predicting crash injury severity. In a study by (J. Zhang et al., 2018), various statistical and machine learning methods were compared, and it was found that machine learning algorithms exhibited better performance. This improvement could be attributed to factors such as the linear utility function and parametric assumptions regarding the error term. (Cerwick et al.,

2014) conducted a comparison between mixed MNL and latent class MNL models. Their analysis revealed that the former model provided better average predictions across different severity levels.

Models predicting crash frequency and severity

The final group of studies focuses on forecasting both crashes and their severity. However, it is important to note that most of these models are primarily designed to forecast crash frequencies rather than the presence or absence of crashes.

Multivariate count data models are commonly employed in these studies, as seen in the works of (Jonathan et al., 2016; Ma & Kockelman, 2006b; Park & Lord, 2007). Additionally, random parameter count data models have been used to account for spatial and temporal heterogeneity, as demonstrated by (Barua et al., 2016; W. Cheng et al., 2017; Dong et al., 2014). Other studies have implemented joint models with two components: (i) a crash prediction component utilizing count data models, and (ii) a crash severity component employing discrete choice models to predict crash counts by severity. This approach has been employed by (Afghari et al., 2020; Pei et al., 2011; Yasmin & Eluru, 2018).

The sequential logit model has also been used to predict the likelihood and severity of crashes. (Xu et al., 2013) developed a model using sequential binary logit models, where crashes were modeled in three stages: Stage 1 (crash vs. non-crash), Stage 2 (property damage only vs. higher severities), and Stage 3 (non-capacitating vs. higher severities). However, a significant drawback of the sequential logit model in the context of proactive crash prediction is that the estimation of multiple models can be computationally demanding and time-consuming, making it impractical for large datasets.

Study contributions

Only a limited number of statistical approaches have been developed to date for proactive crash prediction, apart from the commonly used case-control approach. This study introduces a duration-based prediction model for both crash occurrence and crash severity. The model framework involves dividing the time duration between historical crashes into distinct time periods to create forecasting epochs and time intervals. This allows the model to incorporate dynamic covariates and ascertain the probability of crashes occurring in future epochs and time intervals (Thapa et al., 2022a). While this modeling approach has previously been demonstrated for crash prediction, the current study extends the framework to incorporate crash severities. The major contributions of this paper can be summarized as follows.

1. We expand upon the duration-based proactive crash prediction model by introducing a novel modeling approach that can forecast both crash occurrence and severity. Our model framework is one of handful statistical approaches for proactive crash prediction that does not rely on the case-control approach (Thapa et al., 2022a). Unlike the original model, which solely predicts the likelihood of crashes for discrete future time intervals, our proposed model can also predict the corresponding crash severities.

   Furthermore, the proposed model is implemented using a larger dataset. Specifically, the model is applied to crash data collected from interstates in two cities in Tennessee, thereby achieving a broader geographical coverage in comparison to the previous study that focused on a single city. This expanded geographical scope enhances the generalizability of the crash predictors, as it ensures adequate representation of diverse roadway conditions and traffic patterns across the study areas.

2. The proposed modeling framework demands discretizing the time duration between crashes to create forecasting epochs (more on this in this in the next section). Consequently, the size of the initial crash data expands significantly. Prior studies have indicated that appropriate sampling techniques can address estimation complexities arising from large data size, thereby allowing for parameter estimation with a reasonable degree of accuracy (Thapa et al., 2022a). However, the incorporation of crash severities adds an additional layer of complexity to the model estimation process.

Therefore, this study aims to investigate the influence of sample size on variable coefficients and identify variables that are sensitive to changes in sample size. Understanding the variables that are particularly impacted by sample size variations is crucial for the implementation of the model. Additionally, this information will play a pivotal role in assessing the reliability of the model and guiding future data collection efforts.

Methodology

In this section, we present the methodology under three distinct subsections: the duration-based prediction framework, the nested logit model, and the estimation of the nested logit model. First, we describe the duration-based prediction framework and the process of creating forecasting epochs. This section is followed by the introduction of the two-level nested logit model and its relationship with the duration-based crash prediction framework. Finally, we discuss the estimation processes used in this study to estimate the parameters of the models.

Duration based prediction framework

In the duration-based crash prediction model, the occurrence of a crash at any time interval dt can be modeled using the MNL framework with infinite alternatives, *n* and the hazard rate, *h* given by $U_n = -h(n-1)dt(\text{Thapa et al.}, 2022\text{a})$ (Thapa et al., 2022a). By utilizing this

relationship, the latent propensity function for each time interval can be expressed as a function of static and dynamic covariates (time-varying factors). The application of this concept is illustrated in the following example.

Table 19
Historical crash data with static covariates

s	Crash	Date of crash	Time of crash	Severity	Terrain (Flat=1, Rolling=0)
1	A1	1/1/2023	00:00	Fatal	1
1	A2	1/1/2023	02:30	PDO	1
1	A3	1/1/2023	03:00	Injury	1

Table 20
Dynamic covariates averaged for 15-min intervals: Vehicle speed (in mph)

Date and time	**1/1/2023 00:00**	**1/1/2023 00:15**	**1/1/2023 00:30**	**1/1/2023 01:00**
Speed	49	51	50	49
Date and time	**1/1/2023 01:15**	**1/1/2023 01:30**	**1/1/2023 01:45**	**1/1/2023 02:00**
Speed	47	50	48	49
Date and time	**1/1/2023 02:15**	**1/1/2023 02:30**	**1/1/2023 02:45**	**1/1/2023 03:00**
Speed	51	50	50	51
Date and time	**1/1/2023 03:15**	**1/1/2023 03:30**	**1/1/2023 03:45**	**1/1/2023 04:00**
Speed	49	48	47	48
⋮	⋮	⋮	⋮	⋮

Table 21
Final Crash data after creating forecasting epochs

s	ID	Time to crash (hr)	Epoch	15-min intervals				Next epoch	Speed (mph)				Severity	Terrain			
				1	2	3	4		1	2	3	4		1	2	3	4
1	A1	2.5	1	0	0	0	0	1	49	51	50	49	Fatal	0.25	0.5	0.75	1
1	A1	2.5	2	0	0	0	0	1	47	50	48	49	Fatal	1.25	1.50	1.75	2
1	A1	2.5	3	0	1	0	0	0	51	50	50	51	Fatal	2.25	2.50	2.75	3
1	A2	0.5	1	0	1	0	0	0	49	48	47	48	PDO	0.25	0.5	0.75	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Example:

Consider the duration between crashes in a highway segment, denoted as *s*, which is discretized into epochs, denoted as *e*, each with time intervals, denoted as *i*, and each interval has a duration of *dt*. Using these indices, we can examine historical crash data for a roadway segment, *s*=1,

where three consecutive crashes, denoted as A1, A2, and A3, were observed with durations of 2.5 hours and 0.5 hours apart (see Table 19). Additionally, available are dynamic covariates, speed and volume for the segment and the crash year at a temporal resolution of *dt*, as shown in Table 20. These covariates, as depicted, exhibit time-varying characteristics.

For discretization, let us choose *e*=1 hour and *dt*=0.25 hours. Therefore, the number of time intervals in an epoch, denoted by *C*=4, each identified by the index *i*=(1, 2, 3, 4). After discretization, the forecasting epochs are created as shown in Table 21. Each epoch consists of four 15-minute intervals, and an additional column called "*Next epoch*" is added, indicating whether the next crash occurred in the current or future epoch (0 if in the current epoch, 1 if in future epochs). Based on the table, we can express the time elapsed since the previous crash using the equation $t_{e,i} = (e-1)Cdt + (i-1)dt$. For example, the time between crashes A1 and A2 can be determined as $t_{3,2} = (3-1)1 + (2-1)0.25 = 2.25$ hours. As shown in the table, the dynamic covariate *Speed* varies across different time periods. The static covariate *Terrain*, in this example, does not repeat across the time intervals of a crash. However, to account for the effect of time, the variable is multiplied by $t_{e,i}$. For instance, the *terrain* variable for the first time-interval is 0.25 multiplied by 1, and for the second time interval, it is 0.5 multiplied by 1, and so on. Therefore, all variables vary across epochs and time-intervals. The final data obtained after the creation of forecasting epochs takes the form of panel data with repeated observations for each crash corresponding to the forecasting epochs.

A few observations can be made from Table 21, particularly regarding the increase in data size after the creation of forecasting epochs. The final data size is influenced by three factors. The first factor is the size of the original crash data. The more crashes are observed, the larger the data size will be after creating forecasting epochs. The second factor is the choice of

discretization. When a smaller time discretization is chosen, more detailed information regarding traffic flow can be obtained. However, this also leads to a considerable increase in data size. The third factor is the distribution of inter-crash duration. If the inter-crash durations are longer, more forecasting epochs will be created, resulting in a larger data size. Considering these factors, implementing a model for a wide geographical area with small discretization can become computationally demanding. Even a slight reduction in time discretization significantly increases computational complexity. To reduce computational complexity, it is suggested to use a smaller sample of the expanded data drawn at the epoch level for model training (Thapa et al., 2022a).

Based on the example provided, the latent propensity function for crash severities, *k* and time interval, *i* can be represented as a function of time since crash, static, and dynamic covariates using the utility function, $U_{k,i}$ in equation 17.

$$U_{k,i} = \beta_t t_{e,i} + r' X_{e,i} \tag{17}$$

In equation 17, the coefficient β_t represents the impact of duration on crash severity. The vector of covariates, $X_{e,i}$, captures the effect of covariates, with its values varying across epochs and time intervals. The corresponding vector of coefficients is denoted by r'. Additionally, if we assume that the latent propensity for the upper-level alternatives consists only of an intercept term, the utility equations for each alternative can be formulated using equation 18.

$$\mathrm{V_{e,i}} = \beta_\mathrm{i} \tag{18}$$

Since the occurrence of a crash at a specific time interval is dependent on the absence of crashes in previous time intervals, the conditional probability of observing a crash in a particular time interval within an epoch can be expressed using a random variable denoted as T_s as follows.

$$P\left(T_s = t_{e,i} \middle| T_s > (e-1)Cdt\right) = \frac{exp\ (V_{e,i})}{\sum_{c=1}^{C} exp\ (V_{e,c}) + exp\ (V_{s,e,c+1})} \tag{19}$$

The unconditional probability of a crash at any time interval can be obtained by multiplying the conditional probability in equation 19 with the cumulative product of all probabilities for the C+1th intervals preceding the epoch *e*.

$$P(T_s = t_{e,i}) = \frac{exp\,(V_{e,i})}{\sum_{c=1}^{C} exp\,(V_{e,c}) + exp\,(V_{e,C+1})} \cdot \prod_{e*=1}^{e-1} \frac{exp\,(V_{e*,C+1})}{\sum_{c=1}^{C} exp\,(V_{e*,c}) + exp\,(V_{e*,C+1})} \quad (20)$$

Nested logit model

The crash outcomes in the example are characterized by two aspects: (i) the time interval when a crash happens, and (ii) the severity of the crash. These outcomes can be effectively modeled using a two-level nested logit model, as depicted in Figure 9. In this model, the time intervals, *i* and an additional alternative (*C*+1) serve as nodes representing the upper-level choice alternatives, while the crash severities correspond to the lower-level alternatives. It is important to note that the crash severities at each time interval are conditional upon the occurrence of a crash within that interval. For simplicity, assume the severity levels are comprised of two categories, denoted by *k* = (F/I, PDO), where F/I represents Fatal or Injury crashes, and PDO represents Property Damage Only crashes. The conditional choice probability of the lower-level alternatives, *k* given the upper-level alternatives, *i* can be expressed as follows.

$$P_k = P(k|i) * P(i) \quad (21)$$

where,

$$P(k|i) = \frac{exp\left(U_{k,i}/\theta_i\right)}{\sum_k exp\left(U_{k,i}/\theta_i\right)} \quad (22)$$

$$P(i) = \frac{\exp\left(V_{e,i} + \Gamma_i * \theta_i\right)}{\sum_i \exp\left(V_{e,i} + \Gamma_i * \theta_i\right) + \exp(V_{e,Next\ epoch})} \quad (23)$$

$$\Gamma_i = log\left[\sum_k \exp\left(U_{k,i}/\theta_i\right)\right] \quad (24)$$

The parameter θ_i in equations 22, 23, and 24 represents the logsum parameter or nesting coefficient, which captures the underlying correlations for alternatives within a nest. Equation 24 provides the inclusive values for nodes in the upper level. However, the C+1th alternative, Next epoch, lacks the logsum parameter due to its degenerate branch. Consequently, the probability of this alternative can be determined using the following equation.

$$P(Next\ epoch) = \frac{\exp\left(V_{Next\ epoch}\right)}{\sum_i \exp(V_i + \Gamma_i * \theta_i) + \exp\left(V_{Next\ epoch}\right)} \quad (25)$$

The probability of F/I crashes in equation 22 can be obtained by substituting the value of $U_{k,i}$ from equation 17. Similarly, equation 23 gives the probability of upper-level alternatives, which is equivalent to equation 19 and can be rewritten using equation 26.

$$P(i) = \frac{exp\ (V_{e,i} + \Gamma_i * \theta_i)}{\sum_i \exp\left(V_{e,i} + \Gamma_i * \theta_i\right) + \exp(V_{e,Next\ epoch})} \cdot \prod_{e*=1}^{e-1} \frac{exp\ (V_{e*,C+1})}{\sum_i \exp\left(V_{e*,i} + \Gamma_i * \theta_i\right) + \exp(V_{e*,Next\ epoch})} \quad (26)$$

Assuming each row in the crash data after creation of forecasting epochs is represented using the superscript *n*, the log-likelihood function for the two-level nested logit model can be expressed as the sum of two components using equation 27. These components are associated with the lower and upper-level alternatives, respectively (Brownstone & Small, 1989).

$$L = \sum_n \log P^n(k^n|i^n) + \sum_n \log P^n(i^n) \quad (27)$$

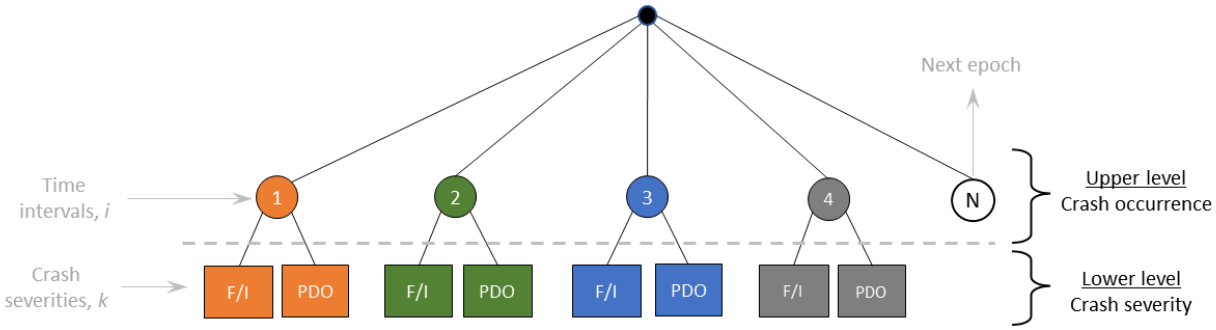


Figure 9 Two-level nested structure of crash occurrence and severity

Estimation of the nested logit model

There are several methods available for estimating parameters in nested logit models, with sequential estimation and simultaneous estimation being the most cited approaches. In sequential estimation, the first component of the log-likelihood function (equation 27) is maximized to estimate the parameters in the lower-level. This step provides estimates of the coefficients scaled by their respective nesting parameter θ_i. To simplify the process, the nesting parameters can be assumed to be constant for all nodes, represented as $\theta_i = \theta$. In the next step, inclusive values are calculated for each node using the scaled estimates obtained from the lower level. These inclusive values are then used in the second component of the log-likelihood function to maximize and obtain the values of θ and intercepts β_i for the upper level. It is important to note that while sequential estimation allows for the maximization and estimation of parameters in a stepwise manner, the estimates obtained are not consistent because the scaled parameters from the lower level are substituted to find parameters in the upper level. An alternative approach is simultaneous estimation, where parameters in both levels are estimated simultaneously using a non-linear maximization algorithm. This method is more rigorous compared to sequential estimation, and the estimates obtained are consistent.

Data

Data source and preparation

The data used in this study was collected from two primary sources. First, historical crash data for the year 2019 was obtained from the Enhanced Tennessee Roadway Information Management System (ETRIMS). This dataset provided information on various crash characteristics such as the date, time, severity, and coordinates of the crash location, as well as details on static covariates such as highway geometry, weather conditions, lighting conditions, land use, and terrain characteristics. The dynamic covariates for the study, namely traffic flow and speed, were obtained from the Radar Detection System (RDS) stations located along the highway segments from which the historical crash data was collected. Since our study aimed to implement a practical time discretization with 15-minute intervals, the RDS data was collected specifically for these 15-minute intervals. To match the RDS data with the corresponding crashes, a geospatial mapping approach was employed, aligning the RDS stations with their respective highway segments.

It is important to note that RDS coverage in Tennessee is limited to major cities. Therefore, for the purposes of this study, the segments of interstates within the city limits of Memphis and Chattanooga were considered. Specifically, the selected segments included I-40 and I-55 in Memphis, and I-24 and I-75 in Chattanooga.

Table 22
Summary of interstate segmentation

Interstate	City	Number of segments	Length (mi)	Number of crashes
I-40	Memphis	146	21.51	905
I-55		94	12.28	268
I-24	Chattanooga	48	14.71	675
I-75		70	13.29	527
Total		358	61.79	2,375

For this study, the interstates were divided into segments based on four criteria including the direction of traffic, number of lanes, posted speed limit, and terrain type. The segmentation details of the interstates are provided in **Table 22**. The table includes information on the total number of segments, their lengths in both directions, and the frequency of crashes observed within each segment. In total, the dataset consisted of 2,375 crashes.

Table 23 presents a breakdown of the crash frequencies based on various categorical variables. Additionally, the table includes descriptive statistics for the continuous variables in the dataset. The table provides a comprehensive overview of the data, highlighting the distribution of crashes across different segments and variable categories.

In this study, the 15-minute traffic volumes were scaled to a range between 0 (minimum value) and 1 (maximum value). This scaling process was applied to avoid the potential influence of larger volumes on the model training process. The duration between crashes exhibited a right-skewed distribution, as indicated by the mean of 516.67 hours (about 3 weeks) being greater than the median of 230.46 hours (about 1 and a half weeks). This suggests that there is a longer average time period between crashes, with occasional instances of shorter durations. A visual representation of the distribution of inter-crash duration for the four interstates is presented by a density plot in Figure 10. The density plot provides a graphical representation of the distribution, highlighting the shape and spread of the duration between crashes for each interstate.

Table 23

Descriptive statistics of crash characteristics

Categorical variables	Frequency of crashes	Relative abundance
Time of day		
Early morning (6 a.m. to 9 a.m.)	447	18.82%
Late morning (9 a.m. to 12 p.m.)	262	11.03%
Early afternoon (12 p.m. to 3 p.m.)	351	14.78%
Late afternoon (3 p.m. to 6 p.m.)	586	24.67%
Evening (6 p.m. to 12 a.m.)	392	16.51%
Night (12 a.m. to 6 a.m.)	337	14.19%
Weather condition		
Clear	1,733	72.97%
Others (Cloudy, rain, fog, or snow)	642	27.03%
Lighting condition		
Daylight	1,612	67.87%
Dark lighted	463	19.49%
Dark, not lighted	300	12.63%
Illumination type		
Illuminated	1,780	74.95%
Not illuminated	595	25.05%
Terrain		
Flat	715	30.11%
Rolling	1,660	69.89%
Land use		
Commercial	1,187	49.98%
Rural	765	32.21%
Mixed	423	17.81%
Crash severities		
Fatal or injury	451	18.99%
Property Damage Only	1,924	81.01%

Continuous variables	Min	Q1	Median	Q3	Max	Mean	SD
Traffic flow characteristics							
Speed (mph)	1.00	59.06	63.47	66.96	91.00	61.41	10.46
Volume (scaled between 0-minimum, and 1-maximum)	0.0002	0.12	0.28	0.46	1.00	0.31	0.22
Highway geometry							
Number of lanes (both directions)	3	6	8	8	12	7.18	1.78
Inter-crash duration (hours)	0	68	230	627	7683	516	783

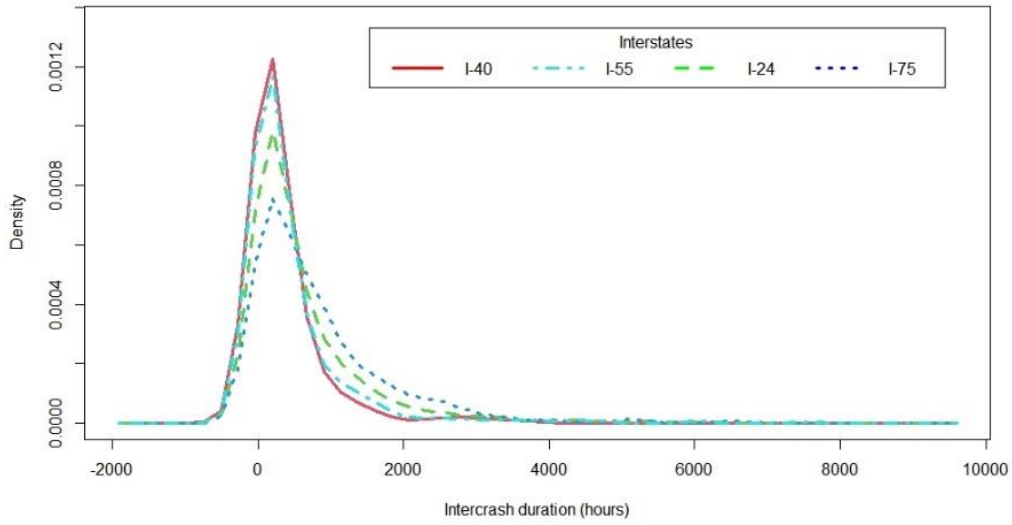


Figure 10 Distribution of inter crash duration for the interstates

From the plot, it can be observed that I-40 has the highest peak, indicating a higher concentration of crashes compared to the other interstates. Furthermore, the density plot reveals that the distribution of crashes on I-40 is less spread out compared to the other interstates. This means that the duration between crashes on I-40 is shorter, indicating a higher frequency of crashes occurring within a shorter period. In terms of increasing spread, the interstates can be ranked as follows: I-40, I-55, I-24, and I-75. This implies that the duration between crashes is longer and more spread out on I-75 compared to the other interstates.

Data sampling

In this study, the models were calibrated using training data and evaluated on testing data. The process of creating training and testing data involved splitting the historical crash data in a 9:1 ratio, where 90% of the data was allocated for training and the remaining 10% for testing. To create forecasting epochs, both the training and testing crashes were expanded. The training data was further sampled at 5% increments up to 25% to investigate whether any sample size below 25% would provide accurate parameter estimates. Thus, the samples used for parameter estimation were 5%, 10%, 15%, and 25% of the training data. This sampling approach is called

epoch level sampling (Thapa et al., 2022a). The sampled training data, along with the complete training data, were used to estimate the parameters for the models. For comparison purposes, the parameter estimates obtained from the complete training data (100% training data) were considered as the "true" estimates.

To evaluate the performance of the trained models, the predicted log-likelihood values were calculated on the training data. In this context, predicted log-likelihood provided a basis for comparing how well the models captured the characteristics of the training data.

Results

All model computations, including estimation and validation, in this study were conducted using R version 4.2.3 on a computer equipped with Intel Core i7-11700K processor and 16 GB of memory. We initially estimated the model parameters using the complete training data, employing both simultaneous and sequential estimation techniques. The objective of estimating with the complete training data was to obtain "true" parameter estimates and compare the results obtained from different estimation techniques. The estimation results are presented in Table 24. In the table, the first column displays the variable groups in the model, along with the corresponding variable categories considered as the base in the models. The second column lists the variables included in the model. The estimation results are then presented, showing the parameter estimates and their respective *t*-statistics for both simultaneous and sequential estimation. The parameter estimates obtained from both estimation methods are comparable, indicating consistency in the results. Additionally, the average values of predicted log-likelihood are also similar between the two methods. When considering estimation complexity, which refers to the time taken for the model to converge from a null model, it was found that sequential estimation offers a considerable advantage. Specifically, using simultaneous estimation, the

model took 51.09 hours (about 2 days) to converge, which was approximately six times the time taken by sequential estimation, which was 8.63 hours. Therefore, sequential estimation may provide consistent estimates with a significant reduction in computational complexity.

The parameters obtained from simultaneous estimation, as shown in the table, can be utilized to express the propensity function for F/I crashes in any time interval using the following utility equation.

$$U_{F/I,e,i} = V_i - 6.46 * t_{e,i} + 1.96 * Early\ morning + 3.50 * Late\ morning \ldots - 1.27 * Volume$$

For example, the utility equation for the first time-interval can be expressed as follows.

$$U_{F/I,e,1} = -8.71 - 6.46 * t_{e,1} + 1.96 * Early\ morning + 3.50 * Late\ morning \ldots - 1.27 * Volume$$

The analysis reveals interesting findings regarding the factors influencing F/I crashes. The duration dynamics coefficient indicates that as the duration between crashes increases, the likelihood of F/I crashes decreases. Moreover, F/I crashes are more likely to occur between 9 am and 3 pm. Clear weather conditions are associated with a higher likelihood of F/I crashes compared to adverse weather conditions such as clouds, rain, fog, or snow.

Dark lighted conditions result in more severe crashes, followed by daytime and dark unlighted conditions. Non-illuminated locations are more prone to F/I crashes compared to illuminated locations. Additionally, locations with flat terrain have a higher likelihood of F/I crashes compared to those with rolling terrain. Higher traffic volume leads to a decrease in F/I crashes, due to stop-and-go conditions during congested conditions. Similarly, higher speeds are associated with a lower likelihood of F/I crashes, although the effect size is small. The coefficients for the upper-level nodes, V_i, have similar magnitudes. The nesting parameter has a value of 4.36, indicating cross nesting of alternatives. It is worth noting that the training data increased significantly after the creation of forecasting epochs, with the original 2,137 crashes expanding to 1,103,104 observations.

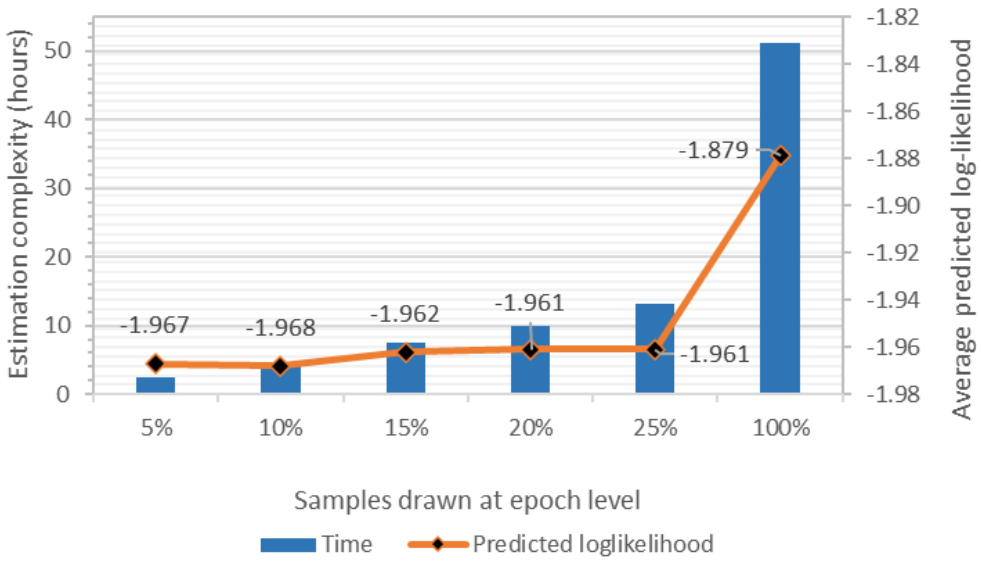
Table 24
Results from estimation using complete training data

Variable groups	Variables	Simultaneous estimation		Sequential estimation	
		Estimate	*t*-stat	Estimate	*t*-stat
Upper level					
Duration dynamics	Time since previous crash	-6.46	-22.60	-7.22	-91.82
Time of day (Evening 6 p.m. to 12 a.m., Night 12 a.m. to 6 a.m.)	Early morning (6 a.m. to 9 a.m.)	1.96	20.75	2.19	44.87
	Late morning (9 a.m. to 12 p.m.)	3.50	22.27	3.91	70.82
	Early afternoon (12 p.m. to 3 p.m.)	3.48	22.39	3.89	75.27
	Late afternoon (3 p.m. to 6 p.m.)	2.47	21.74	2.76	58.19
Weather conditions (Others)	Clear	1.92	22.22	2.15	72.37
Lighting condition	Daytime	0.52	9.79	0.58	10.80
Dark, not lighted)	Dark lighted	3.47	22.11	3.88	67.84
Illumination type (Not illuminated)	Illuminated	-0.88	-19.04	-0.99	-32.94
Terrain type (Rolling)	Flat	0.28	9.58	0.32	10.50
Land use	Commercial	-1.06	-18.82	-1.18	-31.78
(Mixed)	Rural	-2.01	-21.60	-2.24	-56.49
Highway geometry	Number of lanes	0.88	22.23	0.98	72.09
Traffic flow	Speed	-0.08	-23.20	-0.09	-506.52
characteristics	Volume	-1.27	-21.70	-1.42	-55.39
Lower level					
Intercepts (Next epoch)	First 15-min interval	-8.71	-131.18	-8.85	-128.09
	Second 15-min interval	-8.74	-130.92	-8.88	-127.94
	Third 15-min interval	-8.77	-130.53	-8.90	-127.68
	Fourth 15-min interval	-8.71	-131.35	-8.85	-128.26
Nesting coefficient	θ	4.36	23.22	4.87	24.80
Goodness of fit					
Number of observations (Training)		1,103,104			
Average initial LL		-213.98			
Average LL at convergence		-2.052			
Number of observations (Testing)		140,591			
Predicted LL		-1.879			
Estimation complexity	Time (hours)	51.09		8.63	

Table 25

Results from simultaneous model calibrated using epoch level samples

Variable groups	Variables	5% sample	10% sample	15% sample	20% sample	25% sample
Upper level						
Duration dynamics	Time since previous crash	-5.05 (-4.21)	-6.55 (-7.15)	-6.60 (-8.65)	-6.99 (-10.10)	-6.94 (-11.36)
Time of day (Evening 6 p.m. to 12 a.m., Night 12 a.m. to 6 a.m.)	Early morning (6 a.m. to 9 a.m.)	1.58 (3.93)	1.81 (6.39)	1.77 (7.73)	1.78 (8.97)	1.84 (10.13)
	Late morning (9 a.m. to 12p.m.)	2.56 (4.13)	3.20 (6.97)	3.43 (8.51)	3.40 (9.87)	3.45 (11.03)
	Early afternoon (12p.m. to 3p.m.)	2.46 (4.14)	3.18 (7.00)	3.60 (8.58)	3.76 (10.03)	3.70 (11.20)
	Late afternoon (3p.m. to 6p.m.)	1.71 (4.02)	2.40 (6.83)	2.48 (8.32)	2.50 (9.66)	2.52 (10.81)
Weather conditions (Others)	Clear	1.43 (4.12)	1.87 (7.00)	1.97 (8.51)	1.98 (9.91)	1.99 (11.08)
Lighting condition (Dark, not lighted)	Daytime	0.49 (2.27)	0.42 (2.50)	0.31 (2.37)	0.39 (3.33)	0.36 (3.53)
	Dark lighted	2.97 (4.15)	3.31 (6.92)	3.36 (8.41)	3.46 (9.81)	3.52 (11.06)
Illumination type (Not illuminated)	Illuminated	-0.66 (-3.60)	-0.90 (-5.98)	-0.80 (-6.99)	-0.72 (-7.74)	-0.75 (-8.81)
Terrain type (Rolling)	Flat	0.49 (3.22)	0.85 (5.89)	0.80 (7.05)	0.70 (7.65)	0.63 (8.15)
Land use (Mixed)	Commercial	-1.22 (-3.94)	-1.28 (-6.27)	-1.16 (-7.40)	-0.98 (-8.05)	-0.99 (-9.01)
	Rural	-2.03 (-4.14)	-2.15 (-6.87)	-2.08 (-8.28)	-1.98 (-9.54)	-1.96 (-10.63)
Highway geometry	Number of lanes	0.77 (4.19)	0.96 (7.09)	0.86 (8.48)	0.89 (9.89)	0.89 (11.11)
Traffic flow characteristics	Speed	-0.07 (-4.29)	-0.08 (-7.34)	-0.08 (-8.86)	-0.08 (-10.32)	-0.08 (-11.55)
	Volume	-14.94 (-4.26)	-18.77 (-7.26)	-17.74 (-8.77)	-17.98 (-10.23)	-18.02 (-11.47)
Lower level						
Intercepts (Next epoch)	First 15-min interval	-8.68 (-28.18)	-8.94 (-39.92)	-8.89 (-49.03)	-8.90 (-56.71)	-8.85 (-63.84)
	Second 15-min interval	-9.01 (-27.42)	-8.61 (-42.27)	-8.56 (-51.90)	-8.59 (-59.88)	-8.62 (-66.90)
	Third 15-min interval	-8.56 (-28.57)	-8.83 (-40.44)	-8.77 (-49.69)	-8.75 (-57.60)	-8.83 (-64.04)
	Fourth 15-min interval	-8.60 (-28.37)	-8.73 (-40.63)	-8.77 (-49.53)	-8.80 (-57.19)	-8.70 (-64.55)
Nesting coefficient	θ	3.68 (4.30)	4.49 (7.35)	4.42 (8.87)	4.46 (10.34)	4.45 (11.58)
Goodness of fit						
Number of observations (Training)		55,062	110,187	165,378	220,662	275,787
Average initial LL		-212.70	-212.87	-212.94	-213.01	-213.08
Average LL at convergence		-2.052	-2.053	-2.054	-2.053	-2.051
Number of observations (Testing)				140,591		
Predicted log-likelihood		-1.967	-1.968	-1.962	-1.961	-1.961
Estimation complexity	Time (hours)	2.48	4.22	7.49	9.86	13.25



Next, we proceeded to estimate parameters using sampled data to explore the tradeoff between model performance and estimation complexity. The results of this estimation can be found in Table 25, which presents the obtained parameter values along with their respective *t*-statistics. Upon visual inspection, it is apparent that the parameter values obtained using the 25% sample are much closer to the true values compared to the 5% sample. This finding aligns with a previous study conducted by (Thapa et al., 2022a). However, it is also crucial to investigate the impact of sample size within the range of 5% to 25% to determine the sample that offers the optimal balance between model performance and estimation complexity. To address this, we estimated parameters at 5% increments, ranging from 5% to 25%. **Error! Not a valid bookmark self-reference.** presents a graphical representation of estimation complexity and predicted log-likelihood for the various samples. Notably, the figure indicates a significant improvement in prediction performance beyond the 10% sample. Furthermore, the models demonstrate similar performance for the 15%, 20%, and 25% samples.

Figure 11 Improvement in model performance with increase in data size

As expected, estimation complexity increases linearly with the sample size. For instance, the model required 2.48 hours to train on the 5% sample, while it took approximately 20 times or

51.09 hours (about 2 days) for the full 100% dataset. Based on the findings depicted in the figure, it is evident that using a 15% sample can yield comparable estimates and predictive performance to the 25% sample, while reducing the estimation complexity to 60% of that offered by the 25% sample. This suggests that the 15% sample size strikes a favorable balance between model performance and estimation complexity.

Effect of sampling on coefficients

Based on the parameter estimates, it is evident that certain predictor variables are particularly sensitive to sampling. A notable example is the *Volume* variable, where the coefficients exhibit significant differences between the sampled data and the complete data (refer to Figure 12). This discrepancy can be attributed to the sampling approach and the scaling of traffic volumes. Since the volumes are scaled between 0 and 1, random sampling can lead to the exclusion of several observations, resulting in considerable variations in the parameter estimates for this variable. On the other hand, coefficients for the *Speed* variable demonstrate consistency. This consistency may be attributed to the fact that the values of the variable do not fluctuate significantly, as indicated by its descriptive statistics, and are less affected by sampling.

Considering these observations, we aim to identify and report variables whose coefficients are either underestimated or overestimated due to sampling. To visualize this, a bar plot in Figure 12 presents the variable coefficients obtained from the sampled and complete data. From the plot, it can be observed that smaller samples are more likely to overestimate the effect of the following variables: *Time since crash*, *Time of day-Early afternoon* and *Late afternoon*, *Terrain-Flat*, *Land Use-Commercial*, *Number of lanes*, and *Volume*. Conversely, variables such as *Time of day-Early Morning* and *Late Morning*, *Lighting-Daytime*, and *Lighting-Dark lighted* are more likely to be underestimated when smaller samples are used. Overall, these findings

emphasize the importance of considering the impact of sampling on parameter estimates, particularly for variables that exhibit sensitivity to sampling.

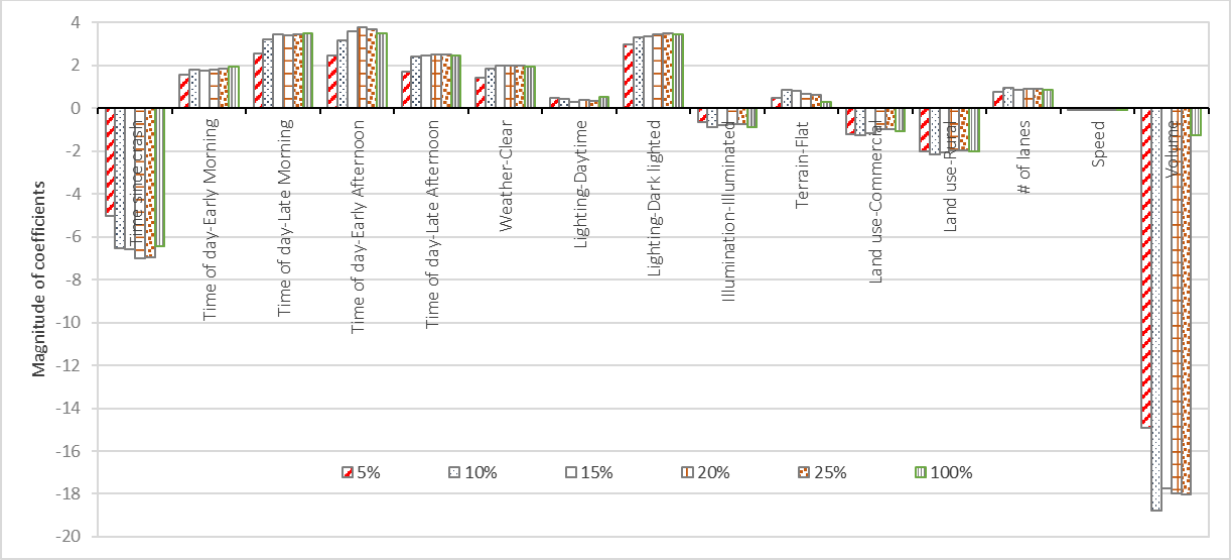


Figure 12 Coefficient of variables for different training samples

Variables whose coefficients are likely to converge towards the true value with an increase in sample size have also been identified. This identification is crucial from a practical standpoint, especially when analysts and planners seek greater accuracy for specific variables. In the following figures, we present two groups of predictors. The first group consists of variables whose coefficients diverge further away from the actual values as the sample size increases. These variables would require larger samples to achieve more accurate estimation. It is important to recognize the limitations in estimating the coefficients for these variables with smaller sample sizes. The second group comprises predictors whose coefficients converge closer to the actual values as the sample size increases. This group includes variables whose coefficients can be obtained with reasonable accuracy, even with small increments in sample size. The results suggest that the estimation of these coefficients becomes more stable and reliable as the sample size grows. These findings serve as valuable insights for researchers and practitioners, allowing

them to prioritize their data collection efforts and allocate resources effectively based on the sensitivity of different predictors to sample size.

The variables whose coefficients diverge despite an increase in sample size, ranging from 5% to 25%, compared to the full data are *Time since crash*, *Time of day-Early afternoon*, *Weather Condition-Clear*, *Lighting Condition-Daytime*, *Illumination-Illuminated*, and *Volume*. These variables are presented in Figure 13, indicating the percentage difference of the coefficients from the complete training data. On the other hand, coefficients for *Time of day-Early morning*, *Time of day-Late morning*, *Time of day-Late afternoon*, *Lighting-Dark lighted*, *Terrain-Flat*, *Land Use-Commercial*, *Land Use-Rural*, *Number of lanes*, and *Speed* converge quicker to the actual values as the sample size increases. These variables are displayed in Figure 14, illustrating the percentage difference compared to the complete training data. These findings highlight the sensitivity of different variables to sample size and provide valuable insights into the accuracy and stability of their coefficient estimates.

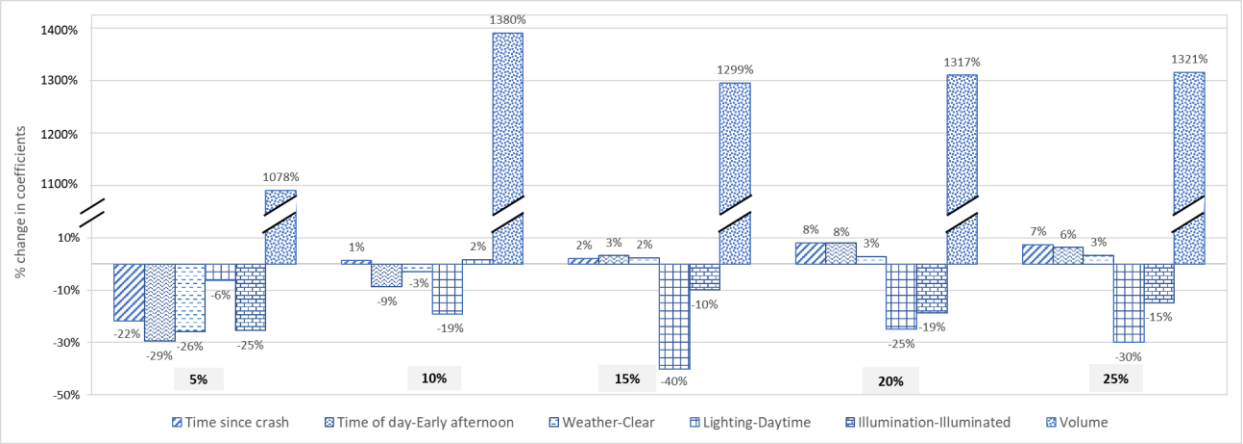


Figure 13 Variable coefficients diverging away from their actual values

Validation

The validation of the proposed nested logit model was carried out to assess its predicted capabilities. All validations were conducted using the simultaneous model trained on 15% data

drawn at the epoch level since our analysis suggested that it provided the best tradeoff between accuracy and estimation complexity. As discussed previously, 10% of the sample was held out for testing. The test sample consisted of 236 crashes, including 39 F/I crashes and 197 PDO crashes. This test sample was used for validation. Similar to the two-step model, validation was conducted to assess predictive abilities for the outcomes considered at the lower and upper levels. These results are discussed in the following subsections.

Upper level: Crashes at epoch level

One of the primary objectives of the proposed framework is to predict the occurrence of future crashes. Therefore, it is crucial to evaluate the temporal accuracy of the predicted crashes. To evaluate this, we measured the proximity between the predicted crash epoch and the actual epoch at which crashes were observed, by introducing a metric called Predicted Temporal Proximity (PTP), represented by equation 28. This metric quantifies how closely the predicted crash epochs align with the observed epochs.

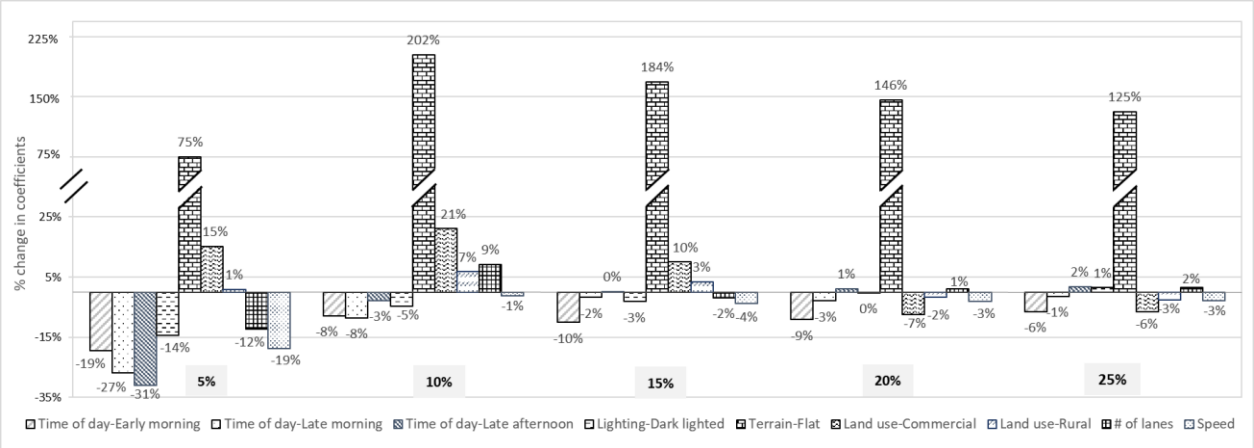


Figure 14 Variable coefficients converging closer to their actual values

Furthermore, we also investigated whether the number of epochs impacted the model's performance in terms of PTP. To accomplish this, we calculated the PTP for different subsets of the testing data by excluding crashes with a substantial number of epochs. This was

accomplished by creating subsets of the test data to include crashes with fewer than 100 to 1000 epochs, with intervals of 100 epochs. The average values of PTP for these subsets of testing data are depicted in Figure 15.

$$PTP = \left|\frac{Predicted\ crash\ epoch - Actual\ crash\ epoch}{Actual\ crash\ epoch}\right| * 100\% \qquad (28)$$

It is important to note that, according to the definition of PTP, a smaller value is desired as it indicates that the predicted crash epoch is closer to the observed epoch. The results depicted in Figure 15 indicate that when there is a substantial number of epochs (i.e., a large intercrash duration), the value of PTP increases. This suggests that epoch-level prediction is more accurate when the duration between crashes is smaller. In other words, the prediction of crash epochs is more reliable for highway segments that experience crashes more frequently. For example, based on the figure, for crashes with intercrash durations less than 100 hours (approximately 4 days), the predicted crash epoch is within 60% of the actual epoch, compared to 74% for durations exceeding 1,000 hours.

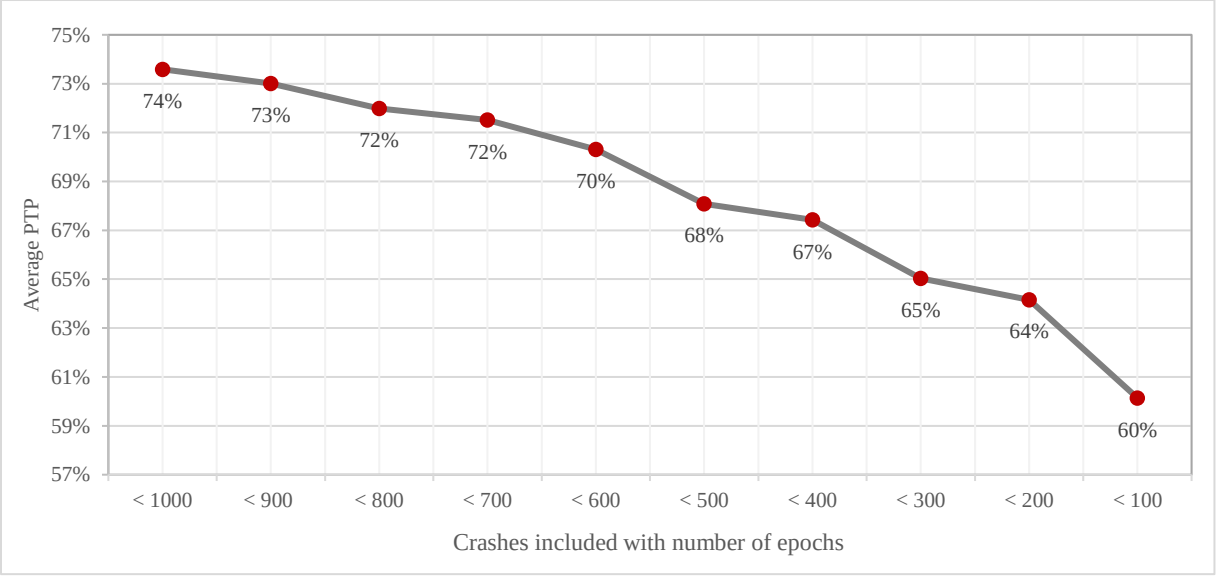


Figure 15 Average PTP for different subsets of test samples

Upper level: Crashes in predicted time-intervals

The accuracy of predicting crash occurrences at specific time intervals can be assessed from two perspectives: i) the accuracy of predicting crashes (true positives), and ii) the accuracy of predicting 'no crashes' (true negatives). Therefore, we relied on the metrics of Specificity and Sensitivity to evaluate the model's predictions. Specificity measures the model's ability to correctly predict 'no crashes' (true negatives) and is defined by equation 29. On the other hand, Sensitivity measures the model's ability to correctly predict crashes (true positives) and is defined by equation 30. It quantifies the proportion of correctly identified positive cases in relation to the actual positive cases. It quantifies the proportion of correctly identified negative cases in relation to the actual negative cases.

The model's prediction accuracy for crash and severity were evaluated using these metrics. The results are summarized in Table 26 and described as follows.

$$Specificity = \frac{True\ Negatives\ (TN)}{True\ Negatives\ (TN) + False\ Positives\ (FP)} \quad (29)$$

$$Sensitivity = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \quad (30)$$

Table 26
Values of Specificity and Sensitivity from model predictions

Predictions	TN	TP	FP	FN	Specificity	Sensitivity
Crash occurrence	539	63	173	169	0.76	0.27
Crash severity						
F/I crashes	887	9	20	28	0.97	0.24
PDO crashes	557	41	192	154	0.74	0.21

The model predictions for the time intervals resulted in the following counts: True Negatives (TN) = 539, True Positives (TP) = 63, False Positives (FP) = 173, and False Negatives (FN) = 169. The Specificity is calculated to be 0.76, indicating a high value. This high value suggests a low rate of false positive predictions. Therefore, the model demonstrates reliability in

predicting crashes. In other words, the likelihood of classifying a time interval without a crash as a time interval experiencing a crash is low. On the other hand, the Sensitivity is calculated to be 0.27, indicating a low value. This low value suggests a high rate of false negatives, or in other words, the chances of classifying true crash intervals as having no crash is high.

Lower level: Crash severity for crashes in predicted time-intervals

The Specificity and Sensitivity measures were also utilized to evaluate the model's ability to predict crash severities at each time interval. For F/I crashes, the following results were obtained: TN = 887, TP = 9, FP = 20, FN = 28, resulting in a Specificity of 0.97 and a Sensitivity of 0.24. Similarly, for PDO crashes, the values obtained were TN = 557, TP = 41, FP = 192, and FN = 154, with a Specificity of 0.74 and a Sensitivity of 0.21.

The results indicate that for both severity types, the Specificity values are high. This suggests that the model is capable of reliably predicting both F/I and PDO crashes with a lower chance of false positive predictions. However, it should be noted that the model also exhibits low Sensitivity values, indicating that the model may not always accurately classify the severity types with a high degree of certainty, leading to a higher occurrence of false negative predictions. This outcome is the result of exceptionally higher prevalence of time-intervals without crashes (0s) in comparison to those with crashes (1s). Future research can improve upon the model by addressing this imbalance in the frequency of outcomes.

Conclusion

This study developed a duration-based model to predict crash occurrence and severity using historical crash and traffic flow data from four interstates in Tennessee. The framework involved the reformulation of crash data to create forecasting epochs and time-intervals, which were used to calculate crash and severity likelihoods. The creation of forecasting epochs significantly

increased the data size and estimation complexity. Additionally, the adoption of a nested structure further contributed to the complexity of model estimation. To address the computational challenges, we suggested sampling the data at the epoch level to reduce estimation complexity. We aimed to find the optimal sampling strategy by considering the tradeoff between model performance and estimation complexity. After evaluating various samples, we determined that a 15% sample drawn at the epoch level provided the best balance in reducing data size. Furthermore, we investigated the impact of sampling on the coefficients of predictor variables to identify those most sensitive to changes in sample sizes. Variables such as *Time since crash, Time of day-Early afternoon, Late afternoon, Terrain-Flat, Land Use-Commercial, Number of lanes,* and *Volume* were found to be more likely to be overestimated by smaller samples. Conversely, variables including *Time of day-Early Morning, Late Morning, Lighting-Daytime* and *Dark lighted* were more likely to be underestimated.

When investigating the stability of coefficients for the predictors, it was found that *Time since crash, Time of day-Early afternoon, Weather Condition-Clear, Lighting Condition-Daytime, Illumination-Illuminated,* and *Volume* exhibited a higher degree of instability. Consistent estimation of these coefficients required larger sample sizes. On the other hand, coefficients for *Time of day-Early morning, Late morning, Late afternoon, Lighting-Dark lighted, Terrain-Flat, Land Use-Commercial* and *Rural, Number of lanes,* and *Speed* demonstrated a tendency to converge towards true estimates with incremental increases in sample size. These findings are crucial for obtaining consistent and reliable estimates when utilizing samples for model estimation and clarify the challenges and considerations associated with implementing the duration-based model, including the impact of data sampling on estimation outcomes and the sensitivity of certain variables to changes in sample sizes.

The proposed framework's validation provided satisfactory results. The measure, Predicted Temporal Proximity (PTP), suggests that the model performs better when implemented on segments where crashes are more frequent. For context, the model, trained on a 15% epoch-level sample, was able to predict crashes within 60% (i.e., average PTP=60%) of the actual epoch for crashes occurring within 100 epochs, or approximately 4 days of each other. On the contrary, the average value of PTP was 74% for crashes occurring within 1,000 epochs of each other. This finding also sheds light on the practical implications of the model, as it is often impractical to predict crashes too far into the future due to potential changes in traffic, weather, and driving conditions. Similarly, the estimated model displayed a satisfactory value of Specificity, indicating a low rate of false positives. In other words, the model is less likely to falsely predict time intervals without crashes as having experienced crashes. This is particularly important as a reasonable degree of certainty is desired to ensure effective allocation of limited safety resources to critical segments. The value of Sensitivity was comparatively smaller, implying a higher rate of false negatives or missed detections. However, it should also be noted that the frequency of time intervals without crashes is several multiples larger than the frequency of time intervals with crashes (preponderance of 0s compared to 1s). Therefore, the low value of Sensitivity is expected in this case.

Future research offers opportunities for notable improvements to the proposed model. Firstly, it would be valuable to investigate alternative nesting structures to determine if they provide a better fit, especially considering that the nesting parameter suggests the presence of alternative nests. More complex nesting structures based on distinct categories such as time of day, weather conditions, and other relevant factors could be explored. Secondly, the model estimates could be enhanced by incorporating random effects. Since the reformulated data, after

the creation of forecasting epochs, takes the form of panel data with repeated observations for crashes and road segments, accounting for segment and crash-specific heterogeneity could lead to more accurate model estimates. Furthermore, data balancing techniques such as Synthetic Minority Over-sampling Technique can be used to balance the frequency of outcomes and study its impact on model estimates. Finally, alternative estimation techniques leveraging parallel and distributed computing can be implemented to reduce estimation time while still retaining information from complete training dataset. Addressing these limitations would contribute to a more comprehensive understanding of crash prediction and severity estimation and improve the accuracy and applicability of the model in real-world scenarios.

Acknowledgements

This research was partially supported by Fulbright Fellowship to second author at Indian Institute of Technology (IIT) Bombay and the Center for Transportation Innovations in Education and Research (C-TIER) at the University of Memphis. Any findings and opinions expressed in this paper are those of the authors and do not necessarily reflect the view of C-TIER.

References

2019 Highway Work Zone Safety Survey. (2019). Associated General Contractors of America. https://www.agc.org/news/2019/05/23/2019-highway-work-zone-safety-survey

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

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Abdel-Aty, M., & Pande, A. (2007). Crash data analysis: Collective vs. Individual crash level approach. *Journal of Safety Research*, *38*(5), 581–587. https://doi.org/10.1016/j.jsr.2007.04.007

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record: Journal of the Transportation Research Board*, *1897*(1), 88–95. https://doi.org/10.3141/1897-12

Afghari, A. P., Haque, M. M., & Washington, S. (2020). Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. *Accident Analysis & Prevention*, *144*, 105615. https://doi.org/10.1016/j.aap.2020.105615

Agent, K. R., & Hibbs, J. O. (1996). *Evaluation of SHRP Work Zone Safety Devices*. 24.

Ahmed, S. S., Cohen, J., & Anastasopoulos, P. Ch. (2021). A correlated random parameters with heterogeneity in means approach of deer-vehicle collisions and resulting injury-severities. *Analytic Methods in Accident Research*, *30*, 100160. https://doi.org/10.1016/j.amar.2021.100160

Al-Ghamdi, A. S. (2002). Pedestrian–vehicle crashes and analytical techniques for stratified contingency tables. *Accident Analysis & Prevention*, *34*(2), 205–214. https://doi.org/10.1016/S0001-4575(01)00015-X

Algomaiah, M., & Li, Z. (2022). Enhancing Work Zone Capacity by a Cooperative Late Merge System Using Decentralized and Centralized Control Strategies. *Journal of Transportation Engineering, Part A: Systems*, *148*(2). https://doi.org/10.1061/JTEPBS.0000632

Ali, Y., Haque, M. M., Zheng, Z., Washington, S., & Yildirimoglu, M. (2019). A hazard-based duration model to quantify the impact of connected driving environment on safety during mandatory lane-changing. *Transportation Research Part C: Emerging Technologies*, *106*(June), 113–131. https://doi.org/10.1016/j.trc.2019.07.015

Ariannezhad, A., Karimpour, A., Qin, X., Wu, Y.-J., & Salmani, Y. (2021). Handling Imbalanced Data for Real-Time Crash Prediction: Application of Boosting and Sampling Techniques. *Journal of Transportation Engineering, Part A: Systems*, *147*(3), 04020165. https://doi.org/10.1061/JTEPBS.0000499

Bagloee, S. A., & Asadi, M. (2016). Crash analysis at intersections in the CBD: A survival analysis model. *Transportation Research Part A: Policy and Practice*, *94*, 558–572. https://doi.org/10.1016/j.tra.2016.10.019

Barua, S., El-Basyouny, K., & Islam, Md. T. (2016). Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic Methods in Accident Research*, *9*, 1–15. https://doi.org/10.1016/j.amar.2015.11.002

Baruya, A. (1998). Road Safety in Europe. *9th Internation Conference: Road Safety in Europe*.

Bashir, S., & Zlatkovic, M. (2021). Assessment of Queue Warning Application on Signalized Intersections for Connected Freight Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(10), 1211–1221. https://doi.org/10.1177/03611981211015247

Benekohal, R. F., Hajbabaie, A., Medina, J. C., Wang, M.-H., & Chitturi, M. V. (2010). *SPEED PHOTO-RADAR ENFORCEMENT EVALUATION IN ILLINOIS WORK ZONES* (FHWA-ICT-10-064). Illinois Department of Transportation.

Berthaume, A. L. (2015). *Microscopic Modeling of Driver Behavior Based on Modifying Field Theory for Work Zone Application* [Doctoral Dissertation, University of Massachusetts Amherst]. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1328&context=dissertations_2

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2011). Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. *2011 World Congress on Information and Communication Technologies*, 1241–1246. https://doi.org/10.1109/WICT.2011.6141426

Bham, G. H., & Mohammadi, M. A. (2011). *Evaluation of Work Zone Speed Limits: An Objective and Subjective Analysis of Work Zones in Missouri Report*. 92.

Brownstone, D., & Small, K. A. (1989). Efficient Estimation of Nested Logit models. *Journal of Business & Economic Statistics*, *7*(1), 67–74. https://doi.org/10.1080/07350015.1989.10509714

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., & Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. *Transportation Research Part C: Emerging Technologies*, *117*, 102697. https://doi.org/10.1016/j.trc.2020.102697

Cerwick, D. M., Gkritza, K., Shaheed, M. S., & Hans, Z. (2014). A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, *3–4*, 11–27. https://doi.org/10.1016/j.amar.2014.09.002

Cestac, J., Paran, F., & Delhomme, P. (2011). Young drivers' sensation seeking, subjective norms, and perceived behavioral control and their roles in predicting speeding intention: How risk-taking motivations evolve with gender and driving experience. *Safety Science*, *49*(3), 424–432. https://doi.org/10.1016/j.ssci.2010.10.007

Chang, H. L., & Jovanis, P. P. (1990). Formulating accident occurrence as a survival process. *Accident Analysis and Prevention*, *22*(5), 407–419. https://doi.org/10.1016/0001-4575(90)90037-L

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, *43*(8), 541–557. https://doi.org/10.1016/j.ssci.2005.04.004

Chang, Y., & Edara, P. (2018). Predicting hazardous events in work zones using naturalistic driving data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-March*, 1–6. https://doi.org/10.1109/ITSC.2017.8317847

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, *90*, 128–139. https://doi.org/10.1016/j.aap.2016.02.011

Cheng, W., Gill, G. S., Dasu, R., Xie, M., Jia, X., & Zhou, J. (2017). Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis & Prevention*, *99*, 330–341. https://doi.org/10.1016/j.aap.2016.11.022

Cheng, Z., Lu, J., Zu, Z., & Li, Y. (2019). Speeding Violation Type Prediction Based on Decision Tree Method: A Case Study in Wujiang, China. *Journal of Advanced Transportation*, *2019*, 1–10. https://doi.org/10.1155/2019/8650845

Choudhary, P., & Velaga, N. R. (2020). Impact of distraction on decision making at the onset of yellow signal. *Transportation Research Part C: Emerging Technologies*, *118*(March 2019), 102741. https://doi.org/10.1016/j.trc.2020.102741

Chung, Y. (2010). Development of an accident duration prediction model on the Korean Freeway Systems. *Accident Analysis and Prevention*, *42*(1), 282–289. https://doi.org/10.1016/j.aap.2009.08.005

Data USA: Highway Maintenance Workers. (2018). https://datausa.io/profile/soc/highway-maintenance-workers

Debnath, A. K., Blackman, R., & Haworth, N. (2015). Common hazards and their mitigating measures in work zones: A qualitative study of worker perceptions. *Safety Science*, *72*, 293–301. https://doi.org/10.1016/j.ssci.2014.09.022

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113

Dissanayake, S., & Akepati, S. R. (2009). *Identification of Work Zone Crash Characteristics*. Federal Highway Administration. https://intrans.iastate.edu/app/uploads/2018/08/Dissanayake_WZCrashChar.pdf

Dissanayake, S., & Lu, J. (2002). Analysis of Severity of Young Driver Crashes: Sequential Binary Logistic Regression Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, *1784*(1), 108–114. https://doi.org/10.3141/1784-14

Dong, C., Clarke, D. B., Yan, X., Khattak, A., & Huang, B. (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate

crash frequencies at intersections. *Accident Analysis & Prevention*, *70*, 320–329. https://doi.org/10.1016/j.aap.2014.04.018

Elliott, M. A., & Thomson, J. A. (2010). The social cognitive determinants of offending drivers' speeding behaviour. *Accident Analysis & Prevention*, *42*(6), 1595–1605. https://doi.org/10.1016/j.aap.2010.03.018

Eseonu, C., Gambatese, J., & Nnaji, C. (2018). *Reducing Highway Fatalities Through Improved Adoption of Safety Technologies*.

Federal Highway Administration. (2009a). *Manual of Traffic Control Devices for Streets and Highways*.

Federal Highway Administration. (2009b). *Manual on Uniform Traffic Control Devices (MUTCD)*. https://mutcd.fhwa.dot.gov/

Federal Highway Administration. (2023). *FHWA Work Zone Facts and Statistics*. Work Zone Management Program. https://ops.fhwa.dot.gov/wz/resources/facts_stats.htm

Flannagan, C. A., Selpi, Baykas, P. B., Leslie, A., Kovaceva, J., & Thomson, R. (2019). *Analysis of SHRP2 Data to Understand Normal and Abnormal Driving Behavior in Work Zones* (FHWA-HRT-20-010). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/48835

Forward, S. E. (2009). The theory of planned behaviour: The role of descriptive norms and past behaviour in the prediction of drivers' intentions to violate. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 198–207. https://doi.org/10.1016/j.trf.2008.12.002

Fountas, G., & Anastasopoulos, P. Ch. (2017). A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. *Analytic Methods in Accident Research, 15*, 1–16. https://doi.org/10.1016/j.amar.2017.03.002

Furth, P. G. (2011). *Sampling and Estimation Techniques for Estimating Bus System Passenger-Miles*. Bureau of Transportation Statistics. https://www.bts.gov/archive/publications/journal_of_transportation_and_statistics/volume_08_number_02/paper_07/index

Furth, P. G., Killough, K. L., & Ruprecht, G. F. (1988). Cluster Sampling Techniques for Estimating Transit Patronage. *Transportation Research Record, 1165*.

Gambatese, J. A., Lee, H. W., & Nnaji, C. A. (2017). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance*. Oregon State University School of Civil and Construction Engineering.

Gambatese, J., & Lee, H. W. (2016). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance II*. (Issue 503).

Gan, H., Wei, J., & Wang, G. (2021). A generic work zone evaluation tool driven by a macroscopic traffic simulation model. *International Journal of Mobile Communications*, *19*(1), 1. https://doi.org/10.1504/IJMC.2021.111884

Garber, N. J., & Ehrhart, A. A. (2000). Effect of Speed, Flow, and Geometric Characteristics on Crash Frequency for Two-Lane Highways. *Transportation Research Record: Journal of the Transportation Research Board*, *1717*(1), 76–83. https://doi.org/10.3141/1717-10

Gelman, A., & Hill, J. (2007). When does a multilevel modeling make a difference? In *Data Analysis Using Regression and Multilevel/Hierarchical Models* (pp. 237–249). Cambridge University Press.

Golob, T. F., Recker, W. W., & Leonard, J. D. (1987). An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention*, *19*(5), 375–395. https://doi.org/10.1016/0001-4575(87)90023-6

Guo, H., Wang, W., Guo, W., & Zhao, F. (2013). Modeling lane-keeping behavior of bicyclists using survival analysis approach. *Discrete Dynamics in Nature and Society*, *2013*. https://doi.org/10.1155/2013/197518

Hamdar, S. H., Khoury, H., & Zehtabi, S. (2016). A simulator-based approach for modeling longitudinal driving behavior in construction work zones: Exploration and assessment. *SIMULATION*, *92*(6), 579–594. https://doi.org/10.1177/0037549716644515

Haque, K., Mishra, S., & Golias, M. M. (2021). Multi-period transportation network investment decision making and policy implications using econometric framework. *Research in Transportation Economics*, *89*, 101109. https://doi.org/10.1016/j.retrec.2021.101109

Haque, M. M., & Washington, S. (2015). The impact of mobile phone distraction on the braking behaviour of young drivers: A hazard-based duration model. *Transportation Research Part C: Emerging Technologies*, *50*, 13–27. https://doi.org/10.1016/j.trc.2014.07.011

Harb, R., Radwan, E., Yan, X., Pande, A., & Abdel-Aty, M. (2008). Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression. *Journal of Transportation Engineering*, *134*(5), 203–214. https://doi.org/10.1061/(ASCE)0733-947X(2008)134:5(203)

Harmon, T., Bahar, G., & Gross, F. (2018). *Crash Costs for Highway Safety Analysis*.

Hashmienejad, S. H. A., & Hasheminejad, S. M. H. (2017). Traffic accident severity prediction using a novel multi-objective genetic algorithm. *International Journal of Crashworthiness*, *22*(4), 425–440. https://doi.org/10.1080/13588265.2016.1275431

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second). John Wiley & Sons, Inc.

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019a). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019b). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

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

Hou, G., & Chen, S. (2019). An Improved Cellular Automaton Model for Work Zone Traffic Simulation Considering Realistic Driving Behavior. *Journal of the Physical Society of Japan*, *88*(8), 084001. https://doi.org/10.7566/JPSJ.88.084001

Hourdos, J. (2012). Portable, Non-Intrusive Advance Warning Devices for Work Zones with or without Flag Operators. *Minnesota Department of Transportation*, *October*.

Imprialou, M. I. M., Quddus, M., Pitfield, D. E., & Lord, D. (2016). Re-visiting crash-speed relationships: A new perspective in crash modelling. *Accident Analysis and Prevention*, *86*, 173–185. https://doi.org/10.1016/j.aap.2015.10.001

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Jonathan, A.-V., Wu, K.-F. (Ken), & Donnell, E. T. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention*, *87*, 8–16. https://doi.org/10.1016/j.aap.2015.11.006

Jovanis, P. P., & Chang, H. L. (1989). Disaggregate model of highway accident occurrence using survival theory. *Accident Analysis and Prevention*, *21*(5), 445–458. https://doi.org/10.1016/0001-4575(89)90005-5

Jovanović, D., Šraml, M., Matović, B., & Mićić, S. (2017). An examination of the construct and predictive validity of the self-reported speeding behavior model. *Accident Analysis & Prevention*, *99*, 66–76. https://doi.org/10.1016/j.aap.2016.11.015

Jung, S., Qin, X., & Noyce, D. A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis & Prevention*, *42*(1), 213–224. https://doi.org/10.1016/j.aap.2009.07.020

Kashyap, A. A., Raviraj, S., Devarakonda, A., Nayak K, S. R., K V, S., & Bhat, S. J. (2022). Traffic flow prediction models – A review of deep learning techniques. *Cogent Engineering*, *9*(1), 2010510. https://doi.org/10.1080/23311916.2021.2010510

Ke, J., Zhang, S., Yang, H., & Chen, X. (Michael). (2019). PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. *Transportmetrica A: Transport Science*, *15*(2), 872–895. https://doi.org/10.1080/23249935.2018.1542414

Keramati, A., Lu, P., Zhou, X., & Tolliver, D. (2020). A Simultaneous Safety Analysis of Crash Frequency and Severity for Highway-Rail Grade Crossings: The Competing Risks Method. *Journal of Advanced Transportation*, *2020*(1). https://doi.org/10.1155/2020/8878911

Khasnabis, S., Mishra, S., & Safi, C. (2012). Evaluation procedure for mutually exclusive highway safety alternatives under different policy objectives. *Journal of Transportation Engineering*, *138*(7), 940–948. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000397

Khattak, A. J., Khattak, A. J., & Council, F. M. (2002). Effects of work zone presence on injury and non-injury crashes. *Accident Analysis and Prevention*, *34*(1), 19–29. https://doi.org/10.1016/S0001-4575(00)00099-3

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, & W. Wong, Eds.; Thid Editi). Springer. https://doi.org/10.1007/978-1-4419-6646-9

Kloeden, C. N., McLean, J., & Glonek, G. F. V. (2002). *Reanalysis of travelling speed and the risk of crash involvement in Adelaide South Australia*. Australian Transport Safety Bureau.

Kock, N., & Lynn, G. S. (2012). Lateral Collinearity and Misleading Results in Variance-Based SEM : An Illustration and Recommendations Lateral Collinearity and Misleading Results in Variance-. *Journal of the Association for Information Systems*, *13*(7), 546–580.

Lee, C., Hellinga, B., & Saccomanno, F. (2003). Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, *1840*(1), 67–77. https://doi.org/10.3141/1840-08

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *Applied Sciences*, *10*(1), 129. https://doi.org/10.3390/app10010129

Lee, J.-T., & Fazio, J. (2005). Influential Factors in Freeway Crash Response and Clearance Times by Emergency Management Services in Peak Periods. *Traffic Injury Prevention*, *6*(4), 331–339. https://doi.org/10.1080/15389580500255773

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis & Prevention*, *135*, 105371. https://doi.org/10.1016/j.aap.2019.105371

Li, Y., & Bai, Y. (2008). Development of crash-severity-index models for the measurement of work zone risk levels. *Accident Analysis and Prevention*, *40*(5), 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012

Li, Y., & Bai, Y. (2009). Highway work zone risk factors and their impact on crash severity. *Journal of Transportation Engineering*, *135*(10), 694–701. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000055

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident Analysis and Prevention*, *111*(November 2017), 354–363. https://doi.org/10.1016/j.aap.2017.11.028

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

Ma, J., & Kockelman, K. (2006a). Crash frequency and severity modeling using clustered data from Washington state. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, October*, 1621–1626. https://doi.org/10.1109/itsc.2006.1707456

Ma, J., & Kockelman, K. M. (2006b). Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record: Journal of the Transportation Research Board*, *1950*, 24–34.

Ma, J., Kockelman, K. M., & Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*, *40*(3), 964–975. https://doi.org/10.1016/j.aap.2007.11.002

Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research*, *25*, 100113. https://doi.org/10.1016/j.amar.2020.100113

Marks, E., Vereen, S., & Awolusi, I. (2017). *Active Work Zone Safety Using Emerging Technologies 2017*. University Transportation Center for Alabama The University of Alabama.

Martin, J., Rozas, A., & Araujo, A. (2016). A WSN-Based Intrusion Alarm System to Improve Safety in Road Work Zones. *Journal of Sensors*, *2016*, 1–8. https://doi.org/10.1155/2016/7048141

Medina-Salgado, B., Sánchez-DelaCruz, E., Pozos-Parra, P., & Sierra, J. E. (2022). Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, *35*, 100739. https://doi.org/10.1016/j.suscom.2022.100739

Meng, Q., & Weng, J. (2011). A Genetic algorithm approach to assessing work zone casualty risk. *Safety Science*, *49*(8–9), 1283–1288. https://doi.org/10.1016/j.ssci.2011.05.001

Mishra, S. (2013). A Synchronized Model for Crash Prediction and Resource Allocation to Prioritize Highway Safety Improvement Projects. *Procedia - Social and Behavioral Sciences*, *104*, 992–1001. https://doi.org/10.1016/j.sbspro.2013.11.194

Mishra, S., Golias, M. M., Sharma, S., & Boyles, S. D. (2015). Optimal funding allocation strategies for safety improvements on urban intersections. *Transportation Research Part A: Policy and Practice*, *75*, 113–133. https://doi.org/10.1016/j.tra.2015.03.001

Mishra, S., Golias, M. M., & Thapa, D. (2021). *Work Zone Alert Systems*. Tennessee Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56274

Mohan, D., Bangdiwala, S. I., & Villaveces, A. (2017). Urban street structure and traffic safety. *Journal of Safety Research*, *62*, 63–71. https://doi.org/10.1016/j.jsr.2017.06.003

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis. *Transportation Research Record*, *2673*(11), 680–692. https://doi.org/10.1177/0361198119845899

Nam, D., & Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, *34*(2), 85–102. https://doi.org/10.1016/S0965-8564(98)00065-2

Novosel, C. (2014). Evaluation of Advanced Safety Perimeter Systems for Kansas Temporary Work Zones. In *Civil, Environmental, and Architectural Engineering, University of Kansas*.

Osman, M., Mishra, S., & Paleti, R. (2018a). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group

differences. *Accident Analysis and Prevention*, *118*(May), 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018b). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018c). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., Paleti, R., & Golias, M. (2019). Impacts of Work Zone Component Areas on Driver Injury Severity. *Journal of Transportation Engineering, Part A: Systems*, *145*(8), 04019032. https://doi.org/10.1061/jtepbs.0000253

Osman, M., Paleti, R., & Mishra, S. (2018a). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis and Prevention*, *111*(May 2017), 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., & Mishra, S. (2018b). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis & Prevention*, *111*, 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis and Prevention*, *97*, 261–273. https://doi.org/10.1016/j.aap.2016.10.020

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051–10072. https://doi.org/10.1007/s12652-020-02759-5

Ozturk, O., Ozbay, K., Yang, H., & Bartin, B. (2013). Crash Frequency Modeling for Highway Construction Zones. *Transportation Research Board's 92nd Annual Meeting, Washington, D.C*, 14p.

Paleti, R., Mahmud, A., Gayah, V., & Pinjari, A. (2021). When and Where does the Next Traffic Crash Occur? A Discretized Duration Based Modeling Approach. *Under Review for Publication*.

Park, E. S., & Lord, D. (2007). Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record*, *2019*, 1–6. https://doi.org/10.3141/2019-01

Pei, X., Wong, S. C., & Sze, N. N. (2011). A joint-probability approach to crash prediction models. *Accident Analysis & Prevention*, *43*(3), 1160–1166. https://doi.org/10.1016/j.aap.2010.12.026

Pham, M.-H., Bhaskar, A., Chung, E., & Dumont, A.-G. (2010). Random forest models for identifying motorway Rear-End Crash Risks using disaggregate data. *13th International IEEE Conference on Intelligent Transportation Systems*, 468–473. https://doi.org/10.1109/ITSC.2010.5625003

Provost, F., Jensen, D., & Oates, T. (2001). Progressive Sampling. In *Instance Selection and Construction for Data Mining* (pp. 151–170). Springer US. https://doi.org/10.1007/978-1-4757-3359-4_9

Qi, Y., Srinivasan, R., Teng, H., & Baker, R. F. (2005). *Frequency of Work Zone Accidents on Construction Projects*.

Quddus, M. (2013). Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, *5*(1), 27–45. https://doi.org/10.1080/19439962.2012.705232

Rahim, M. A., & Hassan, H. M. (2021). A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention*, *154*, 106090. https://doi.org/10.1016/j.aap.2021.106090

Rahman, R., Bhowmik, T., Eluru, N., & Hasan, S. (2021). Assessing the crash risks of evacuation: A matched case-control approach applied over data collected during Hurricane Irma. *Accident Analysis & Prevention*, *159*, 106260. https://doi.org/10.1016/j.aap.2021.106260

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007

Sarker, A. A., Naimi, A., Mishra, S., Golias, M. M., & Freeze, P. B. (2015). Development of a Secondary Crash Identification Algorithm and occurrence pattern determination in large scale multi-facility transportation network. *Transportation Research Part C: Emerging Technologies*, *60*, 142–160. https://doi.org/10.1016/j.trc.2015.08.011

Scott-Parker, B., Hyde, M. K., Watson, B., & King, M. J. (2013). Speeding by young novice drivers: What can personal characteristics and psychosocial theory add to our understanding? *Accident Analysis & Prevention*, *50*, 242–250. https://doi.org/10.1016/j.aap.2012.04.010

Shangguan, Q., Fu, T., & Liu, S. (2020). Investigating rear-end collision avoidance behavior under varied foggy weather conditions: A study using advanced driving simulator and survival analysis. *Accident Analysis and Prevention*, *139*(March), 105499. https://doi.org/10.1016/j.aap.2020.105499

Sharma, A., Bullock, D., & Peeta, S. (2011). Estimating dilemma zone hazard function at high speed isolated intersection. *Transportation Research Part C: Emerging Technologies*, *19*(3), 400–412. https://doi.org/10.1016/j.trc.2010.05.002

Simons-Morton, B. G., Ouimet, M. C., Chen, R., Klauer, S. G., Lee, S. E., Wang, J., & Dingus, T. A. (2012). Peer influence predicts speeding prevalence among teenage drivers. *Journal of Safety Research*, *43*(5–6), 397–403. https://doi.org/10.1016/j.jsr.2012.10.002

Song, J. J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, *97*(1), 246–273. https://doi.org/10.1016/j.jmva.2005.03.007

Stipancic, J., Miranda-Moreno, L., Saunier, N., & Labbe, A. (2019). Network screening for large urban road networks: Using GPS data and surrogate measures to model crash frequency and severity. *Accident Analysis and Prevention*, *125*(February), 290–301. https://doi.org/10.1016/j.aap.2019.02.016

Stout, D., Graham, J., Bryant-Fields, B., Migletz, J., Fish, J., & Hanscom, F. (1993). *Maintenance Work Zone Safety Devices Development and Evaluation*.

Sun, J., & Sun, J. (2016a). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Sun, J., & Sun, J. (2016b). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y., & Huang, H. (2020). Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Analytic Methods in Accident Research*, *27*, 100123. https://doi.org/10.1016/j.amar.2020.100123

Thapa, D., & Mishra, S. (2021a). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., & Mishra, S. (2021b). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*, 106125. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., Paleti, R., & Mishra, S. (2022a). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Thapa, D., Paleti, R., & Mishra, S. (2022b). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Theiss, L., Ullman, G. L., & Lindheimer, T. (2017). *Closed Course Performance Testing of the Aware Intrusion Alarm System*.

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2019). Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention*, *130*, 151–159. https://doi.org/10.1016/j.aap.2017.12.018

Therneau, T. M. (2020). *A Package for Survival Analysis in R. R package version 3.2-7*.

Therneau, T. M., Grambsch, P. M., & Panktatz, S. V. (2003). Penalized Survival Models and Frailty. *Journal of Computational and Penalized Survival Models and Frailty*, *12*(1), 156–175.

Ullman, G. L., Trout, N. D., & Theiss, L. (2016). *Driver Responses to the AWARE Intrusion Alarm System*. Texas A&M Transportation Institute.

Venugopal, S., & Tarko, A. (2000). Safety models for rural freeway work zones. *Transportation Research Record*, *1715*, 1–9. https://doi.org/10.3141/1715-01

Wang, C., Quddus, M. A., & Ison, S. G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention*, *43*(6), 1979–1990. https://doi.org/10.1016/j.aap.2011.05.016

Wang, J., Yamamoto, T., & Liu, K. (2020). Key determinants and heterogeneous frailties in passenger loyalty toward customized buses: An empirical investigation of the subscription termination hazard of users. *Transportation Research Part C: Emerging Technologies*, *115*(July 2019), 102636. https://doi.org/10.1016/j.trc.2020.102636

Wang, X., Katz, R., & Dong, X. S. (2018). *Fatal Injuries at Road Construction Sites among Construction Workers* [Quarterly]. Center for Construction Research and Training.

https://www.cpwr.com/wp-content/uploads/publications/publications_Quarter2-QDR-2018.pdf

Work Zones-Injury Facts-National Safety Council. (2020). https://injuryfacts.nsc.org/motor-vehicle/motor-vehicle-safety-issues/work-zones/

Wu, L., Meng, Y., Kong, X., & Zou, Y. (2020). Incorporating survival analysis into the safety effectiveness evaluation of treatments: Jointly modeling crash counts and time intervals between crashes. *Journal of Transportation Safety and Security*, *0*(0), 1–21. https://doi.org/10.1080/19439962.2020.1786871

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

Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., & Jiang, X. (2020). A machine-learning method for improving crash injury severity analysis: A case study of work zone crashes in Cairo, Egypt. *International Journal of Injury Control and Safety Promotion*, *27*(3), 266–275. https://doi.org/10.1080/17457300.2020.1746814

Yang, H., Ozbay, K., Ozturk, O., & Xie, K. (2015). Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. *Traffic Injury Prevention*, *16*(4), 387–396. https://doi.org/10.1080/15389588.2014.948615

Yasmin, S., & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention*, *59*, 506–521. https://doi.org/10.1016/j.aap.2013.06.040

Yasmin, S., & Eluru, N. (2018). A joint econometric framework for modeling crash counts by severity. *Transportmetrica A: Transport Science*, *14*(3), 230–255. https://doi.org/10.1080/23249935.2017.1369469

Yasmin, S., Eluru, N., Bhat, C. R., & Tay, R. (2014). A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic Methods in Accident Research*, *1*, 23–38. https://doi.org/10.1016/j.amar.2013.10.002

Yasmin, S., Eluru, N., Wang, L., & Abdel-Aty, M. A. (2018). A joint framework for static and real-time crash risk analysis. *Analytic Methods in Accident Research*, *18*, 45–56. https://doi.org/10.1016/j.amar.2018.04.001

Ye, X., Pendyala, R. M., Shankar, V., & Konduri, K. C. (2013). A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention*, *57*, 140–149. https://doi.org/10.1016/j.aap.2013.03.025

Yu, B., Chen, Y., & Bao, S. (2019). Quantifying visual road environment to establish a speeding prediction model: An examination using naturalistic driving data. *Accident Analysis & Prevention*, *129*, 289–298. https://doi.org/10.1016/j.aap.2019.05.011

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, *51*, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis and Prevention*, *73*, 351–358. https://doi.org/10.1016/j.aap.2014.09.006

Zhang, C., He, J., Wang, Y., Yan, X., Zhang, C., Chen, Y., Liu, Z., & Zhou, B. (2020). A Crash Severity Prediction Method Based on Improved Neural Network and Factor Analysis. *Discrete Dynamics in Nature and Society*. https://doi.org/10.1155/2020/4013185

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods. *IEEE Access*, *6*, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*(1), 204–215. https://doi.org/10.1016/j.ijforecast.2010.06.002

Zhang, K., & Hassan, M. (2019a). Crash severity analysis of nighttime and daytime highway work zone crashes. *PLoS ONE*, *14*(8), 1–17. https://doi.org/10.1371/journal.pone.0221128

Zhang, K., & Hassan, M. (2019b). Identifying the Factors Contributing to Injury Severity in Work Zone Rear-End Crashes. *Journal of Advanced Transportation*, *2019*, 1–9. https://doi.org/10.1155/2019/4126102

Zhao, G., Wu, C., & Qiao, C. (2013). A Mathematical Model for the Prediction of Speeding with its Validation. *IEEE Transactions on Intelligent Transportation Systems*, *14*(2), 828–836. https://doi.org/10.1109/TITS.2013.2257757

Zheng, L., & Sayed, T. (2020). A novel approach for real time crash prediction at signalized intersections. *Transportation Research Part C: Emerging Technologies*, *117*, 102683. https://doi.org/10.1016/j.trc.2020.102683

Zimmerman, K., Mzige, A. A., Kibatala, P. L., Museru, L. M., & Guerrero, A. (2012). Road traffic injury incidence and crash characteristics in Dar es Salaam: A population based

study. *Accident Analysis & Prevention*, *45*, 204–210.

https://doi.org/10.1016/j.aap.2011.06.018

5. Assessing Driver Behavior in Work Zones: A Discretized Duration Approach to Predict Speeding

Introduction

Highway construction and maintenance play a crucial role in enhancing and sustaining transportation infrastructure, which experiences increasing use by travelers every year. During these operations, work zones are established to ensure the safety of workers and road users. These work zones can be noisy, distracting, and confusing due to the presence of heavy equipment and machinery. Consequently, they become more susceptible to safety mishaps, particularly from oncoming traffic. In fact, highway construction work is categorized as one of the most hazardous occupations. For instance, among all road construction sites, work zones involving paving/surfacing equipment operators and maintenance workers have the second and third highest fatality rates (X. Wang et al., 2018). Most of these fatalities and crashes can be attributed to adverse driver behavior and non-compliance with work zone safety measures. Among several factors influencing work zone crashes, speeding stands out as the most common. According to the Fatality Analysis Reporting System (FARS) database, in 2021, 32% of work zone fatalities were linked to speeding as a contributing factor, with 24% of fatal crashes resulting from rear-end collisions (Federal Highway Administration, 2023).

The Manual of Uniform Traffic Control Devices (MUTCD) classifies work zones based on their location and duration (Federal Highway Administration, 2009b). It serves as a comprehensive guide for traffic control and enforcement of safety measures in work zones. The main objective of these safety measures is to ensure smooth traffic flow and consistent speeds throughout the work zone, thereby avoiding abrupt changes that could lead to crashes. Most work zone crashes, especially rear-end collisions, occur due to inconsistent traffic flow or sudden

speed variations. To address this issue, the MUTCD provides guidance on implementing various technologies and strategies to maintain a steady traffic flow and enforce safety measures in work zones. The guide focuses on increasing compliance among road users and eliminating adverse driver behavior using regulatory strategies (such as speed photo radar enforcement and police presence) and warning strategies (including warning signs, dynamic message signs, speed feedback systems, etc.). Despite the implementation of existing safety strategies and work zone enforcement measures, work zone crashes have been on the rise in recent years. Between 2020 and 2021, work zone fatalities witnessed a troubling increase of 10.8% (Federal Highway Administration, 2023). In 2020, 39% of all work zone crash fatalities in the US occurred on interstates, with a slight rise to about 40% in 2021. The higher traffic speed and lower work zone compliance on interstates contribute significantly to the number of fatalities. Extensive evidence in the literature suggests that higher speeds are associated with more severe crashes (Osman, Mishra, et al., 2018c; Osman, Paleti, et al., 2018b).

In this context, the ability to predict speeding can bring significant benefits from both traffic safety and operational perspectives. Having prior knowledge of potential speeding events can assist transportation planners and agencies in preparing in advance and taking necessary steps to prevent such occurrences. This proactive approach can lead to a potential reduction in crashes, alleviate congestion resulting from accidents, and optimize resource allocation by identifying critical highway segments. Therefore, this study uses a discretized duration framework to model and predict speeding on highway segments with existing work zones. The implemented duration-based framework is specifically designed to incorporate time-varying covariates into the multinomial logit model (MNL) through time-discretization. This enables the calculation of the risk of speeding, allowing forecasting road users' speeding behavior.

Literature review

In the existing literature, driver behavior within work zones is primarily characterized by compliance with two key factors: (i) the enforced speed limit and (ii) merge behavior. Notably, these two aspects are major contributors to work zone crashes, and as a result, work zone safety measures focus on promoting safer driving behavior by regulating the operating speed limit and merge behavior. Various studies have highlighted the significance of speed compliance in reducing crash risks. Higher speeds have been linked to an increased likelihood of crashes and more severe outcomes (Osman et al., 2016; Osman, Paleti, et al., 2018b; K. Zhang & Hassan, 2019b). Additionally, unsafe and aggressive merge behavior, combined with adverse weather and lighting conditions, has been identified as risky driving behavior (Debnath et al., 2015). Interestingly, aggressive driving and merge behavior are also associated with traffic speed. Drivers encountering slower speeds, congestion, and travel delays tend to become frustrated, leading to more aggressive maneuvers on the road. This highlights the interconnectedness between driving behavior and traffic flow within work zones.

Work zone risk factors and driver behavior

Work zones can lead to sudden disruptions in traffic flow, resulting in slowdowns, queues, lane change maneuvers, traffic conflicts, and speeding, all of which impact driving behavior (Flannagan et al., 2019). Researchers have extensively studied the factors that influence driver behavior in work zones. Nearly half of all work zone crashes occur in the vicinity of the activity area (Dissanayake & Akepati, 2009). Among these crashes, approximately 42% are rear-end collisions. The main contributors to work zone crashes include taking no improper action (32.1%), inattentive driving (19%), and following too closely (9.7%). Driver behavior also varies based on different work zone types and activity levels. For instance, when navigating through

longer work zone closures, drivers tend to travel at higher speeds (Hamdar et al., 2016). The type of barriers used also influences driver headway. Adverse weather, poor lighting conditions, and middle-aged drivers have been associated with risky driving behavior. Workers involved in work zone construction highlighted the most hazardous conditions they face, such as working in wet weather leading to reduced visibility and skid resistance, driver frustration, aggression towards traffic controllers, and distracted driving due to mobile phone use (Debnath et al., 2015). Peak hours and non-daylight hours (dawn, dusk, and night) are considered the most hazardous times for work zone activities, attributed to a higher number of drunk drivers and reduced visibility. Additionally, workers perceive working on freeways and hilly/curved roads as risky. Regarding speed compliance, workers consider police enforcement, the presence of police cars (even without an officer present), installation of speed bumps, and work zone-oriented driver education as the most effective countermeasures. These measures aim to encourage drivers to comply with speed limits and improve safety within work zones.

Predicting driving behavior and traffic flow

As previously mentioned, driving behavior and traffic flow are mutually dependent. Many studies examining driver behavior under various circumstances, such as the implementation of new work zone enforcement measures, have employed three main approaches: i) Field observation and analysis, ii) Traffic micro and macrosimulation, for instance, studies conducted by (Berthaume, 2015; Gan et al., 2021; Hou & Chen, 2019), and iii) Driving simulator experiments, as demonstrated in research conducted by (Algomaiah & Li, 2022; Bashir & Zlatkovic, 2021). The first approach, field observation and analysis, is beneficial when there is no prevalent risk or when adequate safety for road users can be ensured, as seen in previous studies (Benekohal et al., 2010; Mishra et al., 2021; Thapa & Mishra, 2021b). On the other hand,

the latter two approaches, traffic micro and macrosimulation, and driving simulator experiments, are preferred to avoid hazardous conditions and provide controlled environments for studying driver behavior in work zones.

Understanding the impact of work zones and driving behavior on traffic flow is crucial from an Intelligent Transportation Systems (ITS) perspective. Real-time and accurate traffic data play a vital role in various ITS applications, including traffic planning and management, incident detection and management, travel time estimation, traffic predictions, and traffic planning. To achieve these objectives, researchers have focused on accurately forecasting traffic flow, for missing data and future conditions.

Numerous research approaches have been explored in this area, including time series and regression analysis, Kalman filter, machine learning techniques such as neural networks and support vector machines, as well as deep learning techniques like convolutional neural networks, long short-term memory, and graphical convolutional networks. For a detailed description of these methods and relevant literature, readers are encouraged to refer to studies conducted by (Medina-Salgado et al., 2022) and (Kashyap et al., 2022). In summary, the primary goal of these methods is to forecast traffic flow conditions rather than focusing on driving behavior, contributing to the advancement of ITS applications and traffic management.

Predicting speeding behavior

Various approaches have been employed in the existing literature to predict driving intention and behavior related to violating traffic laws. The theory of planned behavior has been widely utilized in multiple studies (e.g., (Elliott & Thomson, 2010; Forward, 2009; Jovanović et al., 2017; Scott-Parker et al., 2013)). For instance, Cestac et al. (2011) investigated young drivers and found that different latent constructs influenced speeding behavior in different driver groups

(Cestac et al., 2011). Novice drivers were influenced by thrill-seeking, beginners by subjective norms, and experienced drivers by the feeling of being in control. In another study, researchers reported that as young drivers are more likely to speed as they gain confidence in their driving abilities (Simons-Morton et al., 2012). Risky peer influences were found to be significant predictors of speeding among novice teenage drivers.

In a different approach, Yu et al. (2019) used naturalistic driving data to develop a speeding prediction model (B. Yu et al., 2019). The study emphasized the role of driver's visual perception as a major factor in speeding. The prediction model was built based on visual road information, environmental variables, vehicle kinematics, and driver characteristics, utilizing a Random Forest algorithm to achieve an accurate prediction rate of 85%.

Zhao et al., (2013) developed a mathematical model to predict intentional and non-intentional speeding (Zhao et al., 2013). The model utilized in-vehicle sensor data and driver characteristics to calculate speeding probabilities. The experiments were conducted using a driving simulator, and the authors reported an average prediction accuracy of over 80%.

In another study, Cheng et al., (2019) adopted a two-step approach to identify and predict speed violations (Z. Cheng et al., 2019). They used a binary logit model to identify variables contributing to speeding violations and then applied a decision tree method to predict specific types of speeding violations, such as "foreign license plate" and "intersection," among others. The study found that country roads had a higher incidence of speeding violations compared to urban roads, primarily due to the lower presence of traffic control infrastructure and lower traffic flow. Higher and more intense rainfall was associated with increased speeding violations, while local drivers were less likely to violate speed limits.

Study contributions

This study contributes to the literature in three major ways:

i. Based on the literature review, numerous studies have examined speeding behavior using the theory of planned behavior. Additionally, a separate body of literature focuses on predicting speeding at the individual driver level, utilizing environmental and in-vehicle data. Furthermore, another set of studies has applied parametric, machine learning, and deep learning techniques to forecast traffic flow and speed, enabling various actions such as crash and congestion prevention, emergency messaging for traffic diversion, rerouting, and queue management, particularly in situations with insufficient or missing disaggregated data. Despite the wealth of research in these areas, we are not aware of any previous study attempting to forecast the likelihood of speeding in the future using historical data and time-varying covariates through a parametric approach. This study aims to fill this research gap by providing insights into predicting speeding behavior using historical data, and time-varying covariates with a parametric approach.

ii. While many prediction models rely on modern data-driven black-box machine learning and artificial intelligence algorithms, our approach is based on exponential models (survival model and MNL). These parametric methods offer the advantage of providing causal inferences through variable effects, including coefficients and marginal/elasticity effects. The enables researchers to gain deeper insights into the relationships between the predictors and speeding.

iii. To the best of our knowledge, apart from Thapa, et al. (2022), there have been no previous implementations of the duration-based model (Thapa et al., 2022b). Additionally, this research stands out as the first to employ this framework for predicting speeding violations,

demonstrating the integration of real-time weather, traffic flow, and congestion data alongside static covariates like highway characteristics. Moreover, our approach considers the presence of unobserved heterogeneity resulting from multiple speeding events occurring in the same highway segment. This aspect of our study allows for a more comprehensive analysis and understanding of the factors influencing speeding behavior and its prediction.

Methodology

The description of the duration-based framework here is taken largely from Thapa, et al. (2022) (Thapa et al., 2022b). Utilizing the duration-based framework, we can determine the likelihood of speeding at a particular time-interval *t*, considering that no speeding has been observed in previous time-intervals. This probability is represented by the hazard function *h(t)*, which can be formulated using a constant hazard rate, *h*.

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{he^{-ht}}{1 - (1 - e^{-ht})} = h \tag{31}$$

In the given equation 31, we represent the probability distribution function and probability density function related to a continuous random variable for time *T* as *f(t)* and *F(t)*, respectively. The probability density function, in this context, indicates the likelihood of observing speeding by time *t*. This is expressed by equation 32 as follows.

$$F(t) = Pr(T \leq t) \tag{32}$$

Assuming that the time duration between consecutive speeding events is discretized into *n* time-intervals, each having a duration of *dt*, we can express the probability of observing the next speeding event at a specific interval *n* since the occurrence of the last speeding event as follows:

$$Pr(T = ndt) = Pr(T \leq ndt) - Pr(T \leq (n-1)dt) \tag{33}$$

$$= F(ndt) - F\big((n-1)dt\big)$$

$$= exp\,(-h(n-1)dt) - exp\,(-hndt)$$

$$= \frac{exp\,(-h(n-1)dt)}{1/(1 - exp\,(-hdt))}$$

Using a Taylor series expansion, i.e., $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots \infty$ -1 < x < 1 in the denominator produces equation 34.

$$Pr(T = ndt) = \frac{exp(-h(n-1)dt)}{1 + exp(-hdt) + exp(-2hdt) + exp(-3hdt) + \cdots \infty} \tag{34}$$

$$= \frac{exp(U_n)}{exp(U_1) + exp(U_2) + exp(U_3) + \cdots \infty}, \text{where } U_n = -h(n-1)dt$$

$$= \frac{exp\,(U_n)}{\sum_1^\infty exp\,(U_c)}$$

Simplifying equation 34 makes it evident that the probability of speeding at the n^{th} interval can be represented as MNL model with infinite alternatives for *n*. The utility equation for the alternatives can be expressed as $U_n = -h(n-1)dt$. As a result, the utility equation can be modified to accommodate non-linear hazard profiles, as demonstrated in equation 35. However, when disregarding all higher-order polynomial terms, it reduces to a simple MNL model.

$$U_n = \beta_1(n-1)dt + \beta_2[(n-1)dt]^2 + \beta_2[(n-1)dt]^3 + \cdots \tag{35}$$

For example, consider the time between consecutive speeding events observed at a specific segment, denoted as s, is discretized into epochs *e*, with *C* number of time-intervals, each lasting for *dt* duration. An illustrative example of this discretization is presented in the table, where two speeding events are observed, 4 hours apart, with *e*=1 hour and *dt*=15 minutes. Consequently, each epoch contains four distinct time-intervals, indexed as *i*={*1,2,3,4*}, and each time-interval provides the corresponding time elapsed since the previous speeding event. This relationship allows us to pinpoint the exact time-interval when the subsequent speeding event

occurred. For example, consider Table 27 which denotes speeding was observed at the fourth time-interval of the fourth epoch, denoted as 1, otherwise 0. As a result, the time elapsed between the two speeding events can be calculated as $t_{e,i} = (e-1)Cdt + (i-1)dt = (4-1) * 4 * 0.25 + (4-1) * 0.25 = 3.75\ hours$. This relationship enables us to construct the utility function considering the time-intervals as choice alternatives in the MNL model. The utility function includes the duration dynamics as the first element and a vector of time-varying covariates as the last element. Notably, even static variables that do not change with time, such as the number of lanes associated with the highway segment, were transformed into dynamic covariates by multiplying them with the corresponding value of $t_{e,i}$ to account for the effect of time elapsed.

$$U_{s,e,i} = \beta_1 t_{e,i} + \cdots + r' X_{s,e,i} \tag{36}$$

Alongside the four time-intervals, there is a fifth alternative to consider, signifying whether the next speeding event will be observed in the current epoch (0) or the next epoch (1). This particular choice alternative serves as the base with an intercept term and can be represented as follows.

$$U_{s,e,C+1} = \beta_{C+1} \tag{37}$$

Since speeding at any time-interval is conditional upon no prior speeding, the conditional probability for any time-interval can be expressed as follows.

$$Pr\big(T_s = t_{e,i} \big| T_s > (e-1)Cdt\big) = \frac{exp\ (U_{s,e,i})}{\sum_{c=1}^{C} exp\ (U_{s,e,c}) + exp\ (U_{s,e,C+1})} \tag{38}$$

The unconditional probability can be obtained by multiplying the conditional probability with the product of probabilities for the fifth alternative, as demonstrated in equation 39.

$$Pr(T_s = t_{e,i}) = \frac{exp\ (U_{s,e,i})}{\sum_{c=1}^{C} exp\ (U_{s,e,c}) + exp\ (U_{s,e,C+1})} \prod_{e*=1}^{e-1} \frac{exp\ (U_{s,e*,C+1})}{\sum_{c=1}^{C} exp\ (U_{s,e*,c}) + exp\ (U_{s,e*,C+1})} \tag{39}$$

Estimating the model parameters, represented as the vector $n = (\beta_1, \dots, r, \beta_{C+1})'$, involves maximizing the likelihood function associated with the probabilities in equation 39 across all speeding events and segments. It's essential to note that the data, after time discretization, takes the form of panel data with multiple speeding events observed at each segment. Thus, it becomes necessary to consider unobserved heterogeneity at the segment level. To address this, the vector of parameters for any segment is assumed to follow a multivariate normal distribution. The resulting mixed logit model is then estimated by integrating the vector of parameters over this distribution. The random parameters for the mixed logit model are obtained as Cholesky parameters by estimating the elements of the unconstrained lower triangular Cholesky matrix, represented as G. This estimation is performed in relation to their variance-covariance matrix, denoted as S, such that $G\ G' = S$.

Table 27
Example demonstrating the discretization of duration between speeding events

Segment	Time to next speeding (hours)	Epoch	First 15-min	Second 15-min	Third 15-min	Fourth 15-min	Next epoch
A	4	1	0	0	0	0	1
A	4	2	0	0	0	0	1
A	4	3	0	0	0	0	1
A	4	4	0	0	0	1	0

Data

Speeding events

This research focused on speeding incidents observed in a work zone set up on I-65 in Robertson County, Tennessee. The I-65 segments within the county are currently undergoing lane expansion in both North and Southbound lanes. The specific location of these interstate segments within Robertson County can be seen in Figure 16. Based on the data obtained from Google Maps Street View, it was determined that the work zone has been active since July 2022.

Therefore, the study period considered for this research spans from July 1, 2022, to May 31, 2023. The work zone consists of 14 INRIX Traffic Message Channel (TMC) segments presented in Table 28 . To identify speeding events during the study period, speed data at 15-minute intervals was collected for these 14 TMCs, along with the reference speed for the highway segments. In the context of INRIX, the reference speed represents the average speed of vehicles over the study period. For this study, speeding was identified for any given time-interval whenever the average speed exceeded the reference speed by 10 mph. Employing this method, a total of 2,444 speeding events were identified. It is essential to note that each speeding event corresponds to a specific 15-minute time interval during which the average vehicle speed exceeded the reference speed by 10 mph.

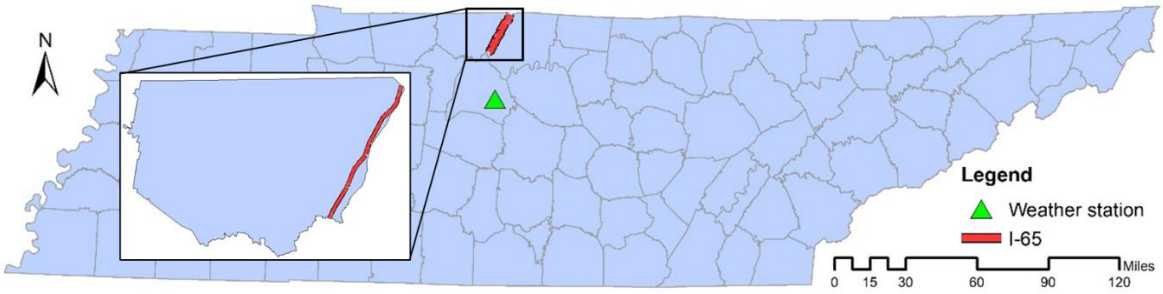


Figure 16 Location of TMC segments within Robertson County and weather station

Table 28
TMC segments within the work zone

Northbound			Southbound		
TMC segments	**Sequence**	**Length (mi)**	**TMC segments**	**Sequence**	**Length (mi)**
121+04668	1	3.49	121-04670	1	3.48
121P04668	2	0.59	121N04670	2	0.55
121+04669	3	4.15	121-04669	3	4.48
121P04669	4	0.59	121N04669	4	0.53
121+04670	5	4.42	121-04668	5	4.07
121P04670	6	0.56	121N04668	6	0.63
121+04671	7	3.39	121-04667	7	3.46

Note: TMC segments with the character “P” or “N” represents segments with exit.

Data aggregation

As the INRIX TMC data only provides traffic flow information, additional datasets were utilized to obtain covariates that could effectively represent the highway and weather conditions in the segments during the study period and at the time of each speeding event. The following datasets were used for this purpose:

i. *Traffic flow data*: Traffic flow data was sourced from the TMC data provided by INRIX, as mentioned earlier. This data contained disaggregated traffic speed information, recorded at 5-minute intervals. This data was aggregated to 15-minute intervals for the purposes of model development and speeding identification. Moreover, the Coefficient of Variation (COV) was calculated for each 15-minute interval to factor in variations in speed. To address the presence of congestion, the travel time index was obtained at 15-minute intervals. Additionally, yearly averages for traffic composition were collected for each segment. All of this data played a crucial role in accurately representing the traffic characteristics conditions and characteristics affecting speeding events during the study period.

ii. *Highway characteristics*: The necessary highway characteristics were collected from the Enhanced Tennessee Roadway Information Management System (ETRIMS), a query-based highway information system managed by the Tennessee Department of Transportation. Among the variables obtained, terrain type, lighting conditions, roadway conditions, and illumination were found to be consistent across all segments and, therefore, were not included in the analysis. The retained variables were the number of lanes and the posted speed limit, as they varied across the segments and were considered essential for the study. A geospatial proximity approach was applied using road inventory data to integrate these

highway characteristics with the traffic flow data. This ensured that the relevant highway characteristics were properly aligned with each segment's corresponding traffic flow data.

iii. *Weather conditions*: Weather data for the freeway segments was collected from the Local Climatological Data (LCD) recorded by the nearest weather station situated at Nashville International Airport. The hourly weather data was then merged with the traffic flow data based on the corresponding date and time. Figure 16 displays the weather station's location relative to the interstate segments. Given that our prediction model's smallest temporal resolution was 15 minutes with 1-hour epochs, we integrated the hourly weather conditions to account for the impact of changing weather conditions. To achieve this, we multiplied the hourly weather conditions for the corresponding epoch with the time elapsed for each time interval. It is important to note that the hourly weather data includes multiple conditions observed within an hour, such as cloudy, rainy, and clear weather. For this study, we categorized the data into two main categories: "Clear" when no adverse weather conditions were observed in a particular hour, and "Other" when at least one adverse weather condition was noted. Additionally, hourly visibility data in miles was obtained as part of the weather conditions dataset. The descriptive statistics for the variables obtained from these datasets are presented in Table 29.

Training and testing data

After identifying the speeding events, 90% of them (2,200 events) were randomly selected to create forecasting epochs for generating the training data, as explained in the Methodology section. This selection process is referred to as Epoch level sampling (Thapa et al., 2022b). All the MNL and mixed logit models were estimated using this training data. The remaining 10% of

the speeding events were used to create the testing data for model validation. This data was used to assess the performance and accuracy of the models.

Table 29
Descriptive statistics of speeding events

Categorical variables	**Frequency of speeding**			**Relative abundance**		
Time of day when speeding was observed						
Early morning (6 a.m. to 9 a.m.)	133			5.44%		
Late morning (9 a.m. to 12 p.m.)	8			0.33%		
Early afternoon (12 p.m. to 3 p.m.)	14			0.57%		
Late afternoon (3 p.m. to 6 p.m.)	10			0.41%		
Evening (6 p.m. to 12 a.m.)	453			18.54%		
Night (12 a.m. to 6 a.m.)	1,826			74.71%		
Weather condition						
Clear	2,130			87.15%		
Others	314			12.85%		
Continuous variables	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Time between speeding (hours)	0.25	1.75	18.62	44.38	44.25	1,559
Driving conditions						
Hourly visibility (miles)	0.12	9.94	10	9.229	10	10
Highway characteristics						
Number of lanes (both directions)	4	4	4	4.58	5	7
Speed limit	55	55	55	61	70	70
Traffic flow characteristics						
Coefficient of variation of speed	0	0.014	0.045	0.055	0.082	0.530
Travel time index	0.7	1	1	1.095	1.1	26.3
Peak hour (%)	7	8	8	7.95	8	9
Passenger vehicles (%)	67	67	69	68.41	69	70
Single unit trucks (%)	3	3	3	3	3	3
Multiple unit trucks (%)	27	28	28	28.59	30	30
% Peak SU trucks	2.31	2.31	3.26	2.87	3.26	3.26
% Peal MU trucks	20.17	20.17	20.17	21.66	23.84	23.84

Results

The model estimation process involved two main steps. In the first step, a fixed parameter model was estimated. In the second step, the estimates obtained from this fixed model were then used as initial values for estimating the random parameter model. The mean parameters of our model estimations are shown in Table 30. The mean parameters exhibit similar values between the

models; however, the mixed MNL model stands out as a superior fit due to its lower log-likelihood value at convergence. The *t*-stat values for the mixed model are notably large suggesting low values of standard error for the respective estimates. Table 31 presents the Cholesky parameters for the mixed model with respective *t*-stat within parenthesis. The reader will note that the significant digits in the value of *t*-stat are reduced in the table to accommodate the results on the same page. The mean parameters presented in Table 30, such as those for the fixed model, can be utilized to construct the utility function for each alternative following equation 36 outlined in the Methodology section. This allows us to calculate the utility for each choice alternative,

$U_{F/I,e,i} = -1.582 * t_{e,i} - 15.813 * Hourly\ visbility\ morning + 2.187 * Number\ of\ lanes$ …

$U_{Next\ epoch} = 4.065$

The findings indicate that higher visibility is negatively correlated with speeding events. As the number of lanes increases, the likelihood of speeding events also increases. As anticipated, a higher travel time index, which reflects congestion, is negatively linked to speeding events. Speeding is most likely during late morning hours. It is worth noting that daytime variables that were statistically insignificant in the fixed parameter model were statistically significant in the mixed model. The effect of posted speed limit was positive in the fixed model but changed to negative in the mixed model (β=0.789, *t*-stat=9.67 versus β=-1.036, *t*-stat=-285.47). This suggests a strong presence of heterogeneity at the segment level. Segments with a higher peak hour percentage are positively associated with speeding. Surprisingly, clear weather conditions were found to be negatively associated with speeding, suggesting that drivers may be more cautious in clear weather compared to adverse weather conditions.

Table 30
Mean parameters for fixed and mixed models

Variable groups	Variables	Coeff. (*t*-stat)	
		Fixed model	**Mixed model**
Intercept	Intercept	4.065 (26.300)	3.863 (117.901)
Duration dynamic	Time since speeding	-1.582 (-6.228)	-1.813 (-196.957)
Driving condition	Hourly visibility (miles)	-15.813 (-6.233)	-18.131 (-209.859)
Highway characteristics	Number of lanes	2.187 (2.876)	2.298 (110.530)
	Posted speed limit (mph)	0.789 (9.667)	-1.036 (-285.468)
	Segment length (mi)	-0.242 (-2.800)	-0.762 (-41.076)
Traffic flow characteristics	Coefficient of variation (Speed)	1.357 (3.276)	0.327 (5.739)
	Travel time index	-0.664 (-4.152)	-1.862 (-25.511)
	DHV %	6.464 (5.404)	3.930 (38.579)
	SU Tracks %	-4.746 (-6.23)	-5.441 (-208.491)
	MU Trucks %	14.383 (6.059)	4.601 (11.322)
	% Peak SU Trucks	-9.645 (-6.158)	-9.275 (-330.679)
	% Peak MU Trucks	-14.564 (-6.406)	-23.581 (-60.451)
Time of day (base=Night (12 a.m. to 6 a.m.))	Early morning (6 a.m. to 9 a.m.)	0.172 (0.274)*	0.320 (37.412)
	Late morning (9 a.m. to 12 p.m.)	9.644 (4.798)	9.713 (2,042.216)
	Early afternoon (12 p.m. to 3 p.m.)	3.989 (1.746)*	4.027 (1,074.780)
	Late afternoon (3 p.m. to 6 p.m.)	1.587 (0.839)*	1.622 (427.829)
	Evening (6 p.m. to 12 a.m.)	0.028 (0.113)*	-1.093 (-18.650)
Weather condition (base = Other)	Clear	-1.582 (-6.228)	-1.862 (-25.511)
	Observations	97,346	
Model fit measures	Average initial LL	-0.381	-0.381
	Average final LL	-0.133	-0.124
	McFadden's R-squared	0.651	0.674

*Indicates the variables were not statistically significant at 5% level of significance

Validation

Validation of the prediction model was performed at two levels, first, the ability of the model to predict the epoch where speeding was observed and second, the accuracy of model's prediction regarding the time-interval at which speeding was observed.

i. Epoch level prediction: At the epoch level, we introduced a new measure called Predicted Temporal Proximity $(PTP) = \left|\frac{Predicted\ over\ speeding\ epoch - Actual\ overspeeding\ epoch}{Actual\ over\ speeding\ epoch}\right| * 100\%$ to assess the model's performance in predicting the epoch when speeding was observed. A

lower value of PTP indicates a more accurate prediction of the epoch of speeding. To evaluate the model's predictive ability across different temporal ranges, we created subsets of the test data by removing speeding events that occurred at higher numbers of epochs. We then calculated the average value of PTP for each subset.

The results from the average PTP, as shown in Figure 17 indicate that speeding events observed within 25 epochs of the last speeding event have notably smaller PTP values. For instance, when considering speeding that occurred within 5 epochs, the average PTP is 61%. However, this average PTP increases to 76% when considering speeding within 25 epochs. These findings suggest that the model's predictions are more accurate for road segments where speeding is more commonly observed. Additionally, the model suggests that predictions become less reliable as the number of epochs increases, indicating a potential decrease in accuracy for predicting speeding events that are farther apart in time.

ii. Time-interval level prediction: The model's predictions at the time-interval level were evaluated by considering the rates of False Positive (FP), False Negative (FN), True Positive (TP), and True Negative (TN) using two key metrics: Specificity and Sensitivity.

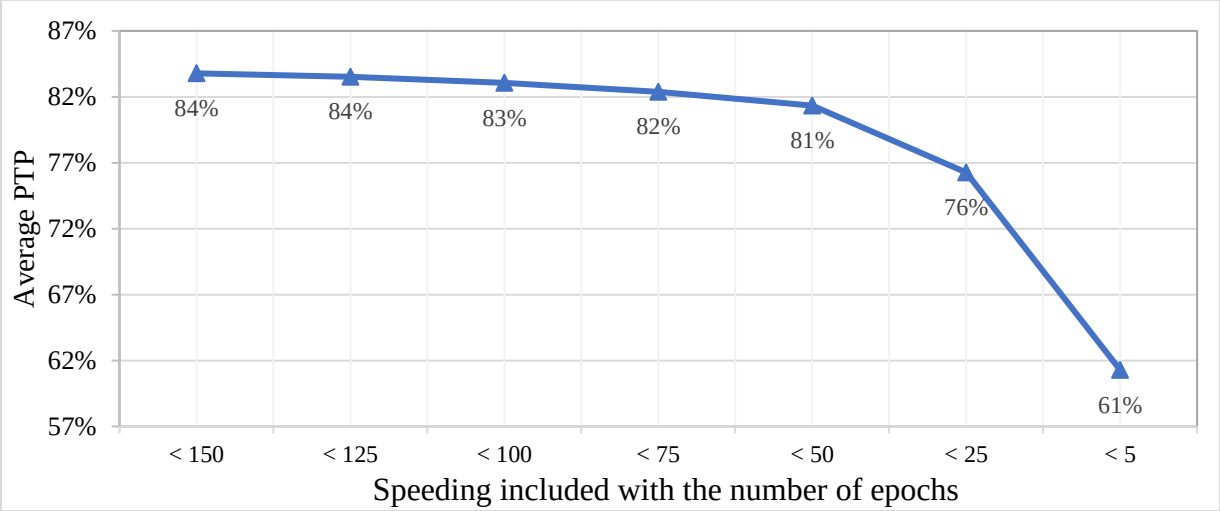
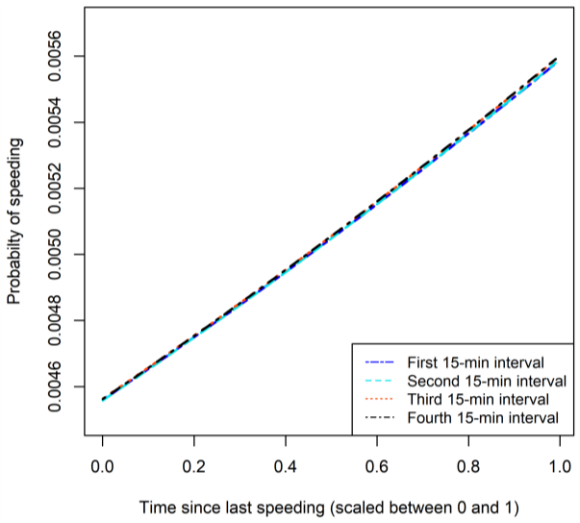
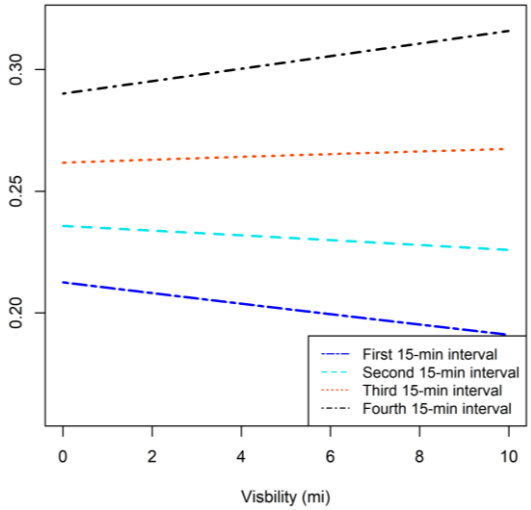


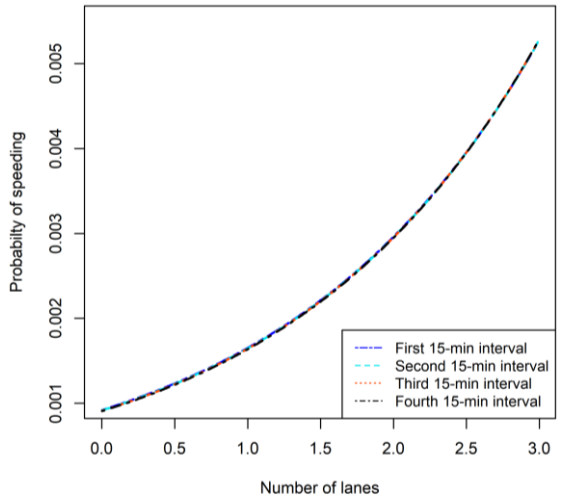
Figure 17 Value of PTP for different subset of test data



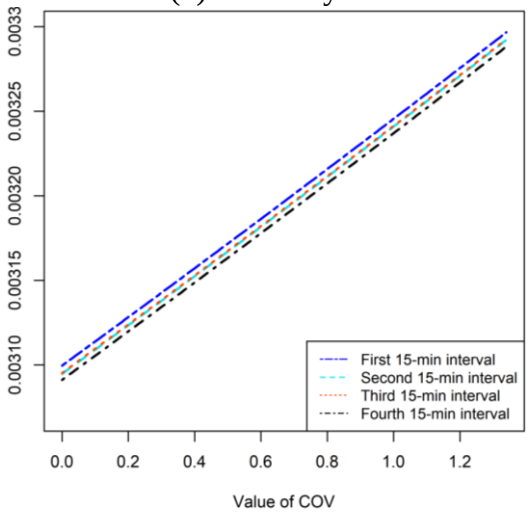
(a) Time since last speeding



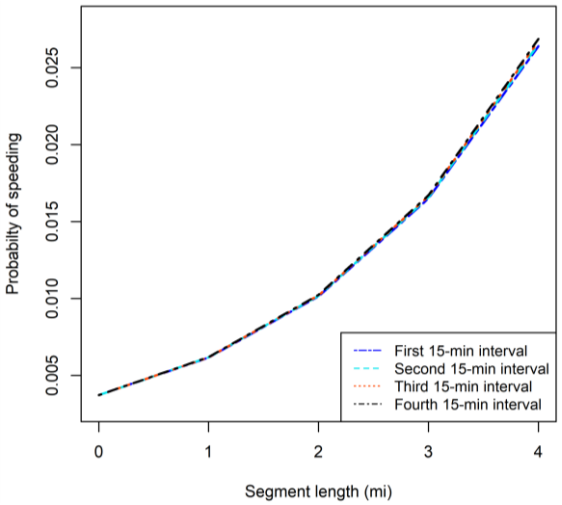
(b) Visibility



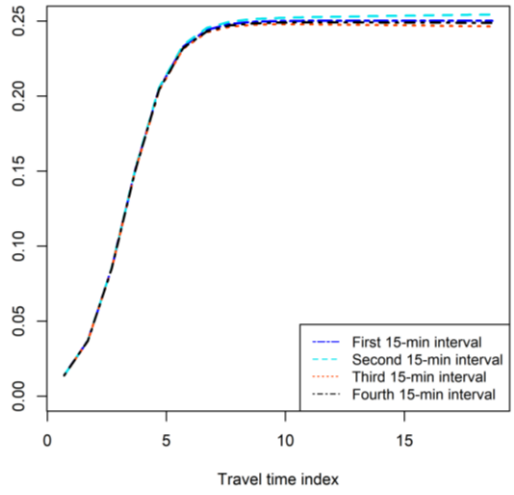
(c) Number of lanes



(d) Coefficient of variation (speed)



(e) Segment length (mi)



(f) Travel time index

Figure 18 Change in probability of outcomes with change in variables

.

Table 31

Correlated random parameter model (Cholesky parameters)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	-1.311 (-24)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	-0.534 (-11)	0.225 (11)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1.245 (23)	0.604 (42)	0.232 (11)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0.759 (41)	0.001 (0)	-1.066 (-16)	-1.856 (-24)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0.028 (1)	0.039 (11)	-1.01 (-18)	1.789 (26)	0.546 (16)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	-0.089 (-3)	0.035 (8)	-0.555 (-13)	-0.905 (-17)	-0.397 (-18)	0.658 (14)	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0.04 (1)	0.049 (10)	0.131 (17)	-0.564 (-16)	-0.107 (-12)	0.627 (14)	0.146 (14)	0	0	0	0	0	0	0	0	0	0	0	0
8	1.31 (30)	0.098 (16)	0.083 (8)	-3.105 (-19)	-0.215 (-15)	0.856 (17)	0.292 (19)	4.086 (18)	0	0	0	0	0	0	0	0	0	0	0
9	1.242 (18)	0.008 (3)	-0.831 (-18)	-1.623 (-28)	-0.225 (-17)	0.45 (32)	0.025 (6)	1.72 (30)	-0.503 (-18)	0	0	0	0	0	0	0	0	0	0
10	-0.072 (-4)	-0.013 (-4)	0.068 (5)	1.064 (17)	0.325 (14)	-0.177 (-19)	-0.04 (-10)	-0.504 (-18)	-0.028 (-6)	-0.328 (-22)	0	0	0	0	0	0	0	0	0
11	1.116 (27)	0.02 (6)	-0.817 (-17)	-1.651 (-20)	-0.122 (-14)	0.147 (13)	0.06 (13)	1.248 (19)	-0.115 (-17)	1.105 (19)	0.892 (20)	0	0	0	0	0	0	0	0
12	1.343 (24)	-0.033 (-8)	-0.874 (-14)	1.265 (13)	0.06 (13)	-0.098 (-9)	-0.098 (-13)	-1.339 (-14)	-0.006 (-2)	-1.045 (-13)	0.202 (48)	-0.09 (-12)	0	0	0	0	0	0	0
13	2.697 (32)	0.037 (8)	0.804 (16)	-3.76 (-17)	-0.464 (-16)	0.834 (14)	0.11 (11)	3.04 (15)	-0.569 (-20)	3.395 (16)	-0.141 (-28)	-0.221 (-29)	-0.045 (-11)	0	0	0	0	0	0
14	0.199 (3)	0.023 (6)	0.464 (16)	2.713 (31)	0.874 (16)	-0.286 (-19)	0.068 (12)	-0.534 (-30)	0.463 (19)	-1.047 (-24)	0.779 (19)	-0.274 (-19)	0.034 (9)	0.015 (4)	0	0	0	0	0
15	1.248 (19)	0.01 (3)	-0.836 (-13)	0.302 (9)	-0.166 (-14)	-0.078 (-9)	0.03 (9)	-0.143 (-6)	0.233 (18)	-0.57 (-14)	1.495 (17)	0.048 (14)	-0.004 (-1)	-0.005 (-1)	0.008 (3)	0	0	0	0
16	0.018 (0)	0.079 (13)	-0.223 (-11)	0.314 (16)	0.16 (18)	0.221 (14)	0.237 (16)	1.318 (15)	0.681 (17)	-0.039 (-2)	-0.035 (-2)	0.127 (10)	0 (0)	0.004 (1)	0.003 (1)	-0.202 (-32)	0	0	0
17	-0.095 (-1)	0.048 (10)	0.016 (1)	2.896 (46)	0.477 (16)	0.357 (11)	0.144 (13)	-0.028 (-1)	0.61 (20)	-0.785 (-42)	0.228 (14)	0.105 (21)	0.07 (13)	0.042 (11)	0.044 (11)	-0.427 (-21)	0.479 (13)	0	0
18	-0.273 (-3)	0.056 (12)	-0.584 (-23)	1.368 (28)	0.794 (16)	0.429 (17)	0.168 (16)	0.748 (17)	0.465 (17)	0.036 (3)	1.487 (19)	0.011 (3)	0.021 (6)	0.004 (1)	0.005 (2)	-1.019 (-15)	0.559 (17)	0.056 (12)	0
19	-0.272 (-8)	0.021 (5)	-0.689 (-14)	-1.154 (-12)	-0.047 (-11)	0.37 (13)	0.062 (9)	1.207 (11)	-0.17 (-13)	1.329 (12)	-0.802 (-18)	-0.01 (-3)	0.002 (1)	0.004 (1)	0.008 (3)	-0.781 (-27)	0.205 (10)	0.021 (5)	-1.283 (-21)

Note: 1=Intercept, 2=Time since speeding, 3=COV (speed) , 4=Posted speed limit, 5=Number of lanes, 6=DHV %, 7=SU Tracks %, 8=MU Trucks %, 9=% Peak SU Trucks, 10=% Peak MU Trucks, 11=Segment length (mi), 12=Early morning (6 a.m. to 9 a.m.), 13=Late morning (9 a.m. to 12 p.m.) , 14=Early afternoon (12 p.m. to 3 p.m.), 15=Late afternoon (3 p.m. to 6 p.m.), 16=Evening (6 p.m. to 12 a.m.), 17=Visibility (miles), 18=Clear weather, 19=Travel time index.

As presented by equation 40, Specificity is defined as the proportion of correctly identified non-speeding intervals (TN) among all the non-speeding intervals (TN + FP). It represents the model's ability to predict the absence of speeding events accurately. Sensitivity, on the other hand, as presented in equation 41 is defined as the proportion of correctly identified speeding intervals (TP) among all the speeding intervals (TP + FN). It measures the model's ability to predict the presence of speeding events correctly.

$$Specificity = \frac{True\ Negatives\ (TN)}{True\ Negatives\ (TN) + False\ Positives\ (FP)} \tag{40}$$

$$Sensitivity = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \tag{41}$$

Model predictions results were: TN=544, TP=56, FP=188, FN=188, Specificity=0.74, Sensitivity=0.23. High Specificity indicates low false positives, while low value of Sensitivity suggests a high rate of false negatives. It is important to consider that after reformulating the speeding data, there is a preponderance of 0s (non-speeding intervals) compared to 1s (speeding intervals). Given this imbalance, the low Sensitivity is expected. The model may tend to predict non-speeding intervals (TN) more accurately but struggle to identify all the instances of speeding intervals (TP). This is a common challenge in models dealing with imbalanced datasets, and further efforts may be required to improve the Sensitivity while maintaining a high Specificity.

Discussion

Utilizing the capability of parametric regression models to derive variable effects, we examined the influence of several variables on speeding probability. The variables considered in this investigation are listed below. The influence of other variables can also be derived in similarly.

Effect of time: Figure 18(a) demonstrates that the effect of time on speeding remains consistent across the four alternatives. As the time since the last speeding event increases, the probability of speeding steadily rises with each time interval. In this analysis, the time since the last speeding event is scaled between a minimum value of 0 and a maximum value of 1 to avoid the undue influence of large differences between the minimum and maximum time values.

Effect of visibility: Figure 18(b) illustrates that as visibility increases, the likelihood of speeding decreases for the first and second time intervals, but steadily increases afterward.

Effect of number of lanes: Figure 18(c) demonstrates that the increase in the probability of speeding is characterized by an upward curve as the number of lanes increases. However, it is worth noting that the differences between the time intervals themselves are not easily distinguishable. In other words, there is a non-linear increase in the probability of speeding with an increase in the number of lanes.

Effect of COV: The effect of the coefficient of variation of speed appears to be linear (see Figure 18(d)). While there is a small difference between the probabilities observed at the second and third time intervals, the difference between the first and fourth time intervals is more noticeable.

Effect of segment length: As discussed in the literature review section, studies have indicated that work zones with longer lane closures are more prone to speeding (Hamdar et al., 2016). The findings from this study align with those observations, showing that with an increase in highway segment length, the average vehicle speed, and consequently the probability of speeding, also increases, as depicted in Figure 18(e).

Effect of travel time index: A travel time index greater than 1 indicates a longer travel time than expected based on the operating speed limit, suggesting a congestion condition. Interestingly,

Figure 18(f) reveals that the probability of speeding rises sharply when the values of the travel time index are less than 6. However, beyond a travel time index of 6, the probability of speeding remains relatively constant. Additionally, no significant difference in probabilities is observed across different time intervals.

Conclusion

A well-functioning transportation system requires significant highway construction and maintenance to ensure efficiency. However, increased construction activities also expose workers to hazardous traffic conditions, which raises the risk of crashes. Despite implementing safety equipment, measures, laws, and policies, ensuring driver compliance with safety measures remains a concern. Enforcing reduced speed limits in work zones proves challenging due to the inherent nature of such work zones, leading to abrupt speed changes and speed violations that significantly contribute to crashes. Moreover, it is well-established that higher speeds are linked to more severe crashes, further amplifying the safety risks associated with work zones.

The objective of this study was to introduce a novel approach for predicting speeding, especially in work zones where it poses a significant threat to road users. Discretized duration framework was used, which allows to consider past speeding trends using historical data and real-time factors like weather, traffic flow, and highway characteristics to estimate the likelihood of speeding in discrete time intervals. Using this modeling framework, we can forecast future speeding behavior by treating these time intervals as choice alternatives in a MNL model. We focused on speeding events on I-65 in Robertson County, Tennessee, to test this approach. We utilized traffic speed and speeding identification data from INRIX, highway characteristics from ETRIMS, and weather data from LCD. The model successfully identified major contributors to speeding and demonstrated reasonably accurate predictive abilities, as evaluated using metrics

like PTP, Specificity, and Sensitivity. The average value of PTP indicates that the model can predict speeding within 61% of its time of occurrence. With Specificity at 0.74 and Sensitivity at 0.26, the model shows low false-positives and high false negatives. Overall, the model predictions suggest that transportation agencies can implement it to predict speeding events in real-time with a fair degree of accuracy.

It is important to note that reformulating speeding events using binary variables for speeding identification resulted in a prevalence of 0s over 1s in the data. This results in low value of Sensitivity. Future research could explore ways to improve model performance by addressing this class imbalance, for example, by using techniques like Synthetic Minority Oversampling Technique (SMOTE). In addition, the effect of different work zone enforcement techniques and strategies can also be studied.

Acknowledgements

This research was funded by the Tennessee Department of Transportation (TDOT) and the Center for Transportation Innovations in Education and Research (C-TIER) at the University of Memphis. The views expressed here are solely those of the authors and do not necessarily represent TDOT and C-TIER.

References

2019 Highway Work Zone Safety Survey. (2019). Associated General Contractors of America. https://www.agc.org/news/2019/05/23/2019-highway-work-zone-safety-survey

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

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Abdel-Aty, M., & Pande, A. (2007). Crash data analysis: Collective vs. Individual crash level approach. *Journal of Safety Research*, *38*(5), 581–587. https://doi.org/10.1016/j.jsr.2007.04.007

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record: Journal of the Transportation Research Board*, *1897*(1), 88–95. https://doi.org/10.3141/1897-12

Afghari, A. P., Haque, M. M., & Washington, S. (2020). Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. *Accident Analysis & Prevention*, *144*, 105615. https://doi.org/10.1016/j.aap.2020.105615

Agent, K. R., & Hibbs, J. O. (1996). *Evaluation of SHRP Work Zone Safety Devices*. 24.

Ahmed, S. S., Cohen, J., & Anastasopoulos, P. Ch. (2021). A correlated random parameters with heterogeneity in means approach of deer-vehicle collisions and resulting injury-severities. *Analytic Methods in Accident Research*, *30*, 100160. https://doi.org/10.1016/j.amar.2021.100160

Al-Ghamdi, A. S. (2002). Pedestrian–vehicle crashes and analytical techniques for stratified contingency tables. *Accident Analysis & Prevention*, *34*(2), 205–214. https://doi.org/10.1016/S0001-4575(01)00015-X

Algomaiah, M., & Li, Z. (2022). Enhancing Work Zone Capacity by a Cooperative Late Merge System Using Decentralized and Centralized Control Strategies. *Journal of Transportation Engineering, Part A: Systems*, *148*(2). https://doi.org/10.1061/JTEPBS.0000632

Ali, Y., Haque, M. M., Zheng, Z., Washington, S., & Yildirimoglu, M. (2019). A hazard-based duration model to quantify the impact of connected driving environment on safety during mandatory lane-changing. *Transportation Research Part C: Emerging Technologies*, *106*(June), 113–131. https://doi.org/10.1016/j.trc.2019.07.015

Ariannezhad, A., Karimpour, A., Qin, X., Wu, Y.-J., & Salmani, Y. (2021). Handling Imbalanced Data for Real-Time Crash Prediction: Application of Boosting and Sampling Techniques. *Journal of Transportation Engineering, Part A: Systems*, *147*(3), 04020165. https://doi.org/10.1061/JTEPBS.0000499

Bagloee, S. A., & Asadi, M. (2016). Crash analysis at intersections in the CBD: A survival analysis model. *Transportation Research Part A: Policy and Practice*, *94*, 558–572. https://doi.org/10.1016/j.tra.2016.10.019

Barua, S., El-Basyouny, K., & Islam, Md. T. (2016). Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic Methods in Accident Research*, *9*, 1–15. https://doi.org/10.1016/j.amar.2015.11.002

Baruya, A. (1998). Road Safety in Europe. *9th Internation Conference: Road Safety in Europe*.

Bashir, S., & Zlatkovic, M. (2021). Assessment of Queue Warning Application on Signalized Intersections for Connected Freight Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(10), 1211–1221. https://doi.org/10.1177/03611981211015247

Benekohal, R. F., Hajbabaie, A., Medina, J. C., Wang, M.-H., & Chitturi, M. V. (2010). *SPEED PHOTO-RADAR ENFORCEMENT EVALUATION IN ILLINOIS WORK ZONES* (FHWA-ICT-10-064). Illinois Department of Transportation.

Berthaume, A. L. (2015). *Microscopic Modeling of Driver Behavior Based on Modifying Field Theory for Work Zone Application* [Doctoral Dissertation, University of Massachusetts Amherst]. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1328&context=dissertations_2

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2011). Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. *2011 World Congress on Information and Communication Technologies*, 1241–1246. https://doi.org/10.1109/WICT.2011.6141426

Bham, G. H., & Mohammadi, M. A. (2011). *Evaluation of Work Zone Speed Limits: An Objective and Subjective Analysis of Work Zones in Missouri Report*. 92.

Brownstone, D., & Small, K. A. (1989). Efficient Estimation of Nested Logit models. *Journal of Business & Economic Statistics*, *7*(1), 67–74. https://doi.org/10.1080/07350015.1989.10509714

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., & Wu, Y. (2020). Real-time crash prediction on expressways using deep generative models. *Transportation Research Part C: Emerging Technologies*, *117*, 102697. https://doi.org/10.1016/j.trc.2020.102697

Cerwick, D. M., Gkritza, K., Shaheed, M. S., & Hans, Z. (2014). A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, *3–4*, 11–27. https://doi.org/10.1016/j.amar.2014.09.002

Cestac, J., Paran, F., & Delhomme, P. (2011). Young drivers' sensation seeking, subjective norms, and perceived behavioral control and their roles in predicting speeding intention: How risk-taking motivations evolve with gender and driving experience. *Safety Science*, *49*(3), 424–432. https://doi.org/10.1016/j.ssci.2010.10.007

Chang, H. L., & Jovanis, P. P. (1990). Formulating accident occurrence as a survival process. *Accident Analysis and Prevention*, *22*(5), 407–419. https://doi.org/10.1016/0001-4575(90)90037-L

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, *43*(8), 541–557. https://doi.org/10.1016/j.ssci.2005.04.004

Chang, Y., & Edara, P. (2018). Predicting hazardous events in work zones using naturalistic driving data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-March*, 1–6. https://doi.org/10.1109/ITSC.2017.8317847

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, *90*, 128–139. https://doi.org/10.1016/j.aap.2016.02.011

Cheng, W., Gill, G. S., Dasu, R., Xie, M., Jia, X., & Zhou, J. (2017). Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis & Prevention*, *99*, 330–341. https://doi.org/10.1016/j.aap.2016.11.022

Cheng, Z., Lu, J., Zu, Z., & Li, Y. (2019). Speeding Violation Type Prediction Based on Decision Tree Method: A Case Study in Wujiang, China. *Journal of Advanced Transportation*, *2019*, 1–10. https://doi.org/10.1155/2019/8650845

Choudhary, P., & Velaga, N. R. (2020). Impact of distraction on decision making at the onset of yellow signal. *Transportation Research Part C: Emerging Technologies*, *118*(March 2019), 102741. https://doi.org/10.1016/j.trc.2020.102741

Chung, Y. (2010). Development of an accident duration prediction model on the Korean Freeway Systems. *Accident Analysis and Prevention*, *42*(1), 282–289. https://doi.org/10.1016/j.aap.2009.08.005

Data USA: Highway Maintenance Workers. (2018). https://datausa.io/profile/soc/highway-maintenance-workers

Debnath, A. K., Blackman, R., & Haworth, N. (2015). Common hazards and their mitigating measures in work zones: A qualitative study of worker perceptions. *Safety Science*, *72*, 293–301. https://doi.org/10.1016/j.ssci.2014.09.022

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113

Dissanayake, S., & Akepati, S. R. (2009). *Identification of Work Zone Crash Characteristics*. Federal Highway Administration. https://intrans.iastate.edu/app/uploads/2018/08/Dissanayake_WZCrashChar.pdf

Dissanayake, S., & Lu, J. (2002). Analysis of Severity of Young Driver Crashes: Sequential Binary Logistic Regression Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, *1784*(1), 108–114. https://doi.org/10.3141/1784-14

Dong, C., Clarke, D. B., Yan, X., Khattak, A., & Huang, B. (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate

crash frequencies at intersections. *Accident Analysis & Prevention*, *70*, 320–329. https://doi.org/10.1016/j.aap.2014.04.018

Elliott, M. A., & Thomson, J. A. (2010). The social cognitive determinants of offending drivers' speeding behaviour. *Accident Analysis & Prevention*, *42*(6), 1595–1605. https://doi.org/10.1016/j.aap.2010.03.018

Eseonu, C., Gambatese, J., & Nnaji, C. (2018). *Reducing Highway Fatalities Through Improved Adoption of Safety Technologies*.

Federal Highway Administration. (2009a). *Manual of Traffic Control Devices for Streets and Highways*.

Federal Highway Administration. (2009b). *Manual on Uniform Traffic Control Devices (MUTCD)*. https://mutcd.fhwa.dot.gov/

Federal Highway Administration. (2023). *FHWA Work Zone Facts and Statistics*. Work Zone Management Program. https://ops.fhwa.dot.gov/wz/resources/facts_stats.htm

Flannagan, C. A., Selpi, Baykas, P. B., Leslie, A., Kovaceva, J., & Thomson, R. (2019). *Analysis of SHRP2 Data to Understand Normal and Abnormal Driving Behavior in Work Zones* (FHWA-HRT-20-010). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/48835

Forward, S. E. (2009). The theory of planned behaviour: The role of descriptive norms and past behaviour in the prediction of drivers' intentions to violate. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 198–207. https://doi.org/10.1016/j.trf.2008.12.002

Fountas, G., & Anastasopoulos, P. Ch. (2017). A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. *Analytic Methods in Accident Research, 15*, 1–16. https://doi.org/10.1016/j.amar.2017.03.002

Furth, P. G. (2011). *Sampling and Estimation Techniques for Estimating Bus System Passenger-Miles*. Bureau of Transportation Statistics. https://www.bts.gov/archive/publications/journal_of_transportation_and_statistics/volume_08_number_02/paper_07/index

Furth, P. G., Killough, K. L., & Ruprecht, G. F. (1988). Cluster Sampling Techniques for Estimating Transit Patronage. *Transportation Research Record, 1165*.

Gambatese, J. A., Lee, H. W., & Nnaji, C. A. (2017). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance*. Oregon State University School of Civil and Construction Engineering.

Gambatese, J., & Lee, H. W. (2016). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance II*. (Issue 503).

Gan, H., Wei, J., & Wang, G. (2021). A generic work zone evaluation tool driven by a macroscopic traffic simulation model. *International Journal of Mobile Communications*, *19*(1), 1. https://doi.org/10.1504/IJMC.2021.111884

Garber, N. J., & Ehrhart, A. A. (2000). Effect of Speed, Flow, and Geometric Characteristics on Crash Frequency for Two-Lane Highways. *Transportation Research Record: Journal of the Transportation Research Board*, *1717*(1), 76–83. https://doi.org/10.3141/1717-10

Gelman, A., & Hill, J. (2007). When does a multilevel modeling make a difference? In *Data Analysis Using Regression and Multilevel/Hierarchical Models* (pp. 237–249). Cambridge University Press.

Golob, T. F., Recker, W. W., & Leonard, J. D. (1987). An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention*, *19*(5), 375–395. https://doi.org/10.1016/0001-4575(87)90023-6

Guo, H., Wang, W., Guo, W., & Zhao, F. (2013). Modeling lane-keeping behavior of bicyclists using survival analysis approach. *Discrete Dynamics in Nature and Society*, *2013*. https://doi.org/10.1155/2013/197518

Hamdar, S. H., Khoury, H., & Zehtabi, S. (2016). A simulator-based approach for modeling longitudinal driving behavior in construction work zones: Exploration and assessment. *SIMULATION*, *92*(6), 579–594. https://doi.org/10.1177/0037549716644515

Haque, K., Mishra, S., & Golias, M. M. (2021). Multi-period transportation network investment decision making and policy implications using econometric framework. *Research in Transportation Economics*, *89*, 101109. https://doi.org/10.1016/j.retrec.2021.101109

Haque, M. M., & Washington, S. (2015). The impact of mobile phone distraction on the braking behaviour of young drivers: A hazard-based duration model. *Transportation Research Part C: Emerging Technologies*, *50*, 13–27. https://doi.org/10.1016/j.trc.2014.07.011

Harb, R., Radwan, E., Yan, X., Pande, A., & Abdel-Aty, M. (2008). Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression. *Journal of Transportation Engineering*, *134*(5), 203–214. https://doi.org/10.1061/(ASCE)0733-947X(2008)134:5(203)

Harmon, T., Bahar, G., & Gross, F. (2018). *Crash Costs for Highway Safety Analysis*.

Hashmienejad, S. H. A., & Hasheminejad, S. M. H. (2017). Traffic accident severity prediction using a novel multi-objective genetic algorithm. *International Journal of Crashworthiness*, *22*(4), 425–440. https://doi.org/10.1080/13588265.2016.1275431

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second). John Wiley & Sons, Inc.

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019a). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019b). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66–84. https://doi.org/10.1016/j.aap.2018.12.022

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

Hou, G., & Chen, S. (2019). An Improved Cellular Automaton Model for Work Zone Traffic Simulation Considering Realistic Driving Behavior. *Journal of the Physical Society of Japan*, *88*(8), 084001. https://doi.org/10.7566/JPSJ.88.084001

Hourdos, J. (2012). Portable, Non-Intrusive Advance Warning Devices for Work Zones with or without Flag Operators. *Minnesota Department of Transportation*, *October*.

Imprialou, M. I. M., Quddus, M., Pitfield, D. E., & Lord, D. (2016). Re-visiting crash-speed relationships: A new perspective in crash modelling. *Accident Analysis and Prevention*, *86*, 173–185. https://doi.org/10.1016/j.aap.2015.10.001

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Jonathan, A.-V., Wu, K.-F. (Ken), & Donnell, E. T. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention*, *87*, 8–16. https://doi.org/10.1016/j.aap.2015.11.006

Jovanis, P. P., & Chang, H. L. (1989). Disaggregate model of highway accident occurrence using survival theory. *Accident Analysis and Prevention*, *21*(5), 445–458. https://doi.org/10.1016/0001-4575(89)90005-5

Jovanović, D., Šraml, M., Matović, B., & Mićić, S. (2017). An examination of the construct and predictive validity of the self-reported speeding behavior model. *Accident Analysis & Prevention*, *99*, 66–76. https://doi.org/10.1016/j.aap.2016.11.015

Jung, S., Qin, X., & Noyce, D. A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis & Prevention*, *42*(1), 213–224. https://doi.org/10.1016/j.aap.2009.07.020

Kashyap, A. A., Raviraj, S., Devarakonda, A., Nayak K, S. R., K V, S., & Bhat, S. J. (2022). Traffic flow prediction models – A review of deep learning techniques. *Cogent Engineering*, *9*(1), 2010510. https://doi.org/10.1080/23311916.2021.2010510

Ke, J., Zhang, S., Yang, H., & Chen, X. (Michael). (2019). PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. *Transportmetrica A: Transport Science*, *15*(2), 872–895. https://doi.org/10.1080/23249935.2018.1542414

Keramati, A., Lu, P., Zhou, X., & Tolliver, D. (2020). A Simultaneous Safety Analysis of Crash Frequency and Severity for Highway-Rail Grade Crossings: The Competing Risks Method. *Journal of Advanced Transportation*, *2020*(1). https://doi.org/10.1155/2020/8878911

Khasnabis, S., Mishra, S., & Safi, C. (2012). Evaluation procedure for mutually exclusive highway safety alternatives under different policy objectives. *Journal of Transportation Engineering*, *138*(7), 940–948. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000397

Khattak, A. J., Khattak, A. J., & Council, F. M. (2002). Effects of work zone presence on injury and non-injury crashes. *Accident Analysis and Prevention*, *34*(1), 19–29. https://doi.org/10.1016/S0001-4575(00)00099-3

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, & W. Wong, Eds.; Thid Editi). Springer. https://doi.org/10.1007/978-1-4419-6646-9

Kloeden, C. N., McLean, J., & Glonek, G. F. V. (2002). *Reanalysis of travelling speed and the risk of crash involvement in Adelaide South Australia*. Australian Transport Safety Bureau.

Kock, N., & Lynn, G. S. (2012). Lateral Collinearity and Misleading Results in Variance-Based SEM : An Illustration and Recommendations Lateral Collinearity and Misleading Results in Variance-. *Journal of the Association for Information Systems*, *13*(7), 546–580.

Lee, C., Hellinga, B., & Saccomanno, F. (2003). Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, *1840*(1), 67–77. https://doi.org/10.3141/1840-08

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *Applied Sciences*, *10*(1), 129. https://doi.org/10.3390/app10010129

Lee, J.-T., & Fazio, J. (2005). Influential Factors in Freeway Crash Response and Clearance Times by Emergency Management Services in Peak Periods. *Traffic Injury Prevention*, *6*(4), 331–339. https://doi.org/10.1080/15389580500255773

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis & Prevention*, *135*, 105371. https://doi.org/10.1016/j.aap.2019.105371

Li, Y., & Bai, Y. (2008). Development of crash-severity-index models for the measurement of work zone risk levels. *Accident Analysis and Prevention*, *40*(5), 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012

Li, Y., & Bai, Y. (2009). Highway work zone risk factors and their impact on crash severity. *Journal of Transportation Engineering*, *135*(10), 694–701. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000055

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident Analysis and Prevention*, *111*(November 2017), 354–363. https://doi.org/10.1016/j.aap.2017.11.028

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

Ma, J., & Kockelman, K. (2006a). Crash frequency and severity modeling using clustered data from Washington state. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, October*, 1621–1626. https://doi.org/10.1109/itsc.2006.1707456

Ma, J., & Kockelman, K. M. (2006b). Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record: Journal of the Transportation Research Board*, *1950*, 24–34.

Ma, J., Kockelman, K. M., & Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*, *40*(3), 964–975. https://doi.org/10.1016/j.aap.2007.11.002

Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research*, *25*, 100113. https://doi.org/10.1016/j.amar.2020.100113

Marks, E., Vereen, S., & Awolusi, I. (2017). *Active Work Zone Safety Using Emerging Technologies 2017*. University Transportation Center for Alabama The University of Alabama.

Martin, J., Rozas, A., & Araujo, A. (2016). A WSN-Based Intrusion Alarm System to Improve Safety in Road Work Zones. *Journal of Sensors*, *2016*, 1–8. https://doi.org/10.1155/2016/7048141

Medina-Salgado, B., Sánchez-DelaCruz, E., Pozos-Parra, P., & Sierra, J. E. (2022). Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, *35*, 100739. https://doi.org/10.1016/j.suscom.2022.100739

Meng, Q., & Weng, J. (2011). A Genetic algorithm approach to assessing work zone casualty risk. *Safety Science*, *49*(8–9), 1283–1288. https://doi.org/10.1016/j.ssci.2011.05.001

Mishra, S. (2013). A Synchronized Model for Crash Prediction and Resource Allocation to Prioritize Highway Safety Improvement Projects. *Procedia - Social and Behavioral Sciences*, *104*, 992–1001. https://doi.org/10.1016/j.sbspro.2013.11.194

Mishra, S., Golias, M. M., Sharma, S., & Boyles, S. D. (2015). Optimal funding allocation strategies for safety improvements on urban intersections. *Transportation Research Part A: Policy and Practice*, *75*, 113–133. https://doi.org/10.1016/j.tra.2015.03.001

Mishra, S., Golias, M. M., & Thapa, D. (2021). *Work Zone Alert Systems*. Tennessee Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56274

Mohan, D., Bangdiwala, S. I., & Villaveces, A. (2017). Urban street structure and traffic safety. *Journal of Safety Research*, *62*, 63–71. https://doi.org/10.1016/j.jsr.2017.06.003

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis. *Transportation Research Record*, *2673*(11), 680–692. https://doi.org/10.1177/0361198119845899

Nam, D., & Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, *34*(2), 85–102. https://doi.org/10.1016/S0965-8564(98)00065-2

Novosel, C. (2014). Evaluation of Advanced Safety Perimeter Systems for Kansas Temporary Work Zones. In *Civil, Environmental, and Architectural Engineering, University of Kansas*.

Osman, M., Mishra, S., & Paleti, R. (2018a). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group

differences. *Accident Analysis and Prevention*, *118*(May), 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018b). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., & Paleti, R. (2018c). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*, 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., Paleti, R., & Golias, M. (2019). Impacts of Work Zone Component Areas on Driver Injury Severity. *Journal of Transportation Engineering, Part A: Systems*, *145*(8), 04019032. https://doi.org/10.1061/jtepbs.0000253

Osman, M., Paleti, R., & Mishra, S. (2018a). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis and Prevention*, *111*(May 2017), 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., & Mishra, S. (2018b). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis & Prevention*, *111*, 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis and Prevention*, *97*, 261–273. https://doi.org/10.1016/j.aap.2016.10.020

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051–10072. https://doi.org/10.1007/s12652-020-02759-5

Ozturk, O., Ozbay, K., Yang, H., & Bartin, B. (2013). Crash Frequency Modeling for Highway Construction Zones. *Transportation Research Board's 92nd Annual Meeting, Washington, D.C*, 14p.

Paleti, R., Mahmud, A., Gayah, V., & Pinjari, A. (2021). When and Where does the Next Traffic Crash Occur? A Discretized Duration Based Modeling Approach. *Under Review for Publication*.

Park, E. S., & Lord, D. (2007). Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record*, *2019*, 1–6. https://doi.org/10.3141/2019-01

Pei, X., Wong, S. C., & Sze, N. N. (2011). A joint-probability approach to crash prediction models. *Accident Analysis & Prevention*, *43*(3), 1160–1166. https://doi.org/10.1016/j.aap.2010.12.026

Pham, M.-H., Bhaskar, A., Chung, E., & Dumont, A.-G. (2010). Random forest models for identifying motorway Rear-End Crash Risks using disaggregate data. *13th International IEEE Conference on Intelligent Transportation Systems*, 468–473. https://doi.org/10.1109/ITSC.2010.5625003

Provost, F., Jensen, D., & Oates, T. (2001). Progressive Sampling. In *Instance Selection and Construction for Data Mining* (pp. 151–170). Springer US. https://doi.org/10.1007/978-1-4757-3359-4_9

Qi, Y., Srinivasan, R., Teng, H., & Baker, R. F. (2005). *Frequency of Work Zone Accidents on Construction Projects*.

Quddus, M. (2013). Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, *5*(1), 27–45. https://doi.org/10.1080/19439962.2012.705232

Rahim, M. A., & Hassan, H. M. (2021). A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention*, *154*, 106090. https://doi.org/10.1016/j.aap.2021.106090

Rahman, R., Bhowmik, T., Eluru, N., & Hasan, S. (2021). Assessing the crash risks of evacuation: A matched case-control approach applied over data collected during Hurricane Irma. *Accident Analysis & Prevention*, *159*, 106260. https://doi.org/10.1016/j.aap.2021.106260

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007

Sarker, A. A., Naimi, A., Mishra, S., Golias, M. M., & Freeze, P. B. (2015). Development of a Secondary Crash Identification Algorithm and occurrence pattern determination in large scale multi-facility transportation network. *Transportation Research Part C: Emerging Technologies*, *60*, 142–160. https://doi.org/10.1016/j.trc.2015.08.011

Scott-Parker, B., Hyde, M. K., Watson, B., & King, M. J. (2013). Speeding by young novice drivers: What can personal characteristics and psychosocial theory add to our understanding? *Accident Analysis & Prevention*, *50*, 242–250. https://doi.org/10.1016/j.aap.2012.04.010

Shangguan, Q., Fu, T., & Liu, S. (2020). Investigating rear-end collision avoidance behavior under varied foggy weather conditions: A study using advanced driving simulator and survival analysis. *Accident Analysis and Prevention*, *139*(March), 105499. https://doi.org/10.1016/j.aap.2020.105499

Sharma, A., Bullock, D., & Peeta, S. (2011). Estimating dilemma zone hazard function at high speed isolated intersection. *Transportation Research Part C: Emerging Technologies*, *19*(3), 400–412. https://doi.org/10.1016/j.trc.2010.05.002

Simons-Morton, B. G., Ouimet, M. C., Chen, R., Klauer, S. G., Lee, S. E., Wang, J., & Dingus, T. A. (2012). Peer influence predicts speeding prevalence among teenage drivers. *Journal of Safety Research*, *43*(5–6), 397–403. https://doi.org/10.1016/j.jsr.2012.10.002

Song, J. J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, *97*(1), 246–273. https://doi.org/10.1016/j.jmva.2005.03.007

Stipancic, J., Miranda-Moreno, L., Saunier, N., & Labbe, A. (2019). Network screening for large urban road networks: Using GPS data and surrogate measures to model crash frequency and severity. *Accident Analysis and Prevention*, *125*(February), 290–301. https://doi.org/10.1016/j.aap.2019.02.016

Stout, D., Graham, J., Bryant-Fields, B., Migletz, J., Fish, J., & Hanscom, F. (1993). *Maintenance Work Zone Safety Devices Development and Evaluation*.

Sun, J., & Sun, J. (2016a). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Sun, J., & Sun, J. (2016b). Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems*, *10*(5), 331–337. https://doi.org/10.1049/iet-its.2014.0288

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y., & Huang, H. (2020). Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Analytic Methods in Accident Research*, *27*, 100123. https://doi.org/10.1016/j.amar.2020.100123

Thapa, D., & Mishra, S. (2021a). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., & Mishra, S. (2021b). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*, 106125. https://doi.org/10.1016/j.aap.2021.106125

Thapa, D., Paleti, R., & Mishra, S. (2022a). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Thapa, D., Paleti, R., & Mishra, S. (2022b). Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. *Accident Analysis & Prevention*, *169*, 106639. https://doi.org/10.1016/j.aap.2022.106639

Theiss, L., Ullman, G. L., & Lindheimer, T. (2017). *Closed Course Performance Testing of the Aware Intrusion Alarm System*.

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2019). Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention*, *130*, 151–159. https://doi.org/10.1016/j.aap.2017.12.018

Therneau, T. M. (2020). *A Package for Survival Analysis in R. R package version 3.2-7*.

Therneau, T. M., Grambsch, P. M., & Panktatz, S. V. (2003). Penalized Survival Models and Frailty. *Journal of Computational and Penalized Survival Models and Frailty*, *12*(1), 156–175.

Ullman, G. L., Trout, N. D., & Theiss, L. (2016). *Driver Responses to the AWARE Intrusion Alarm System*. Texas A&M Transportation Institute.

Venugopal, S., & Tarko, A. (2000). Safety models for rural freeway work zones. *Transportation Research Record*, *1715*, 1–9. https://doi.org/10.3141/1715-01

Wang, C., Quddus, M. A., & Ison, S. G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention*, *43*(6), 1979–1990. https://doi.org/10.1016/j.aap.2011.05.016

Wang, J., Yamamoto, T., & Liu, K. (2020). Key determinants and heterogeneous frailties in passenger loyalty toward customized buses: An empirical investigation of the subscription termination hazard of users. *Transportation Research Part C: Emerging Technologies*, *115*(July 2019), 102636. https://doi.org/10.1016/j.trc.2020.102636

Wang, X., Katz, R., & Dong, X. S. (2018). *Fatal Injuries at Road Construction Sites among Construction Workers* [Quarterly]. Center for Construction Research and Training.

https://www.cpwr.com/wp-content/uploads/publications/publications_Quarter2-QDR-2018.pdf

Work Zones-Injury Facts-National Safety Council. (2020). https://injuryfacts.nsc.org/motor-vehicle/motor-vehicle-safety-issues/work-zones/

Wu, L., Meng, Y., Kong, X., & Zou, Y. (2020). Incorporating survival analysis into the safety effectiveness evaluation of treatments: Jointly modeling crash counts and time intervals between crashes. *Journal of Transportation Safety and Security*, *0*(0), 1–21. https://doi.org/10.1080/19439962.2020.1786871

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

Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., & Jiang, X. (2020). A machine-learning method for improving crash injury severity analysis: A case study of work zone crashes in Cairo, Egypt. *International Journal of Injury Control and Safety Promotion*, *27*(3), 266–275. https://doi.org/10.1080/17457300.2020.1746814

Yang, H., Ozbay, K., Ozturk, O., & Xie, K. (2015). Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. *Traffic Injury Prevention*, *16*(4), 387–396. https://doi.org/10.1080/15389588.2014.948615

Yasmin, S., & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention*, *59*, 506–521. https://doi.org/10.1016/j.aap.2013.06.040

Yasmin, S., & Eluru, N. (2018). A joint econometric framework for modeling crash counts by severity. *Transportmetrica A: Transport Science*, *14*(3), 230–255. https://doi.org/10.1080/23249935.2017.1369469

Yasmin, S., Eluru, N., Bhat, C. R., & Tay, R. (2014). A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic Methods in Accident Research*, *1*, 23–38. https://doi.org/10.1016/j.amar.2013.10.002

Yasmin, S., Eluru, N., Wang, L., & Abdel-Aty, M. A. (2018). A joint framework for static and real-time crash risk analysis. *Analytic Methods in Accident Research*, *18*, 45–56. https://doi.org/10.1016/j.amar.2018.04.001

Ye, X., Pendyala, R. M., Shankar, V., & Konduri, K. C. (2013). A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention*, *57*, 140–149. https://doi.org/10.1016/j.aap.2013.03.025

Yu, B., Chen, Y., & Bao, S. (2019). Quantifying visual road environment to establish a speeding prediction model: An examination using naturalistic driving data. *Accident Analysis & Prevention*, *129*, 289–298. https://doi.org/10.1016/j.aap.2019.05.011

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, *51*, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis and Prevention*, *73*, 351–358. https://doi.org/10.1016/j.aap.2014.09.006

Zhang, C., He, J., Wang, Y., Yan, X., Zhang, C., Chen, Y., Liu, Z., & Zhou, B. (2020). A Crash Severity Prediction Method Based on Improved Neural Network and Factor Analysis. *Discrete Dynamics in Nature and Society*. https://doi.org/10.1155/2020/4013185

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods. *IEEE Access*, *6*, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*(1), 204–215. https://doi.org/10.1016/j.ijforecast.2010.06.002

Zhang, K., & Hassan, M. (2019a). Crash severity analysis of nighttime and daytime highway work zone crashes. *PLoS ONE*, *14*(8), 1–17. https://doi.org/10.1371/journal.pone.0221128

Zhang, K., & Hassan, M. (2019b). Identifying the Factors Contributing to Injury Severity in Work Zone Rear-End Crashes. *Journal of Advanced Transportation*, *2019*, 1–9. https://doi.org/10.1155/2019/4126102

Zhao, G., Wu, C., & Qiao, C. (2013). A Mathematical Model for the Prediction of Speeding with its Validation. *IEEE Transactions on Intelligent Transportation Systems*, *14*(2), 828–836. https://doi.org/10.1109/TITS.2013.2257757

Zheng, L., & Sayed, T. (2020). A novel approach for real time crash prediction at signalized intersections. *Transportation Research Part C: Emerging Technologies*, *117*, 102683. https://doi.org/10.1016/j.trc.2020.102683

Zimmerman, K., Mzige, A. A., Kibatala, P. L., Museru, L. M., & Guerrero, A. (2012). Road traffic injury incidence and crash characteristics in Dar es Salaam: A population based

study. *Accident Analysis & Prevention*, *45*, 204–210.

https://doi.org/10.1016/j.aap.2011.06.018

6. Conclusion

Traffic safety poses a significant concern in industrialized nations, where it remains a leading cause of fatalities, particularly among younger individuals. While advances in vehicular technologies have contributed to a reduction in crashes to some extent, transportation planners and safety officials stand to benefit even more by gaining the ability to predict crashes, or at the very least, identify hazardous traffic conditions and their impact on safety. In this context, the current dissertation aimed to investigate the application of duration models for diagnostic and predictive analysis. Furthermore, it introduced a novel approach for the diagnostic analysis of crashes and developed an innovative econometric framework for a real-time crash prediction model based on duration models. These accomplishments were collectively attained through four distinct studies.

The first study aimed to enhance our understanding of the role of Work Zone Intrusion Alert Systems (WZIAS) in improving work zone safety by conducting field experiments. This research identified three commonly used WZIAS and examined how their placement impacted worker responses. To analyze the results, the study employed duration models. These models were used to assess the time it took for workers to respond when WZIAS were present. Non-parametric Kaplan-Meier estimators and semi-parametric Cox Proportional Hazard models were utilized for this analysis.

The study's findings led to actionable recommendations for optimizing the use and placement of these systems, along with suggested modifications to the existing Manual on Uniform Traffic Control Devices (MUTCD) guidelines regarding work zone setup, all aimed at enhancing work zone safety. Specifically, the study recommends reducing the posted speed limit and increasing the buffer length as necessary measures to improve work zone safety. The

placement of WZIAS within the work zone, including the specification of buffer length, should be determined based on the transmission range of the WZIAS (work zone coverage).

The second study implemented a real-time crash prediction framework based on a duration model. This framework established a direct relationship between choice models and duration models. To achieve this, the study proposed reformulation of historical crash data by discretizing the time duration between crashes into more practical forecasting window, specifically one-hour epochs and 15-minute intervals. By doing so, it is possible to treat the outcomes within each time interval as alternatives in a Multinomial Logit (MNL) model. This innovative approach allowed for the incorporation of dynamic covariates since each alternative could be associated with the respective dynamic covariate values for that time interval. This model framework was put into practice using crash data observed in 2019 on interstates I-40 and I-55 in Memphis, Tennessee. These interstates were segmented based on speed limits, the number of lanes, and terrain types for analysis. Dynamic covariates related to traffic flow, such as speed and volume, for these segment forecasting windows were collected from RDS stations at 15-minute intervals.

In total, 1,174 crashes were utilized for analysis after excluding those near entry and exit ramps. The mean duration of crashes was found to be 652 hours. However, the reformulation of the data significantly increased the dataset size, resulting in approximately 700,000 observations. This complexity in data size posed challenges to model training, which was essential for achieving real-time crash prediction. To address this challenge, the study explored various sampling techniques to identify the most effective strategy for reducing estimation complexity while maintaining reasonably accurate parameter estimates. It was found that sampling 25% of the data at the epoch level proved to be the ideal approach. This approach reduced estimation

time to approximately one-fourth of the original time, with only a marginal 1.25% change in the predicted log-likelihood compared to the model trained on the full dataset.

In a subsequent study, the duration-based crash prediction framework was expanded to encompass crash severities. To achieve this, the MNL model was adapted into a two-level nested logit model, where crash occurrences and severities were considered in the upper and lower levels, respectively. Furthermore, the analysis was extended to include two additional interstates in Chattanooga, namely I-24 and I-75, extending the scope for better generalization of findings. Given the need for smaller samples to address estimation complexity issues, the study investigated the impact of sampling on parameter estimates to identify which covariates were most sensitive to sampling. This investigation was particularly crucial from a stability perspective, helping to better understand the effect of sampling on the consistency and accuracy of parameter estimates.

The findings indicated that drawing a 15% sample at the epoch level struck a balanced approach, reducing the dataset size while maintaining reasonable predictive accuracy. A stability analysis of predictor variables across different samples revealed that certain variables, such as Time of day (Early afternoon), Weather condition (Clear), Lighting condition (Daytime), Illumination (Illuminated), and Volume, required larger samples for more accurate coefficient estimation. Conversely, variables like Daytime (Early morning, Late morning, Late afternoon), Lighting condition (Dark lighted), Terrain (Flat), Land use (Commercial, Rural), Number of lanes, and Speed converged towards true estimates with small incremental increases in sample size. Furthermore, the model's predictions regarding crash occurrence were validated by introducing a novel metric known as Predicted Temporal Proximity (PTP). This metric quantifies how closely the predicted crash time-interval aligns with the actual crash time-interval in the test

sample. A lower PTP value suggests that the predicted time-interval is closer to the actual time-interval. Additionally, Sensitivity and Specificity were employed to evaluate the model's performance in terms of false positive and false negative predictions for crash severities. The results indicate that the model framework provides a reasonably accurate predictive capability. Notably, the rate of false positives is lower when compared to the rate of false negatives, reflecting high Specificity and low Sensitivity. It's worth highlighting that the accuracy of crash occurrence prediction is influenced by the duration between consecutive crashes. In other words, the prediction is more accurate for roadway segments where crashes occur more frequently, resulting in shorter inter-crash durations. For example, a PTP of 60% was achieved for crashes occurring within 100 epochs, whereas a PTP of 74% was obtained for crashes occurring within 1,000 epochs.

In the final study, the duration-based crash prediction framework was applied to forecast instances of speeding in work zones, showcasing its versatility. This involved utilizing speed data from INRIX for 14 Traffic Message Channel segments of I-65. The time intervals between speeding incidents were discretized into one-hour epochs and 15-minute intervals. Notably, this study incorporated local climatological data as a dynamic covariate in the model. Furthermore, to account for segment-specific effects, the prediction framework was estimated as a mixed model.

The study identified several significant predictors of speeding, including visibility, the number of lanes, posted speed limit, segment length, coefficient of variation in speed, and travel time index. Among these variables, the number of lanes, posted speed limit, and coefficient of variation in speed showed positive associations with speeding. In contrast, visibility, segment length, and travel time index exhibited negative associations with speeding. The results also indicated strong evidence of segment-specific unobserved heterogeneity. Similar to the previous

study, the model's validation using the PTP metric suggested that the model's predictions are more accurate for segments where the duration between speeding incidents is shorter. Sensitivity and Specificity metrics also yielded similar findings, with high Specificity and low Sensitivity in the model's predictions.

While this dissertation made significant strides in addressing limitations, there remain several promising avenues for future research. Firstly, an ongoing challenge in crash analysis is the disproportionate presence of "non-crash" events compared to actual "crashes" in the data. This issue, which was observed in all the studies within this dissertation, becomes particularly pronounced in the duration-based crash prediction framework due to the data reformulation. Therefore, future studies should consider implementing data balancing techniques before model training and analysis. One such technique is the Synthetic Minority Oversampling Technique (SMOTE), which can help balance the dataset and potentially improve parameter estimates and prediction accuracy. Secondly, it is worthwhile to explore more efficient approaches to model estimation that allow for the use of complete data rather than relying on sampled data. One avenue to investigate is model training using parallel or distributed computing. This approach could enhance the ability to calculate crash probabilities in real-time and assess the model's effectiveness through real-world implementation. By leveraging advanced computing techniques, researchers may be able to achieve more robust and scalable models for traffic safety prediction.

References

2019 Highway Work Zone Safety Survey. (2019). Associated General Contractors of America. https://www.agc.org/news/2019/05/23/2019-highway-work-zone-safety-survey

Awolusi, I., & Marks, E. D. (2019). Active Work Zone Safety: Preventing Accidents Using Intrusion Sensing Technologies. *Frontiers in Built Environment*, *5*. https://doi.org/10.3389/fbuil.2019.00021

Bagloee, S. A., & Asadi, M. (2016). Crash analysis at intersections in the CBD: A survival analysis model. *Transportation Research Part A: Policy and Practice*, *94*, 558–572. https://doi.org/10.1016/j.tra.2016.10.019

Chang, H. L., & Jovanis, P. P. (1990). Formulating accident occurrence as a survival process. *Accident Analysis and Prevention*, *22*(5), 407–419. https://doi.org/10.1016/0001-4575(90)90037-L

Chang, Y., & Edara, P. (2018). Predicting hazardous events in work zones using naturalistic driving data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-March*, 1–6. https://doi.org/10.1109/ITSC.2017.8317847

Gambatese, J. A., Woo, H., Chukwuma, L., & Nnaji, A. (2017). *Work Zone Intrusion Alert Technologies: Assessment and Practical Guidance*. https://rosap.ntl.bts.gov/view/dot/32574

Hughes, J. E., Kaffine, D., & Kaffine, L. (2023). Decline in Traffic Congestion Increased Crash Severity in the Wake of COVID-19. *Transportation Research Record: Journal of the Transportation Research Board*, *2677*(4), 892–903. https://doi.org/10.1177/03611981221103239

Jovanis, P. P., & Chang, H. L. (1989). Disaggregate model of highway accident occurrence using survival theory. *Accident Analysis and Prevention*, *21*(5), 445–458. https://doi.org/10.1016/0001-4575(89)90005-5

Khattak, A. J., Khattak, A. J., & Council, F. M. (2002). Effects of work zone presence on injury and non-injury crashes. *Accident Analysis and Prevention*, *34*(1), 19–29. https://doi.org/10.1016/S0001-4575(00)00099-3

Li, Y., & Bai, Y. (2008). Development of crash-severity-index models for the measurement of work zone risk levels. *Accident Analysis and Prevention*, *40*(5), 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012

Li, Y., & Bai, Y. (2009). Highway work zone risk factors and their impact on crash severity. *Journal of Transportation Engineering*, *135*(10), 694–701. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000055

Ma, J., & Kockelman, K. (2006a). Crash frequency and severity modeling using clustered data from Washington state. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *October*, 1621–1626. https://doi.org/10.1109/itsc.2006.1707456

Ma, J., & Kockelman, K. M. (2006b). Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record: Journal of the Transportation Research Board*, *1950*, 24–34.

Ma, J., Kockelman, K. M., & Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*, *40*(3), 964–975. https://doi.org/10.1016/j.aap.2007.11.002

Marks, E., Vereen, S., & Awolusi, I. (2017). *Active Work Zone Safety Using Emerging Technologies 2017* (FHWA/CA/OR-; p. 36). University Transportation Center for Alabama The University of Alabama. https://trid.trb.org/view/1483615

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis. *Transportation Research Record*, *2673*(11), 680–692. https://doi.org/10.1177/0361198119845899

Nam, D., & Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, *34*(2), 85–102. https://doi.org/10.1016/S0965-8564(98)00065-2

National Highway Traffic Safety Administration. (2022, May 17). *Newly Released Estimates Show Traffic Fatalities Reached a 16-Year High in 2021*. https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities

National Safety Council. (2020). *Work Zones-Injury Facts*. https://injuryfacts.nsc.org/motor-vehicle/motor-vehicle-safety-issues/work-zones/

Osman, M., Mishra, S., & Paleti, R. (2018). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis and Prevention*, *118*(May), 289–300. https://doi.org/10.1016/j.aap.2018.05.004

Osman, M., Mishra, S., Paleti, R., & Golias, M. (2019). Impacts of Work Zone Component Areas on Driver Injury Severity. *Journal of Transportation Engineering, Part A: Systems*, *145*(8), 04019032. https://doi.org/10.1061/jtepbs.0000253

Osman, M., Paleti, R., & Mishra, S. (2018). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis and Prevention*, *111*(May 2017), 161–172. https://doi.org/10.1016/j.aap.2017.11.026

Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis and Prevention*, *97*, 261–273. https://doi.org/10.1016/j.aap.2016.10.020

Ozturk, O., Ozbay, K., Yang, H., & Bartin, B. (2013). Crash Frequency Modeling for Highway Construction Zones. *Transportation Research Board's 92nd Annual Meeting, Washington, D.C*, 14p.

Qi, Y., Srinivasan, R., Teng, H., & Baker, R. F. (2005). *Frequency of Work Zone Accidents on Construction Projects*.

Song, J. J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, *97*(1), 246–273. https://doi.org/10.1016/j.jmva.2005.03.007

Thapa, D., & Mishra, S. (2021). Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. *Accident Analysis and Prevention*, *156*. https://doi.org/10.1016/j.aap.2021.106125

US Department of Transportation. (2020, September 25). *State by State Crash Data and Economic Cost Index*. https://www.transportation.gov/research-and-technology/state-state-crash-data-and-economic-cost-index

Venugopal, S., & Tarko, A. (2000). Safety models for rural freeway work zones. *Transportation Research Record, 1715*, 1–9. https://doi.org/10.3141/1715-01

World Health Organization. (2018). *Global Status Report on Road Safety*. https://www.who.int/publications/i/item/9789241565684

Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., & Jiang, X. (2020). A machine-learning method for improving crash injury severity analysis: A case study of work zone crashes in Cairo, Egypt. *International Journal of Injury Control and Safety Promotion*, *27*(3), 266–275. https://doi.org/10.1080/17457300.2020.1746814

Ye, X., Pendyala, R. M., Shankar, V., & Konduri, K. C. (2013). A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention*, *57*, 140–149. https://doi.org/10.1016/j.aap.2013.03.025

Yellman, M. A., & Sauber-Schatz, E. K. (2022). Motor Vehicle Crash Deaths—United States and 28 Other High-Income Countries, 2015 and 2019. *MMWR. Morbidity and Mortality Weekly Report*, *71*(26), 837–843. https://doi.org/10.15585/mmwr.mm7126a1

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis and Prevention*, *73*, 351–358. https://doi.org/10.1016/j.aap.2014.09.006

Zhang, K., & Hassan, M. (2019). Crash severity analysis of nighttime and daytime highway work zone crashes. *PLoS ONE*, *14*(8), 1–17. https://doi.org/10.1371/journal.pone.0221128