# USE OF HARSH-BRAKING DATA FROM CONNECTED VEHICLES AS A SURROGATE SAFETY MEASURE

by

Nathaniel Patrick Edelmann


A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Civil Engineering

Boise State University

December 2022

BOISE STATE UNIVERSITY GRADUATE COLLEGE

**DEFENSE COMMITTEE AND FINAL READING APPROVALS**

of the thesis submitted by

Nathaniel Patrick Edelmann

Thesis Title: Use of Harsh-Braking Data from Connected Vehicles as a Surrogate Safety Measure

Date of Final Oral Examination: 19 September 2022

The following individuals read and discussed the thesis submitted by student Nathaniel Patrick Edelmann, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

| | |
|---|---|
| Mandar Khanal, Ph.D. | Chair, Supervisory Committee |
| Kyungduk Ko, Ph.D. | Member, Supervisory Committee |
| Yang Lu, Ph.D. | Member, Supervisory Committee |

The final reading approval of the thesis was granted by Mandar Khanal, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

# DEDICATION

My time in graduate school has been encompassed by the Covid pandemic, with healthcare professionals and first responders working tirelessly to care for the sick and save countless lives. Their work put them in harm's way constantly, and they toiled through often unbearable conditions. The dedication and heroism that healthcare professionals and first responders showed during the worst times and continue to show today have been an inspiration to me and so many others. As a small token of gratitude, I dedicate this thesis to them.

ACKNOWLEDGEMENTS

I am grateful to the Boise State University Department of Civil Engineering for my engineering education. It has opened doors for me and introduced me to so many great people. The sense of community within our department has been a great support and source of joy for me and for many of my fellow students. My advisor, Dr. Mandar Khanal, has been a tremendous source of inspiration and support. His vision for this research work and encouragement to me to pursue graduate school has changed my life for the better, and for that I will always be grateful. I would like to thank Dr. Kyungduk Ko and Dr. Yang Lu for being committee members for this thesis and for all that they taught me during my time in their classes.

This thesis would not have been possible without the constant encouragement and support I have received from my family. For as long as I can remember, they have always cheered me on in my endeavors and taught me to give everything my best effort even when a task is incredibly challenging. These lessons have been extremely valuable during graduate school.

I would also like to thank the PacTrans Region 10 University Transportation Center and acknowledge the support received from them which made it possible to obtain the connected vehicle data that was used in this research.

ABSTRACT

Traffic safety may be analyzed with the use of surrogate safety measures, measures of safety that do not incorporate collision data but rather rely on the concept of traffic conflicts. Use of these measures provides several benefits over use of more traditional analysis methods with historical crash data. Surrogate measures eliminate the need to wait for crashes to occur to conduct a safety analysis. The amount of time required for enough crash data to accumulate can be significant, delaying safety analyses. Similarly, these measures allow for safety analysis to be conducted prior to crashes occurring, potentially calling attention to hazardous areas which may be altered to prevent crashes. In addition to these benefits, traffic conflicts occur much more frequently than collisions, generating many more data points which in turn make statistical methods of analysis more effective.

Evaluating surrogate safety measures for a particular transportation network is most effectively done with the use of traffic microsimulation or with connected vehicle data. Traffic microsimulation (such as the use of PTV VISSIM) will generate kinematic data that may then be used for computation of surrogate safety measures. A significant amount of research has been done on this topic, resulting in the establishment of algorithms for calculation of several different surrogate measures and validation of these measures.

Kinematic data from connected vehicles has also been used for the calculation of surrogate safety measures. One data point collected by connected vehicles is harsh

braking events which could serve as a surrogate safety measure. Because drivers usually brake more gently if given the opportunity to do so, harsh braking events indicate that a traffic conflict has occurred or is about to occur. Such events take away the driver's opportunity to brake gently. This research establishes statistical models which relate harsh braking events to crashes on intersections and segments in Salt Lake City, Utah. The findings indicate that harsh braking events have the effect of reducing expected crashes because they represent traffic conflicts which were remedied through the use of harsh braking as an evasive action. The presence of schools and the presence of left turn lanes were also found to be statistically significant crash predictors. In addition to this research work a paper outlining the existing state of safety analysis with surrogate safety measures and evaluating the usefulness and practicality of various existing measures is presented.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

## LIST OF ABBREVIATIONS

1. AADT: Average Annual Daily Traffic
2. ACPM: Aggregated Crash Propensity Metric
3. ADT: Average Daily Traffic
4. AICC: Autonomous Intelligent Cruise Control
5. CPI: Crash Potential Index
6. DeltaS: Maximum Relative Speed of Vehicles in a Conflict
7. DR: Deceleration Rate
8. DRAC: Deceleration Rate to Avoid Collision
9. ET: Encroachment Time
10. GT: Gap Time
11. IAPE: Initially Attempted Post-Encroachment time
12. MaxS: Maximum of the Speeds of vehicles in a conflict
13. MTTC: Modified Time-To-Collision
14. PET: Post-Encroachment Time
15. PSD: Proportion of Stopping Distance
16. SpaT: Signal Phasing and Timing
17. SPMD: Safety Pilot Model Deployment
18. SSAM: Surrogate Safety Assessment Module
19. SSM: Surrogate Safety Measure
20. TCT: Traffic Conflicts Technique

21. TET: Time Exposed Time-to-collision

22. TIT: Time Integrated Time-to-collision

23. TTC: Time-To-Collision

# CHAPTER 1: INTRODUCTION

Traffic safety is of utmost importance to engineers as they design transportation infrastructure. Automobile collisions cause significant damage in the form of financial damages and human injuries and deaths. In spite of the tremendous cost of automobile collisions, crashes are actually quite rare events from a data collection standpoint. Another type of event, known as a traffic conflict, is quite common, making it more useful than crashes for safety analysis. Traffic conflicts include collisions, but the vast majority of conflicts do not result in a collision. In order to understand the safety level of a particular piece of transportation infrastructure, a safety analysis must be conducted on collected data. This can be difficult to do effectively with crash data because of its infrequent nature, but traffic conflict data can supply many more data points and therefore result in a more effective analysis. One indicator of a traffic conflict is a harsh braking event which is recorded by connected vehicles and can be used as a surrogate safety measure.

## 1.1 Background

Many types of surrogate safety measures exist to measure the severity of a traffic conflict. The severity can then be compared against a threshold value to determine if a conflict indeed occurred, or the severity can be aggregated in some way. Some examples of established surrogate safety measures include time to collision (TTC), deceleration rate to avoid collision (DRAC), post-encroachment time (PET), crash potential index (CPI), and proportion of stopping distance (PSD). In addition to these well-established surrogate

measures, some more recently defined and validated measures include the aggregated crash propensity metric (ACPM) from Wang and Stamatiadis (2013) and another metric from Astarita et al. (2020) which considers single vehicle collisions in addition to multiple vehicle collisions and takes collision energy into account to consider collision severity.

Definitions of these surrogate measures are as follows:

- TTC: the amount of time required for a collision to occur given that two vehicles remain on their current path at their current speed
- DRAC: the required deceleration rate that a following vehicle must use to prevent a collision from occurring
- PET: the duration of time between when a leading vehicle leaves the path of travel and when a following vehicle arrives at the point where a collision would have occurred had the leading vehicle remained in place
- CPI: the probability that a following vehicle's DRAC will exceed its braking capacity
- PSD: a ratio between a following vehicle's stopping distance and the distance to a potential collision location
- ACPM: the sum of probabilities of collisions occurring for several conflict types including crossing, lane-change, and rear-end conflicts
- A surrogate measure from Astarita et al.: a regression model which uses mean collision energy from simulated trajectory deviations along with traffic flow and a dummy variable as explanatory variables

Prior research that makes use of these surrogate safety measures usually employs kinematic data from traffic microsimulations. The microsimulation software programs, such as VISSIM, output kinematic data which can then be used to compute the surrogate safety measures previously defined. Microsimulation employs many assumptions, and safety analyses using microsimulation data is only as accurate as the simulation itself. Because of this limitation, research has been conducted into using real-world data from connected vehicles.

Connected vehicle data has been used for surrogate safety measure analysis using established measures, but there are other measures that may be promising. According to the Federal Highway Administration, some indicators of safety that have not yet been quantitatively related to crash counts are: "delay, travel time, approach speed, percent stops, queue length, stop-bar encroachments, red-light violations, percent left turns, spot speed, speed distribution, and deceleration distribution" (Gettman & Head, 2003). Deceleration distribution may be captured in connected vehicle data in the form of harsh braking events. A harsh braking event is an indicator that a traffic conflict occurred to bring about the harsh braking maneuver. This indication suggests that harsh braking events may be quantitatively linked to crash counts as a surrogate safety measure to be employed using connected vehicle data.

## 1.2 Problem Statement

The research contained in this thesis seeks to incorporate jerk data from connected vehicles into crash prediction models for the purpose of improving the traffic safety analysis process. The availability of connected vehicle data has opened up many possibilities for traffic analysis advancements. These advancements are in many areas of

transportation engineering, including safety analysis. Traffic safety analysis is highly meaningful because of the fact that driving is one of the most dangerous activities that the majority of people undertake due to the frequency with which people drive and the consequent high exposure to risk. Identifying risky infrastructure allows for the area to be studied further and remedied.

Jerk data has been shown to have promise as a predictor of crash rates and is available from connected vehicles. Past research has demonstrated the viability of jerk data from GPS units in a small group of vehicles for traffic safety analysis. Connected vehicles are so much more numerous that use of connected vehicle data should be even more illuminating than a small-scale study with GPS units. More importantly, connected vehicle data is collected automatically, meaning that it may be accessed quickly for any area in the United States when needed.

In researching the existing surrogate safety measures, it became clear that there would be value in producing a paper which summarized and compared the methods developed in the literature written on the topic. Such a paper would need to explain the development of surrogate safety measures over time and how these measures differ in the difficulty of implementation, relevance, theoretical value, and practical value. These findings would provide a resource for researchers and practitioners to evaluate which of the available methods for surrogate safety analysis would be best to investigate further or implement in a real-world application.

### 1.3 Thesis Summary

This research involves a review of the development of surrogate safety measures over time, the identification of jerk thresholds for high-jerk events on intersections and

segments, and the development of crash prediction models using connected vehicle data from Salt Lake City. The crash prediction models produced by this research include models from three different model families: Poisson, negative binomial, and generalized Poisson. The models were selected using the statistical significance of the variables contained within them, the residuals generated, and measures such as the Akaike information criterion. The findings from Salt Lake City may be applicable to other metropolitan areas, and this will need to be validated with future research efforts.

The remainder of this thesis is organized as follows. Chapter 2 contains a manuscript which was submitted to the 2022 International Association of Journals and Conferences which reviews the existing research on surrogate safety measures, outlines the evolution of these measures over time, and evaluates the methods available based on a uniform set of criteria. Chapter 3 contains a manuscript which was submitted to the 2023 Transportation Research Board Annual Meeting and for subsequent publication in the *Transportation Research Record*. This manuscript details the research work undertaken using connected vehicle data from Salt Lake City to develop a set of statistical regression models for intersections and segments. Chapter 4 serves as a conclusion with a summary of the research work completed and ideas for future research to build upon these efforts.

# CHAPTER 2: MANUSCRIPT SUBMITTED TO IAJC CONFERENCE

The following manuscript was written by Nathaniel Edelmann and Mandar Khanal and submitted to the International Association of Journals and Conferences for inclusion in their October, 2022 conference in Orlando, Florida and potential inclusion in one of their journals. The manuscript begins on the next page and is titled "Review of Surrogate Safety Measures for Roadway Safety Analysis." The manuscript submitted to the conference conformed to the established formatting standards of the conference; the formatting has been altered here to conform to the formatting in the rest of this document. The content presented here is identical to that which was submitted to the conference.

## 2.1 Abstract

Vehicular collisions are a source of tremendous cost in the form of financial damages, human injuries, and deaths. Because of this, traffic safety is of utmost importance to engineers as they design transportation infrastructure. Traffic safety analysis informs decisions relating to projects intended to bolster the safety of roadways and intersections, and this analysis uses data that is collected for a transportation system network. The traditional method of safety analysis uses collision data, but a newer set of safety analysis methods instead considers data on traffic conflicts as a replacement for collisions. These methods are known as surrogate safety measures (SSMs) which analyze kinematic data to assess safety. Several SSMs have been developed and validated in an effort to capture the risk exposure of vehicles more fully. SSMs offer a range of benefits over traditional analysis with collision data. First, the data used by SSMs may be

collected more rapidly than collision data. Collisions happen quite infrequently from a data collection standpoint, but vehicular kinematic data may be collected in large quantities within a matter of weeks. Second, surrogate safety measures are a proactive analysis, as they allow for safety analysis without collisions occurring. Improvements may be made based upon the results of an SSM analysis to prevent crashes. Third, kinematic data supplies so many more data points as opposed to collisions that statistical analysis for traffic conflicts is significantly more robust. Analysis with SSMs has evolved over the decades from being measured with manual observations in the field and use of time-lapse imagery to use of microsimulation software, with the most recent advancement being the incorporation of connected vehicle data. This paper serves as a summary of the development of SSMs and establishes the state of the practice for surrogate safety analysis.

## 2.2 Introduction

Surrogate safety measures (SSMs) are a means of measuring the safety of traffic infrastructure using data other than crash data. SSMs are beneficial because they eliminate the need to wait a long time for crashes to occur to generate data. In a similar vein, use of SSMs allows for hazardous areas to be identified prior to a large number of crashes occurring. This may mean that improvements can be made earlier to prevent those crashes. Yet another benefit of using SSMs is the dramatic increase in data points that comes from being able to analyze traffic conflicts instead of collisions. Conflicts are extremely common; whereas collisions are quite rare by comparison. More data points allow statistical analysis to be more effective.

One class of current research pertaining to SSMs involves the investigation of harsh braking events recorded by connected vehicles as an SSM. The validity of harsh braking as an SSM will need to be investigated in order to determine if its use is indeed viable. Such validation may be done through the comparison of results of a safety analysis conducted with the harsh braking SSM to historical crash counts or to the results of a safety analysis conducted using existing SSMs. The articles and reports included within this paper offer insight into how SSMs have evolved over time and how new SSMs may be validated. Similarly, they illuminate the various mechanisms used to compute the safety of road infrastructure as alternatives to crash data.

The following sections provide summaries of pertinent content within a selection of articles and reports published on SSMs in transportation engineering. The literature reviewed includes a paper that uses SSMs in conjunction with connected vehicle data, papers that define SSMs founded in both kinematics and statistics, papers on the use of SSMs with traffic microsimulation software, and even two papers that establish new SSMs and validate them through the use of microsimulation. This information is presented together in one paper to be a resource for researchers and practitioners alike to understand the current state of this research and the means by which these methods have been established.

### 2.3 Paper #1: "Analysis of Traffic Conflicts and Collisions"

The first paper considered was written in 1978 by Brian L. Allen, B. Tom Shin, and Peter J. Cooper of McMaster University in Canada. These researchers set out to improve upon the previously established traffic conflicts technique (TCT). They outline a number of flaws with TCT and propose several SSMs that would ameliorate the

shortcomings of TCT as it existed previously. Their newly established SSMs include proportion of stopping distance, gap time, encroachment time, deceleration rate, post-encroachment time, and initially attempted post-encroachment time.

The paper begins with the discussion of the existing TCT and the areas in which it lacks effectiveness. Prior to this paper, TCT used brake applications as the primary indicator of a traffic conflict. While brake applications are easily identified and counted without subjectivity in the data analysis process, they have several drawbacks. Drivers have variable braking habits with overly cautious drivers applying brakes when it is not necessary and less cautious drivers failing to apply brakes during hazardous encounters. Brake applications are also a binary measurement with no indication of the severity of the evasive maneuver. Finally, deceleration is not always an effective evasive maneuver. Sometimes acceleration is the safest option in order for a vehicle to clear a potential collision location. This can lead to an inaccurate conclusion in TCT. A traffic conflict definition can be misleading if it relies on the presence of an evasive maneuver, such as brake application. This is because collisions can occur without any evasive maneuver taking place. This means that a traffic conflict definition requiring evasive action can lead to collisions that are not preceded by a conflict. This is problematic because collisions are supposed to be a subset of conflicts. The authors present the time-space diagram in Figure 1 which illustrates a left-turning vehicle and a through vehicle. The authors suggest that measurement of various parameters within the time-space diagram would be useful for safety analysis, and they present several measures which they developed.

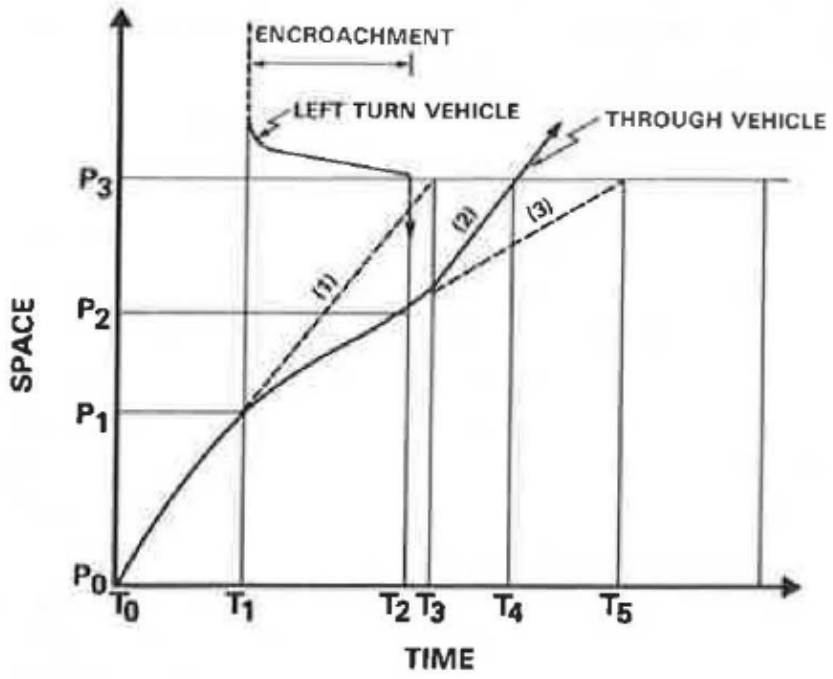


**Figure 1. Time-Space Diagram of a Left-Turn Conflict**

The paper outlines SSMs developed by the authors as alternatives to TCT based primarily on brake applications. The first measure is proportion of stopping distance (PSD), which is the ratio of the remaining distance between a vehicle and a potential collision point to the minimum acceptable stopping distance. The PSD must be at or above a value of 1.0 for a situation to be safe. It may be found with equations 2.1 and 2.2, in which *RD* is the remaining distance between a vehicle and a potential collision point, *MSD* is the minimum acceptable stopping distance, *V* is the following vehicle's velocity, and *D* is the maximum acceptable deceleration rate.

$$PSD = \frac{RD}{MSD} \tag{2.1}$$

$$MSD = \frac{V^2}{2D} \tag{2.2}$$

The next measures described are gap time (GT), encroachment time (ET), and deceleration rate (DR). GT is the difference between $T_3$ and $T_2$ in Figure 1. Time $T_3$ is the

time at which a through vehicle would have arrived at the potential collision point if the through vehicle had not altered its motion. Time $T_2$ is the time at which the left-turning vehicle is no longer encroaching on the through vehicle's path of travel. GT can therefore be positive or negative. A smaller absolute value of GT represents a greater probability of a collision occurring. ET is a measure of the total amount of time that the left-turning vehicle occupies the path of travel of the through vehicle, the difference between $T_2$ and $T_1$ in Figure 1. DR is another SSM which occurs through the development of a conflict and is capable of indicating situational severity. The authors point out that variability between drivers can account for higher or lower DRs to some extent. However, rapid deceleration is a strong indicator of a hazardous situation.

The last two developed measures are: post-encroachment time (PET) and initially attempted post-encroachment time (IAPE). PET is the amount of time that elapses between when an encroaching vehicle leaves the path of travel of another vehicle and when the other vehicle reaches the point where a collision would have occurred. PET may be quantified as the difference between $T_4$ and $T_2$ in Figure 1. PET represents how narrowly drivers avoided colliding. The measure represents the cumulative effects of the initial situation and the actions taken by the drivers to avoid colliding. PET suffers from drivers often accelerating as soon as a conflict ends. For this reason, the authors developed IAPE which eliminates the effects of early acceleration. IAPE may be calculated with equations 2.3 and 2.4, in which $T_1$ is the beginning time of encroachment, $P_1P_3$ is the distance between the potential collision point and the initial location of the through vehicle, and $V_2$ is the average through vehicle velocity.

$$IAPE = T_5 - T_2 \tag{2.3}$$

$$T_5 = T_1 + (P_1 P_3 / V_2) \quad (2.4)$$

The authors suggest that the most controversial part of their paper is the rejection of the TCT brake application method. They concede the point that their evaluation of their new SSMs is not highly effective in confirming an advantage in these measures. This is because the correlation coefficients obtained for the SSMs were low in spite of an active collision history at the study intersection. Little hope exists to have higher correlation coefficients at any intersection. Low correlation coefficients may be something to be expected, and arguments for a transition away from brake applications should rely on the conceptual weaknesses of brake applications. The fact that not all collisions are preceded by braking should alone bar brake application from being an acceptable measure.

### 2.4 Paper #2: "Extended Time-to-Collision Measures for Road Traffic Safety Assessment"

The next article, by Michiel M. Minderhoud and Piet H.L. Bovy and published in 2001 in the journal *Accident Analysis and Prevention*, outlines the development of two new modifications to a surrogate safety measure known as time-to-collision (TTC). The authors call these modifications "Extended Time-to-Collision" together, and individually these modifications are called "Time Exposed Time-to-Collision" and "Time Integrated Time-to-Collision." The paper addresses the use of these measures with vehicles that are equipped with autonomous intelligent cruise control (AICC). These new measures are intended to provide a comparative measure which may be used in conjunction with microsimulation to understand the impacts to safety of the use of AICC.

The authors describe the TTC SSM. TTC is the amount of time that would need to elapse in order for two vehicles to collide if their trajectories remain unchanged. TTC may be calculated with equation 2.5, in which *X* is the vehicle's position, *X'* is the vehicle speed, and *l* is the vehicle length. The leading vehicle is denoted as *i-1*, and the following vehicle is denoted as *i*.

$$TTC_i = \frac{X_{i-1}(t) - X_i(t) - l_{i-1}}{X_i'(t) - X_{i-1}'(t)} \quad \forall X_i'(t) > X_{i-1}'(t) \tag{2.5}$$

TTC may only be calculated for situations in which the speed differential between the vehicles is such that the leading vehicle is traveling more slowly than the following vehicle. The safety of a TTC value is tied to a critical TTC safety threshold. TTC values above this threshold are safe situations, and TTC values beneath this threshold are unsafe. Past research has produced threshold values ranging from 2.6 seconds to 4 seconds. The article presents a TTC profile which will be used to illustrate what time exposed time-to-collision and time integrated time-to-collision are measuring. This profile is shown in Figure 2.

The authors present their modifications, beginning with time exposed time-to-collision (TET). TET is a summation of the time that the TTC is beneath the safety threshold value. A low TET value indicates a safe situation because the overall exposure to a hazardous situation is small. It does not consider how severely the safety threshold is being violated. Calculation of TET requires position and speed data for all vehicles on a road section within the study time period. This data is typically collected at discrete moments, separated by a time scan interval. TET may be calculated with equation 2.6, in which $TTC^*$ is the safety threshold value of TTC, $TTC_i(t)$ is the value of TTC at a discrete time *t* for vehicle *i*, $\delta_i(t)$ is a switch variable that indicates if the threshold TTC is

exceeded, and $\tau_{sc}$ is a time scan interval indicating the time step resolution. For a N number of drivers, the population TET* may be found with equation 2.7.

$$TET_i^* = \sum_{t=0}^{T} \delta_i(t) \cdot \tau_{sc} \quad where \quad \delta_i(t) = \begin{cases} 0 & else \\ 1 & \forall\, 0 \le TTC_i(t) \le TTC^* \end{cases} \tag{2.6}$$

$$TET^* = \sum_{i=1}^{N} TET_i^* \tag{2.7}$$

Next, the paper presents the time integrated time-to-collision (TIT) SSM. The TIT measure addresses one drawback of the TET metric, its inability to consider the amount by which the safety threshold TTC is not met. In this way, TIT is capable of capturing the severity of the hazard better than TET. TIT may be calculated for continuous time with equation 2.8. Analysis using continuous time is not practically possible, so equation 2.8 represents a theoretical abstraction. For discrete time, TIT may be calculated with equation 2.9.

$$TIT^* = \sum_{i=1}^{N} \int_0^T [TTC^* - TTC_i(t)]dt \quad \forall 0 \le TTC_i(t) \le TTC^* \tag{2.8}$$

$$TIT^* = \sum_{i=1}^{N} \sum_{t=0}^{T} [TTC^* - TTC_i(t)] \cdot \tau_{sc} \quad \forall\, 0 \le TTC_i(t) \le TTC^* \tag{2.9}$$

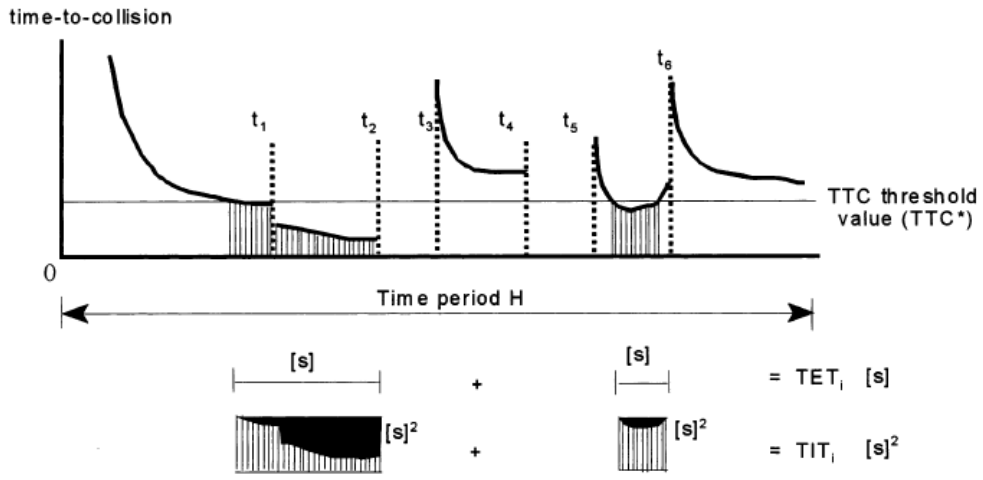


**Figure 2. Illustrations of TET and TIT**

The researchers use the comparative power of TET to demonstrate the impact of incorporating various levels of AICC. AICC is capable of adapting vehicle speed to keep proper distance from leading vehicles. To analyze the effectiveness of AICC, the researchers used an applied microscopic simulation with an individual driving behavioral model. They ran models for 50% partial AICC, 100% partial AICC, 50% complete AICC, and 100% complete AICC. Their analysis was also done for 1 second, 2 second, and 3 second safety threshold TTC values. They suggest that a shorter safety threshold TTC is possible for AICC because of its increased reaction ability over humans. Partial AICC requires driver intervention at speeds below 30 km/h or when the necessary deceleration is at or above 3 $m/s^2$. Complete AICC supports the driver fully. Again, there was high exposure to high TTC values and low exposure to small TTC values. Choice of threshold TTC value has a large effect on the total exposure time. Choice of a realistic threshold depends upon the design of the AICC system and will need to wait until AICC systems are more established and empirical data is available. This may eventually be accomplished in future research.

## 2.5 Paper #3: "Surrogate Safety Measures from Traffic Simulation Models – Final Report"

This report, written in 2003, is a summary of a project by the Federal Highway Administration which sought to evaluate the efficacy of the use of simulation software in conjunction with SSMs to determine the safety of intersections. It also identifies algorithms for determining SSMs from simulation models, known as the Surrogate Safety Assessment Methodology. This methodology allows for evaluation of various alternatives and is applicable to both signalized and unsignalized intersections.

The authors present descriptions of the following SSMs that are a part of the traffic conflicts technique: GT, ET, DR, PSD, PET, IAPE, and TTC. Field measurement is possible for these measures, but it introduces subjectivity which can compromise the quality of the safety analysis. Microsimulation can be used to simulate conflicts more precisely. There are other SSMs that have also been suggested. These measures include "delay, travel time, approach speed, percent stops, queue length, stop-bar encroachments, red-light violations, percent left turns, spot speed, speed distribution, and deceleration distribution" (Gettman & Head, p. 10). Although these measures have not been quantitatively related to crash frequency, they may be used as indicators of higher or lower crash frequency. These informal measures exist for two-lane roads as well and include design features such as curvature and superelevation.

The report gives an overview of traffic simulation models. Microsimulations analyze traffic at the level of the individual vehicle over time steps. Vehicles in the simulation have varying characteristics, but they always drive safely and never crash. The authors favor microsimulations that are commonly used in industry and have analyses

that are simple to implement. They also prefer the simulation to have a graphic network editor and analysis tools that may be used after processing. The analysis must model driver behaviors such as car following, lane changing, and gap acceptance and should have particularly realistic behavioral components to be useful. Most microsimulation programs do not readily allow extraction of data to output files, but this would be necessary for computing SSMs. The behavior and driver performance parameters need to be able to be manipulated, and the ability for a user to make or request modifications to the software itself at a relatively low cost is preferable. With these preferences established, the authors evaluate nine microsimulation programs: CORSIM, VISSIM, Simtraffic, Paramics, HUTSIM, Texas, WATSIM, Integration, and AIMSUN.

The evaluation of the various microsimulation software programs did not identify any clear best choice and revealed that using any microsimulation program for the computation of SSMs would require at least some modification of the program. Because of this, the authors recommend using a surrogate safety assessment module (SSAM) after the simulation is run. The workflow for conducting a safety analysis would involve running a simulation model, importing event files to the SSAM from the simulation, and then running the SSAM to generate reports and graphics detailing the computed SSMs. The authors go on to outline algorithms which allow SSMs to be computed for conflict events. Conflict events that may be modeled include crossing flows, merging crossing flows, adjacent flows (lane changing), and following flows (rear-end collisions). Some conflict events that are not modeled are sideswipe, head-on, and swerve-out-of-lane collisions as well as U-turn related and pedestrian collisions. The authors call for future research to improve the modeling of pedestrian collisions.

The report concludes with a discussion of validating SSMs as computed from microsimulations. One method of validation is determining if an SSM analysis with microsimulation may be used to decide between two different intersection design alternatives. The next way is to determine a correlation between SSMs and traditionally gathered crash data. The goal here is to determine if an SSM analysis with microsimulation may be used to replace traditional data gathering procedures. The third way suggested by the authors to validate SSMs is to determine if it is possible to predict the benefits to safety caused by the implementation of safety-oriented intersection improvements. The report outlines methods for validating SSMs with microsimulation in these ways.

### 2.6 Paper #4: "Comparing Safety Performance Measures Obtained from Video Capture Data"

This paper, written in 2010 by Giuseppe Guido, Frank Saccomanno, Vittorio Astarita, Demetrio Festa, and Alessandro Vitale and published in the *Journal of Transportation Engineering*, details a study in which SSMs were calculated for a roundabout in an urban area of Cosenza, Italy. The SSMs used in this study include TTC, TIT, deceleration rate to avoid collision (DRAC), PSD, and crash potential index (CPI). The different outcomes of the safety analysis according to the particular safety measure used, traffic conditions, and roundabout geometry variations are discussed with the purpose of demonstrating the usefulness of SSMs and highlighting the impact of using different measures on the outcome of safety analysis.

Next, the authors discuss the SSMs that they use in this study. The first SSM described is DRAC. DRAC is based on the idea that a leading vehicle will execute some

initial action such as braking, changing lanes, or accepting a gap. The following vehicle, in turn, decelerates in order to avoid a rear-end collision. The authors use a DRAC safety threshold of 3.35 m/s$^2$. DRAC is an effective safety measure because it considers the effects of differential speeds and evasive action in the form of braking. It may be calculated for rear end collisions using equation 2.10, in which $t$ is the time interval, $X$ is the position of the vehicle, $L$ is the vehicle length, and $V$ is the velocity. The subscript $FV$ refers to the following vehicle, and the subscript $LV$ refers to the leading vehicle.

$$DRAC_{FV,t+1}^{REAR} = \frac{(V_{FV,t}-V_{LV,t})^2}{(X_{LV,t}-X_{FV,t})-L_{LV,t}} \quad (2.10)$$

The authors next discuss TTC and PSD as defined previously. We have seen the definitions of TTC and PSD before. The safety threshold for TTC was set at 1.5 seconds in this paper. The paper goes on to define and discuss TIT, which we have seen in the discussion of Paper #2.

The final SSM discussed in the paper is CPI. CPI was developed in response to the identification of concerns with the original DRAC measure. DRAC has the drawback of not considering the variability of vehicle braking capacity based on mechanical variations in vehicles or environmental factors. To address these variations, the CPI was developed, which takes braking capacity variations into consideration. The DRAC and the maximum available deceleration rate, MADR, are calculated at every time step considered. CPI may be calculated using equation 2.11, in which $\Delta t$ is the observation time interval, $b$ is a state variable which equals 1 if the gap between the leading and following vehicles is closing and 0 otherwise, $T_i$ is the total observed time for vehicle $i$, $ti_i$ is the initial time interval observed, and $tf_i$ is the final time interval observed.

$$CPI_i = \frac{\sum_{t=ti_i}^{tf_i} P(DRAC_{i,t} > MADR_i) \cdot \Delta t \cdot b}{T_i} \quad (2.11)$$

The paper goes on to outline the methods employed to measure the interactions of vehicles within the study roundabout. A camera was set up on the roof of a close building and was used to record traffic operations on a weekday during off-peak hours. Off-peak hours were selected because the vehicular speeds are not reduced by congestion. Radar measurements revealed the average speed of vehicles to be 25 kph during the off-peak conditions. The Adobe Premier software program was used to process the video to obtain trajectories. In addition to the video footage, 176 virtual detectors were spaced 1 meter apart to collect individual trajectories. Following and leading vehicle trajectories were then linked, resulting in 77 pairs of vehicles. The authors verified the values they estimated for vehicle speeds by measuring speeds with laser guns and comparing the results. Laser guns were set up at six reference stations, including four stations at the roundabout entrances/exits. A statistical analysis of the speeds calculated from the video footage and the speeds measured using the laser guns revealed that no statistically significant difference exists between the two methods of measuring vehicle speed.

The paper next details the computation of SSMs from the 77 identified interactions. For CPI, two values were used for the MADR. The first definition is based on the coefficient of friction and cross grade of the pavement. The second definition is based upon a truncated normal distribution with a minimum value of 4.2 m/s$^2$ and a maximum value of 12.7 m/s$^2$. Potential conflicts are defined as interactions with a DRAC exceeding 3.35 m/s$^2$, a TTC lower than 1.5 seconds, a PSD less than or equal to 1, a TIT greater than zero, or a CPI greater than zero. The authors used a standardized U-statistic

to compare the safety measures. This statistic was calculated using equation 2.12, in which $x$ is the observed exposure time to a conflict value, $x_{min}$ is the minimum observed exposure time to a conflict value, and $x_{max}$ is the maximum observed exposure time to a conflict value.

$$u = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.12)$$

The paper concludes with a discussion of the characteristics of the SSMs that were highlighted in this study. The authors found very similar results with TIT and TTC in the safety analysis. PSD resulted in a higher time exposure to hazardous situations and nebulous results as to where safety problems exist. The authors found similar results for both of the CPI measures; although the CPI using the distributed MADR led to more localized results. Relative to other measures, CPI underestimated risk possibly due to CPI's consideration of braking capacity. The measures all identified areas with significant merging activity. This exposes vehicles to more abrupt acceleration and deceleration rates as well as traffic flow turbulence. Overall, the study revealed that measures which require a larger number of inputs, such as CPI, yielded more focused results regarding locations of safety hazards. These more focused results may potentially be of greater use to decision-makers in determining which safety improvements ought to be prioritized.

### 2.7 Paper #5: "Comparing Simulated Road Safety Performance to Observed Crash Frequency at Signalized Intersections"

This 2011 paper was written by Janailson Souza, Marcos Sasaki, and Flávio Cunto, Ph.D. and was submitted to the International Conference on Road Safety and

Simulation. The paper details a study in which the researchers conducted one of the validation efforts suggested in the FHWA report, namely validation by correlating SSMs and traditionally gathered crash data. The study considered intersections in Fortaleza, Brazil, and the comparison of simulation results with real-world data was done for both peak and off-peak two-hour periods. The SSMs evaluated in this paper include TTC, DRAC, and CPI. The number of rear-end collisions was observed to decrease over a period of approximately three years (2007, 2008, and 2009), but the SSMs as computed with microsimulation programs did not reflect this decrease.

The authors state that SSMs fall into three categories: time-based measures, measures of required braking power, and safety indices. All of these categories serve to provide a proactive approach to safety analysis. Another benefit of SSMs over crash data is the significantly greater frequency of high-risk situations in comparison to crashes which means that statistical methods are more reliable. The authors point out a limitation in time-based measures in that multiple scenarios may result in the same value. For instance, a low speed at a close following distance may have the same TTC as a high speed at a longer following distance. This makes it difficult to use time-based measures effectively to determine crash severity. For this reason, measures of required braking power and safety indices can be more useful. The authors' study considers one measure from each category (TTC, DRAC, and CPI). The researchers used the geometric and traffic characteristics of three intersections to build six scenarios in PTV VISSIM: peak and off-peak models for each of the three intersections.

The results of the simulation include numbers of conflicts over a three-year period and the number of conflicts per vehicle over a three-year period. The results are for three

years so that they may be compared to the crash data from 2007, 2008, and 2009. The crash data exhibits a downward trend over the three-year timespan which is not predicted by the SSMs. The authors suggest that the simulation environment's simplicity and the rareness of collisions may account for this discrepancy. The TTC and DRAC measures resulted in a much higher number of conflicts than CPI did. CPI also exhibited the highest variability which is due to the inclusion of two stochastic components: random seed generation and a distribution of maximum available deceleration rates. Crashes and conflicts increase with increased traffic volume, and the three-approach intersection in the study had significantly fewer conflicts and collisions than the other two intersections with four approaches. This supports the idea that increased exposure increases conflict and crash numbers.

The paper concludes with ideas regarding the correlation of the SSMs with actual crash data. In spite of the microsimulation not capturing the downward trend in collisions, the SSMs did find the differences in the numbers of crashes at each of the three intersections considered. This suggests that microsimulations may be used for proactive safety analysis. The authors suggest further research in incorporating parking maneuvers in safety analyses and including more types of vehicles, such as motorcycles which were excluded from this study. Another potential research area is the use of safety performance models to improve crash estimates.

### 2.8 Paper #6: "Use of Crash Surrogates and Exceedance Statistics to Estimate Road Safety"

This 2012 article was written by Andrew P. Tarko at Purdue University and published by the journal *Accident Analysis and Prevention*. The article presents a new

type of safety model which is a combination of multiple previous safety models and expands the narrow abilities of existing models. Tarko writes that the narrow abilities of prior models are due to the use of poor-quality data to estimate complicated safety factors. Data quality has improved because of better sensing techniques and technology and naturalistic driving data collection. The new model presented in this article improves upon past techniques by including crash precursor events into an estimation method that makes use of the Generalized Pareto distribution.

The paper begins with an overview of past methods for determining what events should be classified as traffic conflicts. A pyramid may be used as a representation of the frequency of traffic interactions based on their riskiness level. The pyramid is broken into sections representing, in order of increasing riskiness level: undisturbed passages, potential traffic conflicts, light traffic conflicts, serious traffic conflicts, and collisions. Less risky interactions comprise larger portions of the total volume of the pyramid than riskier interactions, indicating the higher frequency of less risky interactions. This representation is illustrated in Figure 3.

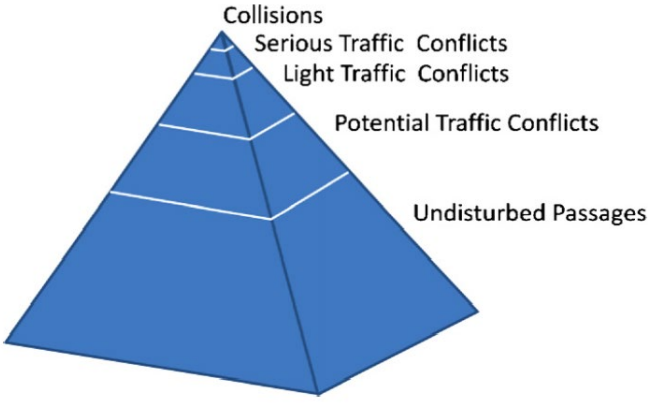


**Figure 3. Pyramid Representation of Traffic Interaction Frequency**

The author then proposes an approach to developing a better model for traffic interactions. The article defines *n* number of traffic interaction classes. The assumptions for this model are that the interaction severity is continuous, an event belongs to a particular interaction class if its severity is above a particular threshold, and the distribution of the severity of events has a right tail that converges to zero. The collision proximity, in turn, may be determined by finding the difference between the event severity and the collision severity threshold.

Tarko explains that this model is an exceedance distribution which may be used in conjunction with the Extreme Value Theory. An equivalent form of the generalized extreme values distribution is the Generalized Pareto distribution which is applicable to values in exceedance of a large fixed threshold. This distribution, and the Extreme Value Theory in general, has been used in areas concerning safety analysis, such as natural disasters, financial losses, and engineering failures. According to Tarko, the generalized extreme values and Generalized Pareto distributions may be used to estimate how frequently a car will depart from a roadway. The riskiness of events is broken into the following categories: all events, risky events, and actual road departures which may or may not be crashes. A fourth category, representing crashes following road departure, may be incorporated into a complete safety model.

Tarko defines some terms including *threshold*, *risky event range*, and *event severity*. The threshold of a risky event is the lateral clearance below which a driver would feel uncomfortable. The risky event range is the longitudinal distance over which a vehicle is too close to the edge of the road and signifies the distance required for a driver to become uncomfortable and make a corrective motion. Event severity is the proximity

of a risky event to an actual road departure and is useful for fitting the Generalized Pareto distribution.

The article outlines experiments conducted with a driving simulator and four test drivers. The track in the simulator featured many horizontal curves as well as accurate signing, billboards, a realistic rural background landscape, and traffic traveling in the same direction as the driver. The test subjects drove 2,052 miles, departing the road four times and experiencing 2,500 risky events. Using the bootstrap method, Tarko found ninety percent confidence intervals for road departures based on the number of risky events. The actual number of road departures was not used in the determination of these confidence intervals but did fall within the interval, which gives credence to the intervals and the methods used to find them. Tarko developed models for the probabilities of risky events, crashes, and crash severity as well as a model that computes the frequency of collisions of varying severity levels. Tarko's models are SSMs, taking data other than crash data and producing crash count and risky event count estimates. The consideration of the breakdown of collision severity is a valuable contribution to the literature on SSMs. Tarko calls for subsequent research into application of Pareto models to suitable data. Pareto models could potentially be used in conjunction with connected vehicle data or data from microsimulation software to determine the expected crash frequency along roadways.

### 2.9 Paper #7: "Surrogate Safety Measure for Simulation-Based Conflict Study"

This paper by Chen Wang and Nikiforos Stamatiadis was published in the journal *Transportation Research Record* in 2013. It outlines the development of an SSM called the aggregated crash propensity metric (ACPM), which may be used with

microsimulation software programs to evaluate intersection safety. The authors also describe a probabilistic model which was developed to incorporate the distributions of driver reaction times and deceleration rates during braking. This serves to compute the probability of crashes which fit into three categories: rear-end, crossing, and lane change. The measure was validated using VISSIM models which found that the ACPM performs better than the Highway Safety Manual methods in determining the relative safety of intersection designs. Attempts to correlate ACPM with real crash data was in its early stages at the time this paper was published, but the early findings suggest the potential for ACPM to be used to predict actual crash numbers.

The article begins with a discussion of SSMs and the apparent need for a new metric which more fully uses the detailed data produced by microsimulations. According to this paper, SSMs have not grown in complexity sufficiently with advancements in microsimulation. The SSAM, for instance, uses TTC with an arbitrary threshold of 1.5 seconds to measure safety. The authors intend to bridge the gap with the ACPM.

The ACPM measures the probability for each conflict at an intersection to result in a collision while considering human and vehicular variations. For every conflict, there exists a portion of the driver population that has a reaction time longer than the TTC, and there exists a portion of vehicles that have a maximum available braking rate that is lower than the required braking rate. This is illustrated in Figures 4 and 5. The reaction time distribution is a lognormal distribution with parameters that depend on the type of collision (crossing, lane-change, and rear-end). The maximum available braking rate distribution is a truncated normal distribution with a mean of 9.7 $m/s^2$, a standard

deviation of 1.3 m/s$^2$, a lower limit of 4.2 m/s$^2$, and an upper limit of 12.7 m/s$^2$, as determined in prior research.

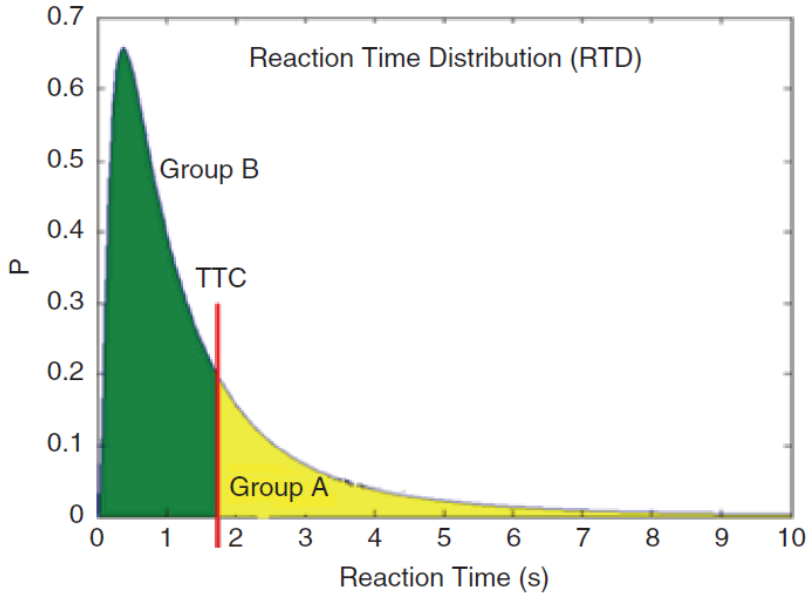


**Figure 4. Reaction Time Distribution with TTC**

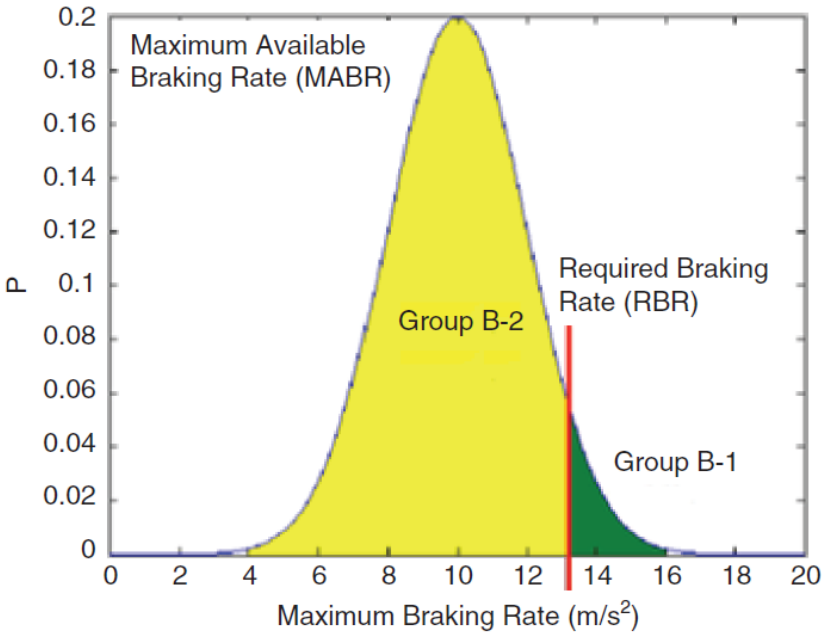


**Figure 5. Maximum Available Braking Rate Distribution with Required Braking Rate**

The groups created by the graphs shown in Figures 4 and 5 are used to determine the crash propensity metric and ultimately the ACPM. Drivers in group A do not react in time to avoid a collision. Drivers in group B-2 react quickly enough to initiate an evasive maneuver but are unable to perform the evasive maneuver successfully due to vehicular limitations. The sum of these groups (A and B-2) are all the conflicts that will result in a collision. The probability of a collision for an individual conflict is the crash propensity metric, and the sum of all propensity metrics for conflicts within a particular category is the ACPM.

The required braking rate is derived for each of the three types of collisions using kinematics. The researchers derived equations in which $l_i$ and $w_i$ are the length and width of vehicle $i$, $V_i$ is the velocity of vehicle $i$, $D$ is the distance between conflicting vehicles, $\theta$ is the conflict angle, and $x$ is the reaction time. Equations 2.13 and 2.14 are for crossing conflicts. Equation 2.13 calculates the total time $t$ during which the leading vehicle is at the conflict point. Equation 2.14 finds the required braking rate for a crossing conflict and uses the output of equation 2.13. For rear-end conflicts, equation 2.15 may be used find the required braking rate, and, for lane change conflicts, equation 2.16 may be used to find the required braking rate.

$$t = \frac{l_1 + \frac{w_1}{\tan\theta} + \frac{w_2}{\sin\theta}}{V_1} \quad (2.13)$$

$$\mathrm{RBR(crossing)} = \frac{V_2 * t}{\left(\mathrm{TTC} + \frac{t}{2} - x\right)^2} \quad (2.14)$$

$$\text{RBR(rear end)} = \frac{(V_2 - V_1)}{2*(\text{TTC} - x)} \quad (2.15)$$

$$\text{RBR(lane change)} = \frac{\frac{2V_2 l_1}{V_1} + l_2 - l_1 * \cos\theta - \frac{w_1}{\sin\theta} - \frac{w_2}{\tan\theta}}{\left(\text{TTC} + \left(\frac{l_1}{V_1}\right) - x\right)^2} \quad (2.16)$$

The crash propensity metric illuminates the differences between two scenarios which may appear identical when looking only at the TTC. In two scenarios with an identical TTC, the required braking rates may be quite different. This makes one scenario more likely to result in a collision, and the crash propensity metric will indicate just how much more likely it is.

The researchers validated the ACPM using experimentation with VISSIM models of twelve intersections on three arterials in Kentucky. The ACPM was computed for each of the three collision types at all of the intersections. The total ACPM is the sum of the three collision type-specific ACPM values. The authors ranked the intersections according to their relative safety according to the ACPM and then predicted the annual numbers of crashes at each of the intersections using the methods presented in the Highway Safety Manual. Spearman rank tests showed high rank correlation coefficients, indicating that the ACPM is a good indicator of relative intersection safety. The researchers also used the leave-one-out cross-validation method to test the ability of the ACPM to predict crash numbers at each of the intersections. The actual crash numbers fell within the 95% confidence interval of the crash predictions most of the time, indicating that ACPM is promising for use as a crash predictor.

The authors conclude the paper by reiterating that the ACPM is an SSM to be used for determining the relative safety of transportation infrastructure. The metric successfully determines the probability of crashes using TTC without an arbitrary cutoff value, a weakness of previous use of TTC. The authors point out that VISSIM's simulation operates based on the assumption that all drivers will follow the rules regarding right-of-way. Of course, this is not always the case and may lead to crossing conflicts being underrepresented. Practitioners may benefit from using another method to characterize crossing conflicts.

### 2.10 Paper #8: "Identifying High Crash Risk Roadways through Jerk-Cluster Analysis"

This paper is a thesis written by Seyedeh Maryam Mousavi and submitted to the Louisiana State University in 2015 as part of the requirements for a master's degree. It details a study in which the author uses naturalistic driving data from GPS sensors to identify locations in which high concentrations of abnormal driving events occur and correlate crash rates to these abnormal events. These events involve sudden and unusual movements of vehicles that may be detected through a measurement of the vehicle's first derivative of acceleration, known as jerk. The author mentions the importance of this work as a means of computing estimates for crash occurrence without crashes actually having to occur to produce data. This is in contrast to the standard methods of safety analysis which are retroactive in nature, relying on long-term historic crash data to identify locations that are less safe than others for the purposes of prioritizing improvements.

The author explains the methodology conducted in this research. Data collection was done through the use of GPS to generate naturalistic driving data. GPS units were placed in 31 study participants' vehicles. The GPS data was filtered to remove erroneous data points. These errors include noise, wandering, and gaps in the GPS data. Noise was the most prevalent error and includes clusters of points around intersections where vehicles are moving slowly. Wandering occurred when GPS points appear in locations where no road exists and were seemingly random. Gaps were places along roadways where data points were missing due to loss of signal between the GPS units and satellites. These errors were removed with the use of the Savitzky-Golay filter.

The next step in the methodology was differentiating the vehicles' velocity values twice to obtain jerk values. Because data was collected at discrete time intervals, jerks were computed for each interval. Because the research in this thesis intended to conduct a microscale analysis, the roadways were segmented to obtain smaller study areas. Three different scales were tested: eighth-mile, quarter-mile, and half-mile segments. These segment lengths played a role in the calculation of road segment crash rates for each of the segments. This rate, expressed for 100 million vehicle-miles, was calculated using equation 2.17 from the US Department of Transportation, in which $C$ is the number of crashes on a segment, $V$ is the average daily traffic (ADT) on the segment, $N$ is the number of years of crash data, and $L$ is the road segment length.

$$R = \frac{C \times 100{,}000{,}000}{V \times 365 \times N \times L} \qquad (2.17)$$

Input values for this equation were obtained to calculate the segment crash rates. Crash counts were obtained for a 5-year period between the beginning of 2009 and the

end of 2013. There were 1,352 crashes on LA 1248 and 1,188 crashes on LA 42. The segment length varied between eighth-mile, quarter-mile, and half-mile segments depending upon the scale being tested. The ADT for each segment was computed by using data from the Louisiana Department of Transportation and the Inverse Distance Weighted interpolation tool within GIS software. With these input values, the crash rates could be calculated.

The thesis discusses a sensitivity analysis that was done to determine the proper jerk value to use as a threshold between normal and abnormal events. Because there is no clear threshold value to use for a continuous variable like jerk, a data-driven sensitivity analysis determined the best threshold value to use from a selection of test values. Threshold values tested began at -0.5 $ft/s^3$ and decreased in increments of 0.5 $ft/s^3$ until a final test threshold value of -10.5 $ft/s^3$ was reached. A count of the number of abnormal events was then obtained and normalized based upon the total number of data points to obtain a jerk ratio for each of the segments. Again, three segment lengths were considered for both of the roadways included in the study. Pearson's correlation coefficients were computed, and this analysis revealed that a jerk threshold of -2.5 $ft/s^3$ and a segment length of one-quarter mile are most highly correlated with crash counts. A graph of the correlation coefficients is presented in Figure 6.

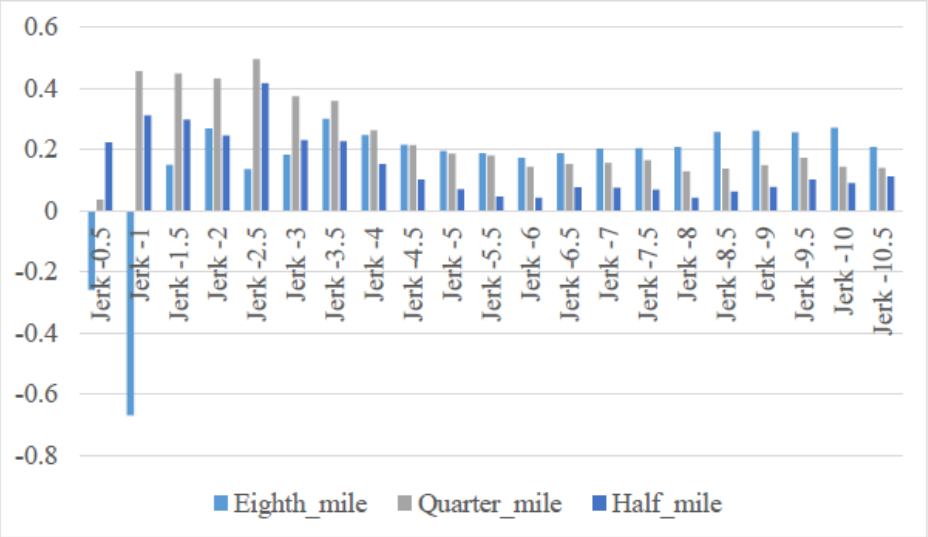


**Figure 6. Pearson's Correlation Coefficients for Various Jerk Threshold Values**

The next part of the analysis is crash frequency modeling. Two crash frequency models were created for each of the roads studied, thus resulting in four total models. The first type of model created includes only the jerk ratio as an independent variable. The second type of model includes both the jerk ratio and the presence of horizontal curvature as explanatory variables. Negative binomial regression was used to create all four models. Crash frequency modeling found that the jerk ratio was highly significant and possesses a positive correlation with crash rate. In contrast, the presence of curvature was only significant for one of the roads (LA 42) at a 95% level of significance. Therefore, presence of curvature may not be established as a meaningful predictor of crash occurrence. The value of the coefficient for the presence of curvature variable was computed to be negative, indicating that the presence of curvature tends to decrease the number of crashes that occur. This suggests that drivers adjust their behavior to drive more cautiously when curves are present, leading to fewer crashes.

The thesis concludes with a discussion of the limitations of the methodology and ideas for future research. The author states that the GPS data was of low quality and low

frequency. To capture braking information requires a high sampling rate. This problem may potentially be solved with the use of connected vehicle data. Additionally, the ADT values were interpolated using the Inverse Distance Weighted interpolation tool. This is a powerful tool, but it is possible that the interpolated values for ADT were not accurate. Having actual ADT counts would lead to a more accurate analysis. The author calls for further research into the ideal segment length for jerk-cluster safety analysis with the use of a spatial analysis tool. The author also suggests that detailed curve information, such as sharpness and radius, be included as explanatory variables in future safety models.

### 2.11 Paper #9: "Assessing Surrogate Safety Measures using a Safety Pilot Model Deployment Dataset"

This 2018 article was written by Zhaoxiang He, Xiao Qin, Pan Liu, and Md Abu Sayed and published in the journal *Transportation Research Record*. The article details a study in which SSMs were used in conjunction with data collected by the Safety Pilot Model Deployment (SPMD) program in Ann Arbor, Michigan. The SPMD program used connected vehicles and thirty items of roadside equipment to collect a variety of types of data on the vehicles involved in the program. The authors of this article used the kinematic data to evaluate the risk of mid-block rear-end crashes using SSMs.

The authors used three different measures: TTC, modified TTC (MTTC), and DRAC. The difference between TTC and MTTC is the inclusion of acceleration in MTTC. TTC is based on the assumption of a constant vehicle speed, but MTTC allows for acceleration or deceleration to be considered. These measures were used as a safety index to determine the level of danger present on various links in Ann Arbor. The measures were then compared to actual crash data to determine the goodness of fit, using

a statistical analysis with negative binomial regression. This statistical analysis reveals that the MTTC is the best of the SSMs.

The authors include the equations they used to calculate the SSMs. These equations may be incorporated into other research that uses connected vehicle data. The authors also present a map with the locations of crashes indicated as points and the safety index shown along links in the roadway network. Similar maps may be generated by other researchers using various GIS software programs such as ArcGIS. This could be a valuable addition to a study that investigates harsh braking events as SSMs.

The authors end the paper with some suggestions for future research. One area in which researchers can build upon this study is in the data processing approach. The authors acknowledge that their method of data processing may not be ideal due to some of the assumptions made. They call for research into finding other effective approaches as well as incorporating additional SSMs, such as PET and the difference in vehicle velocities. They also suggest that future research take place regarding the use of signal phasing and timing (SPaT) data. This research could potentially illuminate relationships between red light running and safety.

### 2.12 Paper #10: "Surrogate Safety Measures from Traffic Simulation: Validation of Safety Indicators with Intersection Traffic Crash Data"

The final paper considered, written by Vittorio Astarita, Ciro Caliendo, Vincenzo Pasquale Giofré, and Isidoro Russo and published in 2020 in the journal *Sustainability*, covers a study which proposes and validates a new SSM. This new measure uses vehicle trajectories and the mean energy of a vehicle to determine a safety metric and is capable of considering the dangers single vehicle crashes into roadside objects. These

considerations have not been incorporated in SSMs prior to this paper. The researchers validate their new metric by comparing its results to both historical data and measures produced by other means, such as TTC and PET.

The authors begin by reviewing the published literature on SSMs and highlighting concerns with the existing measures. They describe measures such as TTC, PET, and DRAC. They also describe a new traffic microsimulation program called TRITONE which evaluates road safety and has been validated through comparison with the SSAM. The authors list four topics which cause them concern with the existing measures: human factor modeling, traffic simulation packages, traffic safety indicators, and friction and shear forces in traffic flows.

The article goes on to describe the researchers' reasoning behind each of the concerns. In terms of human factor modeling, the prior measures did not consider human error or human distraction. These are usually caused by drivers multitasking and account for approximately 30% of crashes in the United States. In considering traffic simulation packages, the authors raise concerns about the inability of most programs to compute SSMs as well as the SSAM's inability to characterize crash severity or map locations of conflicts. In terms of traffic safety indicators, the authors point out that SSMs do not consider the outcome of a traffic conflict should it become a collision. Finally, the authors are concerned by the lack of consideration for potential conflicts between vehicles that are on trajectories that do not intersect or between vehicles and roadside objects.

With these concerns in mind, the researchers describe how their new SSM will address these lacunae. Beginning with a starting dataset for vehicle trajectories within a

network, the researchers extract both the position and speed for every single vehicle in the dataset for every second of their simulation. For each of these vehicle speeds and locations, the researchers calculated deviated trajectories that are a particular angle to the right and to the left of the vehicle's neutral trajectory along the road. The angle is generated with a Gaussian distribution. The deviated trajectories are then followed by the vehicle for a particular distraction time, which the researchers assumed to be five seconds. With these distracted paths calculated for the vehicles, potential collisions with other vehicles or roadside objects are determined, and the energy of impact in the crash is calculated using the physics of inelastic collisions.

This methodology solves the concerns of the researchers in a number of ways. First, it takes human error and distraction into account through the deviated courses. This allows for conflicts between vehicles on paths that do not overlap to be considered, such as conflicts between vehicles traveling in opposite directions along a roadway. This also allows for single vehicle crashes to be represented as long as the location, shape, and material properties of roadside objects are included in the analysis. Finally, the crash dynamics are represented in the simulation, which means that impact energy is known. The researchers ran their simulation using TRITONE for four scenarios involving nine intersections in Salerno, Italy. The results of the simulations included numbers of crashes and mean collision energy. For comparison, the researchers also computed numbers of collisions from TTC and PET with threshold values.

With these results, the article details a statistical analysis of the two methods of estimating crash counts. This analysis involves the computation of the root mean square error and likelihood ratio test statistic for each method. The researchers developed two

models, Model A and Model B. Model A uses TTC, PET, traffic flow, and a dummy variable as explanatory variables. Model B uses mean collision energy, traffic flow, and a dummy variable as explanatory variables. The statistical analysis demonstrated that both of these models are statistically equivalent and Model B is able to estimate crash counts accurately as evidenced by a comparison to five-year crash counts. The findings of this paper suggest that crash counts may be successfully estimated using trajectory deviations to calculate mean collision energy and then fitting a model with that as an explanatory variable. The authors mention that their simulations made use of many default values for parameters, so they recommend further research into ways to calibrate this methodology to a specific area.

## 2.13 Discussion

Practitioners may apply the findings of the papers considered here to conduct a variety of types of surrogate safety analysis which vary in both the measures used and the way in which data is collected for the analysis. SSMs include time-based measures, deceleration-based measures, and safety indices. Time-based measures include TTC, TET, TIT, PET, IAPE, and PSD. Deceleration-based measures include DRAC and the use of harsh braking data as indicated by the jerk values experienced by vehicles. Safety indices include CPI and ACPM. The method used by Astarita et al. involved developing a crash prediction equation that uses a combination of time-based measures and traffic flow characteristics. Although it is possible to collect data for SSMs with in-person observations or video data, the standard methods at this point in time include simulation with VISSIM or TRITONE models or use of connected vehicle data. Additional processing is necessary for both of these methods.

In order for practitioners to conduct surrogate safety analysis, they must first collect data on the intersection or link being considered. If using microsimulations, such data would include the roadway geometry and traffic characteristics. Analysis with simulation involves building models in VISSIM, TRITONE, or another suitable microsimulation program and then processing the output kinematic data with the Federal Highway Administration's SSAM. The equations presented throughout this paper may also be used with such kinematic data to compute SSMs. Connected vehicle data is currently available through vendors but may eventually be available to the public in the future. Practitioners can use the kinematic data from connected vehicles as inputs to the SSM equations included throughout this paper as was done by He et al. (2018). Simulation and connected vehicle data allow for proactive safety analysis and preemptive safety improvements.

These papers vary in their levels of usefulness at this point in time. Newer papers, of course, have an advantage over older papers due to their authors having the benefit of a greater amount of prior research. However, some older papers, such as Allen et al.'s 1978 paper, still offer useful insights and SSMs. Table 2.1 is a rubric of the usefulness of the papers considered here with different categories that may concern practitioners looking into implementing their methods.

**Table 2.1. Rubric of Usefulness for Papers Considered**

| Paper Number - Year | Theoretical Value | Practical Value | Relevance | Difficulty of Implementation |
|---|---|---|---|---|
| 1 – 1978 | Most | Most | Medium | Medium |
| 2 – 2001 | Medium | Medium | Medium | Medium |
| 3 – 2003 | Least | Most | Medium | Least Difficult |
| 4 – 2010 | Most | Least | Least | Most Difficult |
| 5 – 2011 | Least | Medium | Medium | Least Difficult |
| 6 – 2012 | Most | Least | Least | Most Difficult |
| 7 – 2013 | Most | Medium | Most | Medium |
| 8 – 2015 | Medium | Most | Most | Least Difficult |
| 9 – 2018 | Medium | Most | Most | Medium |
| 10 – 2020 | Most | Medium | Most | Most Difficult |

## 2.14 Conclusions

Traffic safety analysis with surrogate safety measures has evolved over the past several decades both in terms of the measures themselves and in the technology used to compute them. The articles and reports summarized in this paper range in publication year between 1978 and 2020, illustrating this evolution. The earliest method of computing surrogate safety measures was the use of time-lapse imagery which eventually gave way to microsimulation and, more recently, the implementation of connected vehicle data. Use of microsimulation greatly improved the precision with which surrogate safety measures may be computed but was also an abstraction. Connected vehicle data supplies both the realism of on-site measurement and the precision that is available with microsimulation, making it the preferred technology at this point in time. The measures have evolved from simply counting brake applications to taking kinematics into account or measuring rates of deceleration to determine where safety hazards exist. Surrogate safety measures are a means of preventing crashes and the damages, injuries, and loss of

life that crashes cause. Research on surrogate safety measures has made great strides, as demonstrated by the articles considered in this paper. Continuing research into making these methods more easily implemented and more effective through the use of connected vehicle data could lead to much more effective safety analysis, more targeted infrastructure improvements, and safer roads and intersections for the public.

## 2.15 Biographies

Nathaniel Edelmann is a graduate student pursuing a Master of Science degree in civil engineering at Boise State University. He earned a BS in civil engineering from Boise State University in 2020. Edelmann's thesis research concerns traffic safety analysis through the use of surrogate safety measures. Edelmann may be reached at nathanieledelmann@u.boisestate.edu

Mandar Khanal is a professor at Boise State University, having joined the Department of Civil Engineering in August, 1997. Dr. Khanal earned an MS degree from Northwestern University and a Ph.D. from the University of California, Irvine. He is a registered civil engineer in California and Idaho. Dr. Khanal was a research associate at the Louisiana Transportation Research Center at Louisiana State University and served as the chair of the Technical Review Team for the Mobile Video Surveillance & Communication Project. At Boise State University, Dr. Khanal is responsible for the Transportation Engineering program, continuously developing new courses within this field. His external grant projects have been sponsored by local, state, and federal agencies. Dr. Khanal may be reached at mkhanal@boisestate.edu

# CHAPTER 3: "USING HARSH BRAKING DATA FROM CONNECTED VEHICLES AS A SURROGATE SAFETY MEASURE"

The following manuscript was written by Nathaniel Edelmann and Mandar Khanal and submitted on July 30, 2022 to the Transportation Research Board for presentation at their January, 2023 Annual Meeting in Washington, D.C. and potential publication in the journal *Transportation Research Record*. The manuscript is titled "Using Harsh Braking Data from Connected Vehicles as a Surrogate Safety Measure." The manuscript submitted to the conference conformed to the established formatting standards of the conference; the formatting has been altered here to conform to the formatting in the rest of this document. The content presented here is identical to that which was submitted to the conference.

The study discussed in this manuscript involves the development of statistical crash prediction models which use a count of high-jerk events from connected vehicles as an explanatory variable. Jerk is the first derivative of acceleration with respect to time, meaning that it is the rate of change of acceleration. It is therefore the second time derivative of velocity and the third time derivative of position. A jerk value of 3 $ft/s^3$ indicates that a vehicle's acceleration is increasing by 3 $ft/s^2$ during every second that elapses. For example, a vehicle which begins with zero acceleration but experiences a jerk of 3 $ft/s^3$ will experience an acceleration of 3 $ft/s^2$ after one second, 6 $ft/s^2$ after two seconds, 9 $ft/s^2$ after three seconds, and so on.

The statistical models used in this study include Poisson, Negative Binomial, and Generalized Poisson regression models. These particular models were selected because they possess the characteristic of having a discrete count as a dependent variable. This trait has led to Negative Binomial regression being commonly used in traffic safety modeling, as the discrete crash count may be used as the dependent variable in the model. Because of this, Negative Binomial regression was investigated first as a potential model. The mean and variance of the crash counts used in the modeling were found to be similar, a requirement for Poisson regression models. Because this requirement was approximately satisfied, Poisson modeling was investigated too. Generalized Poisson regression models have been used by past researchers to fit both underdispersed and overdispersed data while Negative Binomial regression models have been used to fit overdispersed data. In our case the variance of the dependent variable was slightly smaller than the mean; in other words, the data set was slightly underdispersed. Because of this reason all three model types were explored in this thesis.

### 3.1 Abstract

Surrogate safety measures are a means of safety analysis for the purpose of identifying high-risk road infrastructure. Surrogate safety measures allow for proactive safety analysis, meaning that the analysis may take place prior to crashes occurring. Safety improvements may in turn be implemented proactively to prevent crashes and the associated injuries and property damage. Existing surrogate safety measures primarily rely on data generated by microsimulations, but the advent of connected vehicles has allowed for the incorporation of data from actual cars into safety analysis with surrogate safety measures. In this study, commercially available connected vehicle data is used to

develop crash prediction models for crashes at intersections and segments in Salt Lake City, Utah. Harsh braking events are identified and counted within the influence areas of sixty study intersections and thirty segments and then used to develop crash prediction models. Other intersection characteristics are considered as regressor variables in the models. These models may be used as a surrogate safety measure to analyze intersection safety proactively. The findings are applicable to Salt Lake City, but similar research methods may be employed by researchers to determine if these models are applicable in other cities and to determine how the effectiveness of this method endures through time.

Keywords: Safety, Surrogate Safety Measure, Crash, Prediction, Connected Vehicle, Harsh Braking

## 3.2 Introduction

Surrogate safety measures (SSMs) offer benefits over traditional safety analysis methods that use historical crash data. SSMs are a type of safety analysis that make use of data other than crash data, typically vehicle kinematic data. The first benefit of SSMs is that they use data which may be collected more rapidly than historical crash data. Crashes are rare events, and historical data may require years of accumulation to conduct a safety analysis. The second benefit is that SSM analysis is proactive, allowing for safety analysis prior to crashes occurring. An unsafe location may therefore be identified and improved before crashes occur, preventing injuries and property damage and possibly saving lives. The third benefit of SSMs is that the kinematic data used in a safety analysis with SSMs is much more voluminous, allowing statistical methods to be more effective.

The kinematic data employed by SSMs may come from several sources. In the past, manual measurement at the study site was used. This method of data collection was

problematic because it allowed for subjectivity and was difficult to perform accurately due to the fleeting nature of traffic interactions. Manual observation was replaced with video recordings which made it possible for traffic interactions to be replayed and offered the chance for multiple observers to analyze interactions, thus improving the problem of subjectivity. More recently, microsimulation technology has allowed for simulation to be used as a source of kinematic data. This method eliminates subjectivity, as the computer running the simulation provides the data rather than human observers (Gettman & Head, 2003). Microsimulation produces highly detailed and precise data and can produce large volumes of data with relatively little effort in comparison with manual collection. The fault of microsimulation lies in its being an abstraction rather than reality. While microsimulations are still highly useful, there has been research into the use of connected vehicle (CV) data with SSMs, meaning the use of data from the physical world rather than simulation.

CVs are a source of traffic data that allows for the high level of precision offered by microsimulation along with the realism of being generated by human drivers. CVs are automobiles sold to the public that include a transceiver which allows data to be collected regarding the vehicle's motion. For the sake of privacy, no individually identifiable information about the vehicle is visible. Vendors offer CV data to clients who wish to use the data for research and engineering projects. One such vendor is Wejo Data Services Inc., which was the source of CV data for this study. The main drawback of using data from CVs is that they currently comprise a small percentage of the total number of vehicles in the United States. A study from October, 2021 found the median CV penetration rate to be approximately 4.5% (Hunter et al., 2021). CVs therefore do not

offer a full picture of traffic. They are gradually becoming more common, though, as older vehicles are retired and replaced with new vehicles that are connected. Research into effective analysis methods with CV data will become more valuable as time goes on, speaking to the need for this research to take place now for a future increase in CVs.

One metric that is available from CVs is harsh braking event counts which form the basis for the models developed in this study. Data points from CVs include information about braking and acceleration. The braking data may be filtered so that harsh braking events are identified and counted and then used as a regressor variable in a crash prediction model. This method is investigated in this paper. The significance of other regressor variables, such as CV volume and intersection geometric characteristics, was also investigated. The proposed crash prediction models may be used to estimate monthly counts of intersection-related crashes and offer all of the benefits of SSMs mentioned above.

### 3.3 Literature Review

Researchers have developed many SSMs which tend to fall into three categories. SSMs can be a time-based measure, a deceleration-based measure, or a safety index. Although most SSMs consider collisions involving two vehicles, it is possible to model single-vehicle crashes due to distraction or error (Astarita et al., 2020). SSMs operate upon the concept that events with greater risk tend to happen less frequently, with the riskiest and rarest events being those events that result in collision (Tarko, 2012). By analyzing less risky events that occur significantly more frequently, a safety analysis with SSMs can offer more insight into safety than analysis with crash data alone.

Time-based measures consider the kinematics of vehicles and how much of a time gap exists between vehicles. Time-to-collision (TTC), post-encroachment time (PET), and proportion of stopping distance (PSD) are time-based SSMs. TTC is a measure of the amount of time required for the space between two vehicles to close. TTC on its own is transient, but Minderhoud and Bovy developed aggregation methods in the form of their extended TTC measures: time-integrated TTC and time-exposed TTC (Minderhoud & Bovy, 2001). Post-encroachment time is the difference in time between when an encroaching vehicle exits the path of travel and when a following vehicle first occupies the location where a collision would have occurred. A modified form of PET exists as initially attempted PET (IAPE). IAPE corrects the measure to account for the acceleration that commonly occurs when a driver determines that a conflict has ended (Allen et al., 1978). PSD is a ratio between the distance a vehicle is from a potential collision location and the minimum stopping distance. These distances depend upon the velocity of the vehicles involved, making PSD a time-based measure.

There are both strengths and weaknesses associated with time-based SSMs. The strength of time-based SSMs lies in their simplicity and intuitiveness. TTC and PET may be implemented with kinematic data supplied by either on-site measurements or microsimulation. PSD also requires such kinematic data, but it also requires information on vehicles' possible deceleration rates. This deceleration rate can be an established value or distribution of values or may be derived from environmental conditions. Drivers are aware of the importance of following distance and time headway, making these measurements intuitive for researchers and practitioners alike. A weakness of time-based SSMs is the possibility of multiple encounters producing identical measures (Souza et al.,

2011). TTC may evaluate to the same solution for both an encounter with a large speed differential between vehicles and a long following distance and another encounter with a small speed differential but a short following distance. This has made it difficult to establish particularly meaningful safety thresholds for these measures. Another weakness is the inability of time-based SSMs to evaluate the severity of a potential collision. In the encounters just described which both result in an identical TTC, the severity of a resulting collision will be very different because of the differing speed differentials.

Deceleration-based measures consider braking action and the braking capacity of vehicles and are better equipped than time-based measures to evaluate potential crash severity. Additionally, this type of measure considers a driver's evasive action, an important component of traffic conflicts. Deceleration-based measures include braking applications and deceleration rate to avoid collision (DRAC). Brake applications have been found to be a poor SSM due to the variability in braking habits among drivers. Brake applications are such a common act, even in benign situations, that they are not highly indicative of a conflict (Allen et al., 1978). Brake applications as an SSM fail to consider the severity of each particular braking action, something that DRAC and harsh braking are able to capture to their benefit. DRAC is a measure of the deceleration rate that a following vehicle would need to apply to avoid colliding with a leading vehicle. That measurement is compared to a safety threshold, commonly given as 3.35 $m/s^2$, to determine whether a conflict occurred (Guido et al., 2010).

Harsh braking events have also been suggested as an indicator of a conflict, which would also fall under the category of deceleration-based measures. A 2015 study found a high level of correlation between crash counts and harsh braking events, defined as

events with a large absolute value of the first derivative of acceleration, known as jerk. These events were collected by vehicles with GPS units which collected data on the vehicles' location over time, allowing the jerk value to be computed. Mousavi found a threshold of -2.5 $ft/s^3$ to be the most effective to define harsh braking but also noted that this threshold is lower than expected. Further investigation of a proper jerk threshold was recommended (Mousavi, 2015).

Safety indices are the third category of SSM. These indices consider various factors and produce an indirect safety metric. Two examples are crash potential index (CPI) and the aggregated crash propensity metric (ACPM). CPI was developed to improve upon the drawbacks of the DRAC measure. While a constant safety threshold value is typically used with DRAC, the braking capacity of vehicles is variable for mechanical and environmental reasons. CPI considers this variability through the use of a maximum available deceleration rate (MADR) distribution. The probability that DRAC is greater than MADR is a term in the computation of CPI. ACPM also considers the MADR distribution in conjunction with a distribution of driver reaction times to compute the probability that each vehicle interaction will result in collision. These probabilities are aggregated to produce the ACPM (Wang & Stamatiadis, 2013). CPI and ACPM indicate the safety level of a study location and time period without being a single measure of some observable quality.

Of the SSMs discussed, analysis of harsh braking events holds potential because of its compatibility with CV data. Previous studies, such as Mousavi's thesis (Mousavi, 2015) and the work of Bagdadi and Várhelyi (2011) have analyzed harsh braking data from GPS units because of the lack of availability of large-scale CV data when these

studies were conducted. He et al. investigated the use of CV data for SSMs, using a safety pilot model dataset to compute TTC, DRAC, and a modified form of TTC (He et al., 2018). Their study demonstrated the effectiveness of computing these measures with kinematic data from CVs. Development of a crash prediction model that uses harsh braking data from CVs would bridge the gap between these two studies and provide another tool for safety analysis.

## 3.4 Methods

The methods undertaken in this study include three phases: selection of study intersections, data collection, and statistical modeling. CV data collection was enabled by the automobile companies that manufactured the CVs and provided that data to Wejo Data Services Inc. The data used in this study was obtained from Moonshadow Mobile, a company that works in tandem with Wejo to present the raw data from CVs in a mobility analytics platform that facilitates filtering and querying. This study uses data within Salt Lake City, Utah for the months of March 2019, January 2021, and August 2021. These months were selected because of the availability of reports from the Utah Department of Transportation (UDOT) which presented detailed crash and traffic information for intersections over the course of these specific months.

### 3.4.1 Intersection Selection

The intersection selection process involved the collection of crash counts for all major intersections in Salt Lake City, amounting to 370 intersections. Crash counts for the three study months were obtained from the UDOT database and summed to find the total number of crashes for the intersections. The crashes within the UDOT system were filtered to include only those deemed to be intersection related. The sixty intersections

with the most crashes were selected. The total crashes ranged from one to six. The sixty chosen study intersections included both signalized and unsignalized intersections.

### 3.4.2 Data Collection

CV data for the intersections was collected from Moonshadow Mobile's interface. The interface comprises an interactive map and a control pane. The map displays waypoints that are produced by the CVs. When a CV is in motion, waypoints are produced once every three seconds. The waypoints are grouped by the overall trip of which it is a part by a journey ID number, making it possible to collect CV volumes. The waypoints also include such data as geographical location, a timestamp, speed, acceleration, jerk, heading, and information about the origin and destination of the trip that includes the particular waypoint. Harsh braking events were identified using the jerk values of these waypoints. Jerk is the first derivative of acceleration and is recorded for each of the waypoints. A geospatial filter was applied to limit the waypoints to those within the influence area of the study intersections, the main intersection square and the legs of the intersection 250 ft behind the stop bar as displayed in Figure 7 (TRB, 2016). Another filter was applied to limit waypoints to only those that possess a jerk value that is above the threshold that differentiates a regular braking event from a harsh braking event. This jerk threshold varied in this study to test the effectiveness of several harsh braking definitions. Thresholds tested varied between -0.5 $ft/s^3$ and -10.5 $ft/s^3$ in increments of 0.5 $ft/s^3$. The query tool was used to obtain counts of harsh braking events for each of the jerk thresholds.

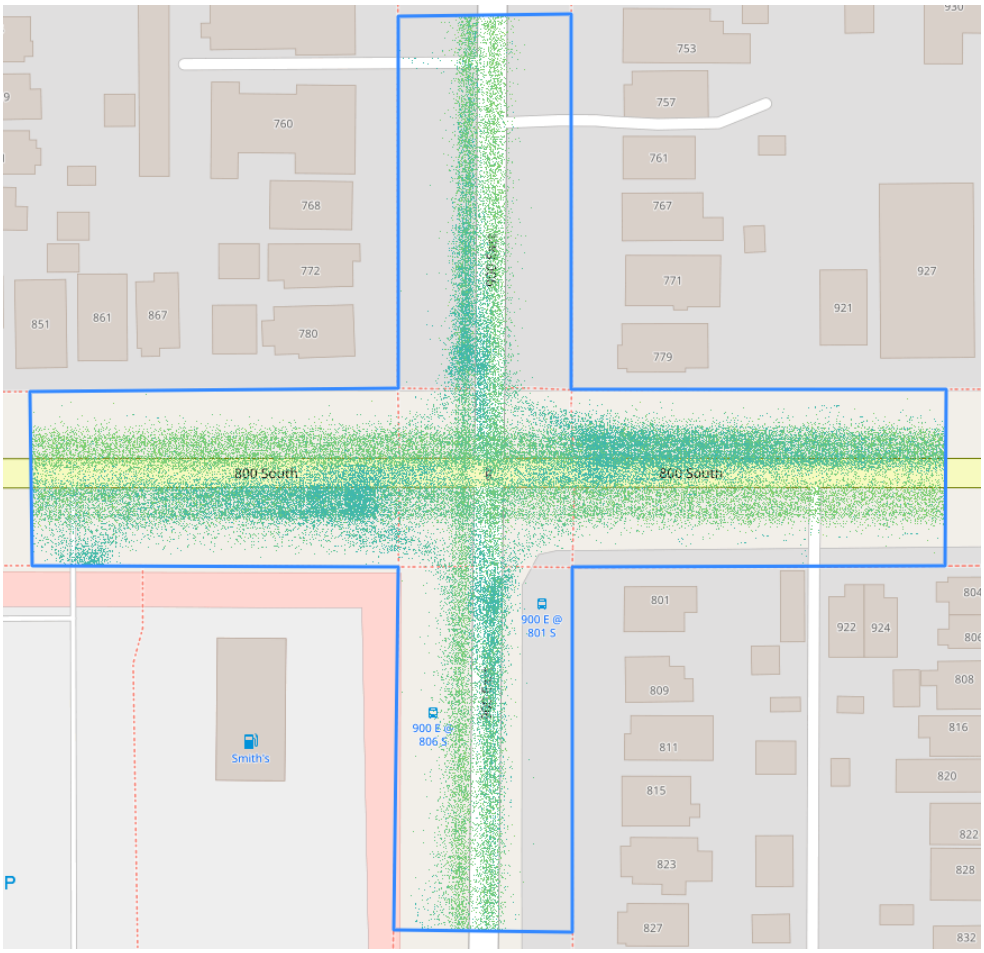


**Figure 7. Intersection Influence Area with Waypoints Displayed**

Other metrics collected from the CV data included the CV volumes and the average jerk value for each of the intersections. The CV volumes were obtained by querying the unique count of the journey ID numbers. This counts the numbers of groups of waypoints that belong to trips that pass through the intersection. Thus, the volume of vehicles passing through the intersection is obtained. The total monthly CV volume was collected as was the total monthly volume that used the intersection between the hours of 7 AM and 9 AM and between the hours of 4 PM and 6 PM. The average jerk value among all waypoints within the intersection influence area was obtained on a monthly basis for each of the three study months for each of the intersections.

In addition to the crash data and CV data, information regarding the geometry and geography of each of the intersections was collected. The number of approaches with left turn lanes, the number of approaches with right turn lanes, and the maximum number of lanes that a pedestrian would have to cross were collected using Google Earth. Historical imagery was employed to ensure that these values were correct for the study months in question. ArcGIS Pro was used to determine the number of bus stops and the number of schools within a 1,000 ft radius of the center point of each of the intersections. These metrics were included in this study because they are used in the safety performance functions within the Highway Safety Manual (AASHTO, 2010). Table 3.1 is a summary of the dependent, exposure, and regressor variables collected for analysis in this study.

**Table 3.1. Summary of Variables**

| Variable | Definition | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| Monthly Crashes | Number of intersection-related crashes within the study month | 0.7389 | 0.8347 | 0 | 6 |
| Jerk1 | Harsh braking events with the threshold being -0.5 ft/s$^3$ | 93813 | 80218 | 1074 | 310321 |
| Jerk2 | Harsh braking events with the threshold being -1.0 ft/s$^3$ | 85134 | 73590 | 660 | 285385 |
| Jerk3 | Harsh braking events with the threshold being -1.5 ft/s$^3$ | 78212 | 68040 | 450 | 263804 |
| Jerk4 | Threshold = -2.0 ft/s$^3$ | 71926 | 62794 | 342 | 242067 |
| Jerk5 | Threshold = -2.5 ft/s$^3$ | 65580 | 57440 | 282 | 221542 |
| Jerk6 | Threshold = -3.0 ft/s$^3$ | 60261 | 52913 | 234 | 204843 |
| Jerk7 | Threshold = -3.5 ft/s$^3$ | 55593 | 48948 | 204 | 189637 |
| Jerk8 | Threshold = -4.0 ft/s$^3$ | 51219 | 45112 | 189 | 174165 |
| Jerk9 | Threshold = -4.5 ft/s$^3$ | 46951 | 41349 | 177 | 159274 |
| Jerk10 | Threshold = -5.0 ft/s$^3$ | 42724 | 37538 | 171 | 148008 |
| Jerk11 | Threshold = -5.5 ft/s$^3$ | 38761 | 34031 | 156 | 136518 |
| Jerk12 | Threshold = -6.0 ft/s$^3$ | 34769 | 30507 | 144 | 124191 |
| Jerk13 | Threshold = -6.5 ft/s$^3$ | 31090 | 27190 | 135 | 112042 |
| Jerk14 | Threshold = -7.0 ft/s$^3$ | 27816 | 24255 | 117 | 100044 |
| Jerk15 | Threshold = -7.5 ft/s$^3$ | 24782 | 21517 | 105 | 88199 |
| Jerk16 | Threshold = -8.0 ft/s$^3$ | 21835 | 18918 | 87 | 77143 |
| Jerk17 | Threshold = -8.5 ft/s$^3$ | 19475 | 16857 | 78 | 67701 |
| Jerk18 | Threshold = -9.0 ft/s$^3$ | 17350 | 14978 | 69 | 59011 |
| Jerk19 | Threshold = -9.5 ft/s$^3$ | 15522 | 13436 | 54 | 53277 |
| Jerk20 | Threshold = -10.0 ft/s$^3$ | 13963 | 12054 | 45 | 49674 |
| Jerk21 | Threshold = -10.5 ft/s$^3$ | 12623 | 10943 | 42 | 46832 |
| Jerk Avg | Average jerk value among all CV waypoints within the study month | -1.437 | 1.4076 | -16.92 | -0.01 |
| Monthly CVs | Number of unique CV trips through the intersection in the study month | 9488 | 8041.9 | 187 | 29481 |
| Monthly AM CVs | Number of unique CV trips through the intersection in the study month between the hours of 7 AM and 9 AM | 947.6 | 872.85 | 9 | 3412 |

| | | | | | |
|---|---|---|---|---|---|
| Monthly PM CVs | Number of unique CV trips through the intersection in the study month between the hours of 4 PM and 6 PM | 1425 | 1204.7 | 20 | 4720 |
| Left Turn Approaches | Number of intersection approaches with a designated left-turn lane | 3.267 | 1.1987 | 0 | 4 |
| Right Turn Approaches | Number of intersection approaches with a designated right-turn lane | 1.733 | 1.3684 | 0 | 4 |
| Maximum Lanes Crossed by Ped | Maximum number of lanes a pedestrian must traverse to cross any of the intersection legs | 6.383 | 1.7378 | 2 | 9 |
| Bus Stops | Number of bus stops within a 1,000 ft radius of the intersection center point | 5.45 | 3.5709 | 0 | 13 |
| Schools | Number of schools within a 1,000 ft radius of the intersection center point | 0.2667 | 0.5135 | 0 | 2 |

### 3.4.3 Statistical Analysis

Once these data points were collected for each of the study intersections during each of the study months, a statistical regression analysis was performed to produce crash prediction models for Salt Lake City. Poisson regression, negative binomial regression, and generalized Poisson regression were considered in the analysis. Poisson regression requires that the mean and variance are equal for the dependent variable in the regression. The mean and variance of the monthly crashes at the intersections were approximately equal, making Poisson regression a viable option.

#### 3.4.3.1 Poisson Regression

Poisson regression is applicable when the variable of interest is assumed to follow the Poisson distribution, which is a model of the probability that a particular number of

events will occur. The dependent variable is the event count, which can be any of the nonnegative integers. Large counts are assumed to be uncommon, making Poisson regression similar to logistic regression, with a discrete response variable. Poisson regression, unlike logistic regression, does not limit the response variable to specific values. The Poisson distribution model takes the form given in Equation 3.1, in which $Y$ is the dependent variable, $y$ is a count from among the nonnegative integers, and $\mu$ is the mean incidence rate for an event per unit of exposure.

$$Pr(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!} \quad (y = 0,1,2, \dots) \tag{3.1}$$

If the Poisson incidence rate, $\mu$, is assumed to be determined by a set of regressor variables, then Poisson regression is possible through the expression displayed in Equation 3.2 and the regression model displayed in Equation 3.3. In these equations, $X$ is a regressor variable, $\beta$ is a regression coefficient, and $t$ is the exposure variable.

$$\mu = exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{3.2}$$

$$Pr(Y_i = y_i|\mu_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!} \tag{3.3}$$

The regression coefficients in Equation 3.2 may be estimated by maximizing the log-likelihood for the regression model. This is done by setting the derivative of the log-likelihood equal to zero to generate a system on nonlinear equations which may be solved with an iterative algorithm. The reweighted least squares iterative method is typically able to converge to a solution within six iterations (NCSS, 2016b).

3.4.3.2 Negative Binomial Regression

The negative binomial distribution is a generalization of the Poisson distribution that includes a gamma noise variable. This allows negative binomial regression to be

performed even if the dependent variable's mean and variance are not equal (NCSS, 2016a). Negative binomial regression is commonly used for traffic safety applications because it has loosened restrictions in comparison to Poisson regression but is still capable of estimating an observed count, such as crash counts (Wang et al., 2017). The negative binomial distribution takes the form presented in Equation 3.4, in which $\alpha$ is the reciprocal of the scale parameter of the gamma noise variable and other variables are as defined previously.

$$Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i} \quad (3.4)$$

The mean of $y$ in negative binomial regression depends upon the exposure variable and the regressor variables which are related by the expression displayed in Equation 3.5. Negative binomial regression is possible with the regression model displayed in Equation 3.6. In these equations, $x$ is a regressor variable, and the other variables are as defined previously. As with Poisson regression, maximizing the log-likelihood may be used to estimate the regressor coefficients through an iterative algorithm (NCSS, 2016a).

$$\mu_i = exp(ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \quad (3.5)$$

$$Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \quad (3.6)$$

3.4.3.3 Generalized Poisson Regression

Generalized Poisson regression, like negative binomial regression, is applicable in a broader set of circumstances than Poisson regression. This is because it does not have the requirement that the mean and variance of the dependent variable in the regression be equal. There are two types of generalized Poisson regression models: Consul's

generalized Poisson model and Famoye's restricted generalized Poisson regression model. Consul's model, also known as the Generalized Poisson-1 (GP-1) model, is the regression model that was employed in this study. The GP-1 model operates on the assumption that the dependent variable, *y*, is a random variable following the probability distribution presented in Equation 3.7, in which $\lambda$ is the number of events per unit of time and $\alpha$ is the dispersion parameter which can be estimated using Equation 3.8 (Hilbe, 2011). In Equation 3.8, *N* is the number of samples, *k* is the number of regression variables, $y_i$ is the $i^{th}$ observed value, and $\hat{y}_l$ is the Poisson rate $\lambda_i$ predicted for the $i^{th}$ sample (Date, n.d.).

$$Pr(Y = y_i) = \frac{e^{-(\lambda+\alpha*y_i)} * (\lambda + \alpha * y_i)^{y_i-1} * \lambda}{y_i!} \tag{3.7}$$

$$\alpha = \frac{\sum_{i=1}^{N} \left( \frac{|y_i - \hat{y}_l|}{\sqrt{\hat{y}_l}} - 1 \right)}{N - k - 1} \tag{3.8}$$

Poisson, negative binomial, and GP-1 regression techniques were explored by model generation in R. Models with many different combinations of regressor variables were created to find the model that performed best. In all models, the number of monthly crashes was used as the dependent variable, and the monthly CV volume was used as the exposure variable. The statistical models were evaluated based on the significance of the regressor variables used in the models, on the basis of the Akaike Information Criterion, and based on the residuals generated by the models. The best-performing models were selected and are summarized and discussed in the Results and Discussion sections.

### 3.4.4 Segment Analysis

The preliminary results from the intersection study prompted interest in how the results of an intersection-based study would compare to the results of a segment-based

study. To address this, a segment analysis was conducted. CV data was collected for thirty road segments in the Salt Lake City area. These segments include sections of interstate highway within the Salt Lake City limits and sections of interrupted state highways outside of the influence area of any intersections. The segments were all made to be approximately one-quarter mile in length.

The segment CV data was collected in the same manner as the intersection CV data with a couple of key differences. First, the intersection CV data was all collected from within intersection influence areas. The segment CV data was all collected from areas entirely outside of intersection influence areas. Second, the geometric information and information related to schools and bus stops were not collected for the segments. Rather, the segment data included only harsh braking events for jerk thresholds ranging between -1 $ft/s^3$ and -10 $ft/s^3$ in increments of 1 $ft/s^3$, as well as monthly CV counts, monthly CV counts between the hours of 7 AM and 9 AM, and monthly CV counts between the hours of 4 PM and 6 PM. As with the intersection analysis, crash data was collected for the segments from the UDOT database.

Statistical analysis was conducted in the same manner as the intersection analysis, with Poisson, negative binomial, and generalized Poisson models generated and evaluated for the segment dataset. The best-performing models were selected and are summarized and discussed in the following sections.

### 3.5 Results

The collected intersection data was used for a statistical regression analysis, and the best regression model for each of the model families was found that had a high level of significance among the regressor variables and the intercept. The best Poisson model

uses *Jerk18* and *Schools* from Table 3.1 as regressor variables. The best negative binomial model also uses *Jerk18* and *Schools* as regressor variables. The best generalized Poisson model uses *Jerk18* as a regressor variable. All of these models have a better than 0.1% significance level for their regressor variables and the intercept. In the case of the generalized Poisson model, both intercepts are significant at a better than 0.1% level. These models are summarized in Table 3.2.

**Table 3.2. Summary of Regression Models for Intersection Analysis**

| **Poisson Regression Model** | | | | | |
|---|---|---|---|---|---|
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -8.576 | 0.2483 | -34.544 | < 2e-16 | < 0.1% |
| Jerk18 | -4.056e-5 | 9.593e-6 | -4.228 | 2.35e-5 | < 0.1% |
| Schools | 1.103 | 0.3193 | 3.455 | 5.51e-4 | < 0.1% |
| Akaike Information Criterion | | 242.58 | | | |
| Log Likelihood | | -118.29 | | | |
| RMSE | | 0.9468 | | | |
| **Negative Binomial Regression Model** | | | | | |
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -8.526 | 0.2664 | -31.999 | < 2e-16 | < 0.1% |
| Jerk18 | -4.127e-5 | 1.037e-5 | -3.981 | 6.87e-5 | < 0.1% |
| Schools | 1.190 | 0.3566 | 3.337 | 8.46e-4 | < 0.1% |
| Akaike Information Criterion | | 242.67 | | | |
| Log Likelihood | | -117.337 | | | |
| Theta | | 3.87 | | | |
| RMSE | | 0.9627 | | | |
| **Generalized Poisson Regression Model** | | | | | |
| Parameter | Estimate | Std. Error | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept 1 | -8.264 | 0.2342 | -35.288 | < 2e-16 | < 0.1% |
| Intercept 2 | -11.81 | 1.741 | -6.782 | 1.19e-11 | < 0.1% |
| Jerk18 | -4.890e-5 | 1.019e-5 | -4.797 | 1.61e-6 | < 0.1% |
| Log Likelihood | | -122.8911 | | | |
| Degrees of Freedom | | 197 | | | |
| RMSE | | 0.9671 | | | |

The segment analysis also yielded three statistical models: a Poisson regression model, a negative binomial regression model, and a generalized Poisson regression model. The best Poisson, negative binomial, and generalized Poisson models identified use *Jerk2* as a regressor variable. All models have a better than 0.1% significance level for their regressor variable and intercept(s). These models are summarized in Table 3.3.

**Table 3.3. Summary of Regression Models for Segment Analysis**

| **Poisson Regression Model** | | | | | |
|---|---|---|---|---|---|
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -8.881 | 0.2338 | -37.991 | < 2e-16 | < 0.1% |
| Jerk2 | -1.345e-5 | 1.661e-6 | -8.098 | 5.57e-16 | < 0.1% |
| Akaike Information Criterion | | 199.32 | | | |
| Log Likelihood | | 97.658 | | | |
| RMSE | | 1.5102 | | | |
| **Negative Binomial Regression Model** | | | | | |
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -8.911 | 0.3329 | -26.768 | < 2e-16 | < 0.1% |
| Jerk2 | -1.212e-5 | 2.126e-6 | -5.702 | 1.19e-8 | < 0.1% |
| Akaike Information Criterion | | 181.21 | | | |
| Log Likelihood | | -87.604 | | | |
| Theta | | 0.880 | | | |
| RMSE | | 1.5621 | | | |
| **Generalized Poisson Regression Model** | | | | | |
| Parameter | Estimate | Std. Error | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept 1 | -8.878 | 0.2416 | -36.750 | < 2e-16 | < 0.1% |
| Intercept 2 | -12.84 | 1.057 | -12.149 | < 2e-16 | < 0.1% |
| Jerk2 | -1.335e-5 | 1.799e-6 | -7.425 | 1.13e-13 | < 0.1% |
| Log Likelihood | | -96.9352 | | | |
| Degrees of Freedom | | 117 | | | |
| RMSE | | 1.5143 | | | |

The estimates for the coefficients of the harsh braking variable in each of these regression models (*Jerk18* and *Jerk2*) are all negative, indicating that an increase in hard braking events decreases the estimate for the number of crashes that will occur within the intersection area of along the segment in question. This suggests that hard braking events are an indication of safety. This is true at intersections as well as on segments away from the influence of intersections.

Validation efforts conducted with the models produced the following graphs, displayed in Figure 8 and Figure 9. These graphs display the expected monthly crash counts for each of the three models on the vertical axis. The horizontal axis represents the observed monthly crash counts that correspond to each of the expected crash counts. The "jitter" function in R has been used to generate these plots; hence, there is scatter around the integer counts of observed crashes.

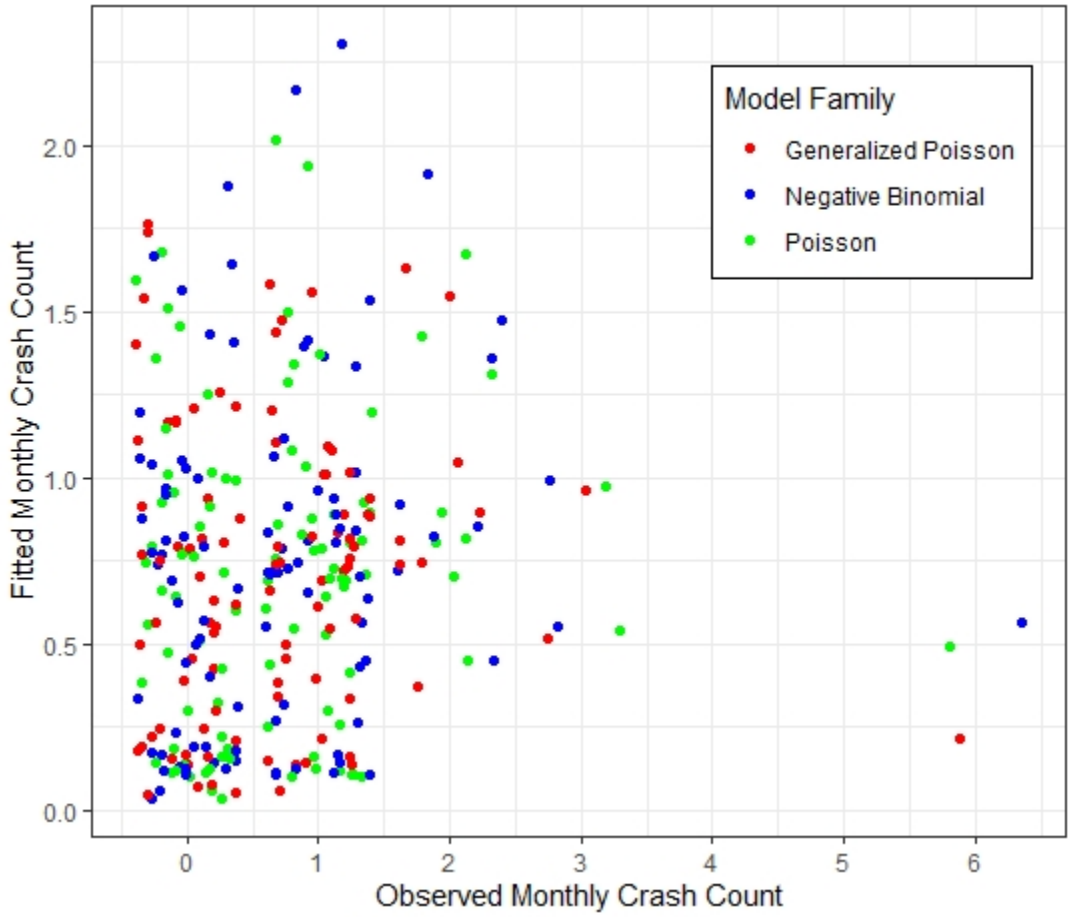


**Figure 8. Intersection Fitted Crash Counts versus Observed Crash Counts**

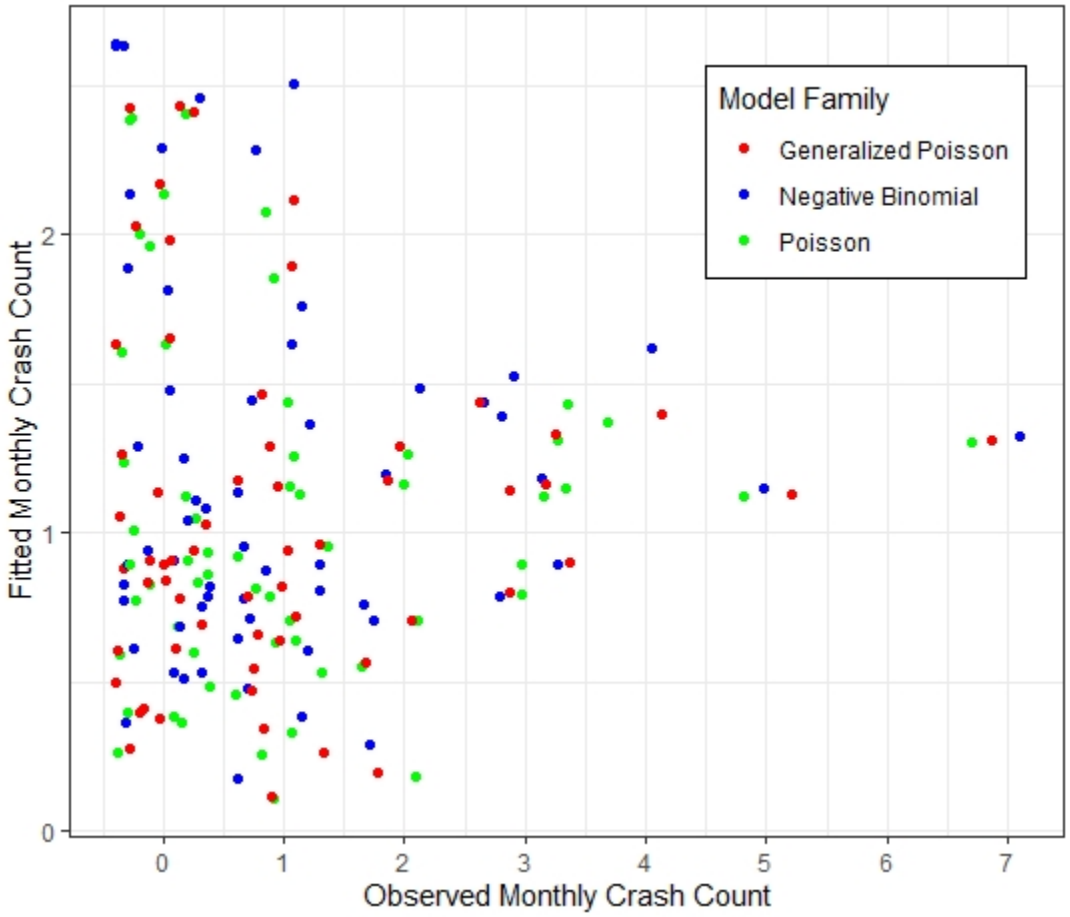


**Figure 9. Segment Fitted Crash Counts versus Observed Crash Counts**

An additional analysis was conducted in the same manner as that which yielded the results presented up to this point, except with outlier crash counts removed from the intersection and segment datasets. The outliers were identified using boxplots generated for the observed crash counts. These boxplots are presented in Figure 10. The outliers are denoted as black points in Figure 10. The best identified Poisson, negative binomial, and generalized Poisson models are summarized in Tables 3.4 and 3.5.

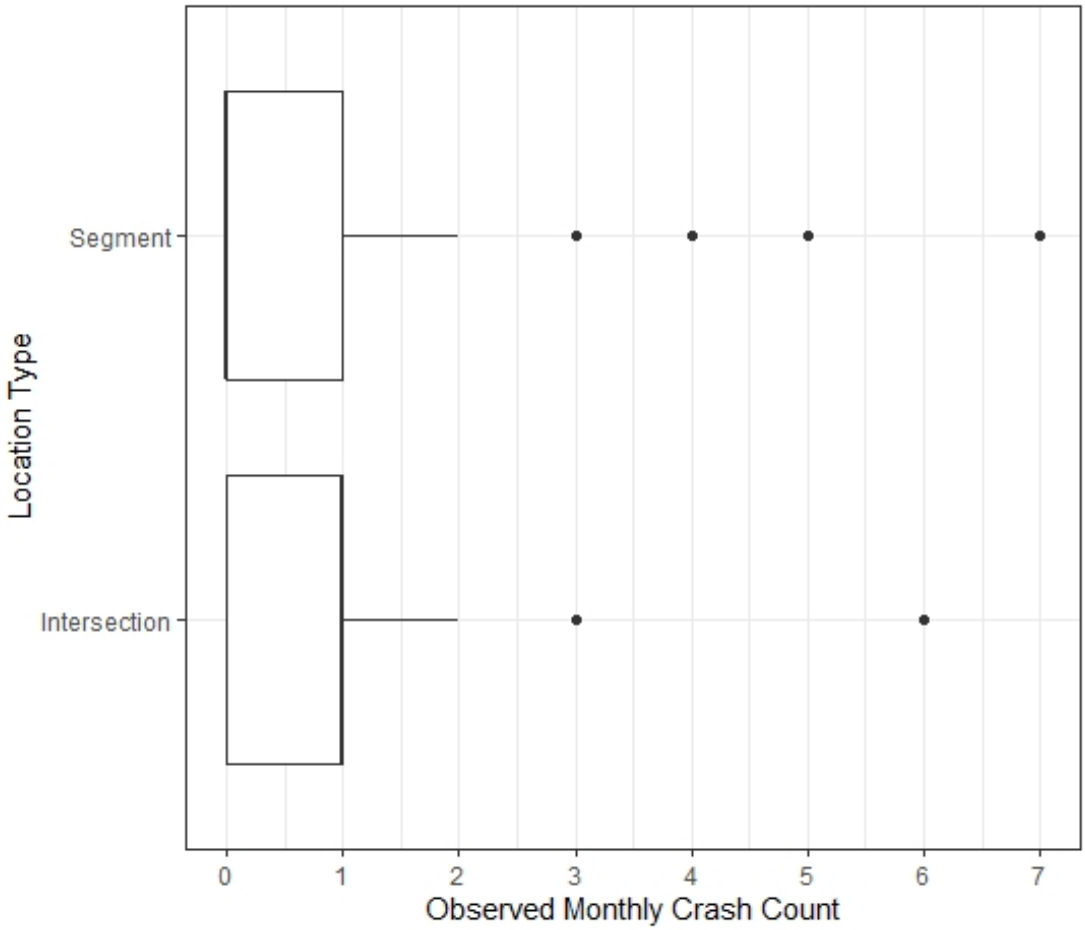


**Figure 10. Boxplots of the Observed Monthly Crash Counts at Intersections and Segments**

**Table 3.4. Summary of Regression Models for Intersection Analysis with Outliers Removed**

| **Poisson Regression Model** | | | | | |
|---|---|---|---|---|---|
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -7.722 | 0.3456 | -22.345 | < 2e-16 | < 0.1% |
| Jerk1 | -9.144e-6 | 1.820e-6 | -5.024 | 5.05e-7 | < 0.1% |
| Left Turn Approaches | -0.2181 | 9.278e-2 | -2.351 | 0.0187 | < 5% |
| Akaike Information Criterion | | 203.97 | | | |
| Log Likelihood | | -98.9848 | | | |
| RMSE | | 0.6348 | | | |
| **Negative Binomial Regression Model** | | | | | |
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -7.722 | 0.3456 | -22.342 | < 2e-16 | < 0.1% |
| Jerk1 | 9.145e-6 | 1.820e-6 | 5.024 | 5.06e-7 | < 0.1% |
| Left Turn Approaches | -0.2181 | 9.279e-2 | 2.350 | 0.0188 | < 5% |
| Akaike Information Criterion | | 205.97 | | | |
| Log Likelihood | | -98.9875 | | | |
| Theta | | 4676 | | | |
| RMSE | | 0.6348 | | | |
| **Generalized Poisson Regression Model** | | | | | |
| Parameter | Estimate | Std. Error | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept 1 | -7.722 | 0.3456 | -22.345 | < 2e-16 | < 0.1% |
| Intercept 2 | 38.95 | 7.415e+4 | -0.001 | 0.9996 | None |
| Jerk1 | -9.144e-6 | 1.820e-6 | -5.024 | 5.05e-7 | < 0.1% |
| Left Turn Approaches | -0.2181 | 9.278e-2 | -2.351 | 0.0187 | < 5% |
| Log Likelihood | | -98.9848 | | | |
| Degrees of Freedom | | 196 | | | |
| RMSE | | 0.6348 | | | |

**Table 3.5. Summary of Regression Models for Segment Analysis with Outliers Removed**

| **Poisson Regression Model** | | | | | |
|---|---|---|---|---|---|
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -10.55 | 0.3682 | -28.643 | < 2e-16 | < 0.1% |
| Jerk2 | 6.322e-6 | 2.083e-6 | -3.035 | 2.41e-3 | < 1% |
| Akaike Information Criterion | | 137.01 | | | |
| Log Likelihood | | -66.5050 | | | |
| RMSE | | 0.7653 | | | |
| **Negative Binomial Regression Model** | | | | | |
| Parameter | Estimate | Std. Err. | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept | -10.47 | 0.3765 | -27.812 | < 2e-16 | < 0.1% |
| Jerk2 | -6.638e-6 | 2.147e-6 | -3.091 | 1.99e-3 | < 1% |
| Akaike Information Criterion | | 138.93 | | | |
| Log Likelihood | | -66.4645 | | | |
| Theta | | 6.7 | | | |
| RMSE | | 0.7730 | | | |
| **Generalized Poisson Regression Model** | | | | | |
| Parameter | Estimate | Std. Error | Z-Score | Pr(>\|z\|) | Significance Level |
| Intercept 1 | -10.55 | 0.3682 | -28.644 | < 2e-16 | < 0.1% |
| Intercept 2 | -38.46 | 9.572e+4 | 0.000 | 0.99968 | None |
| Jerk2 | -6.322e-6 | 2.083e-6 | -3.035 | 0.00241 | < 1% |
| Log Likelihood | | -66.505 | | | |
| Degrees of Freedom | | 117 | | | |
| RMSE | | 0.7653 | | | |

Validation efforts were conducted for the models generated with outlier crash counts removed from the datasets. These validation efforts produced the graphs displayed in Figures 11 and 12. These graphs display the expected monthly crash counts for each of the three models on the vertical axis. The horizontal axis represents the observed monthly crash counts that correspond to each of the expected crash counts.

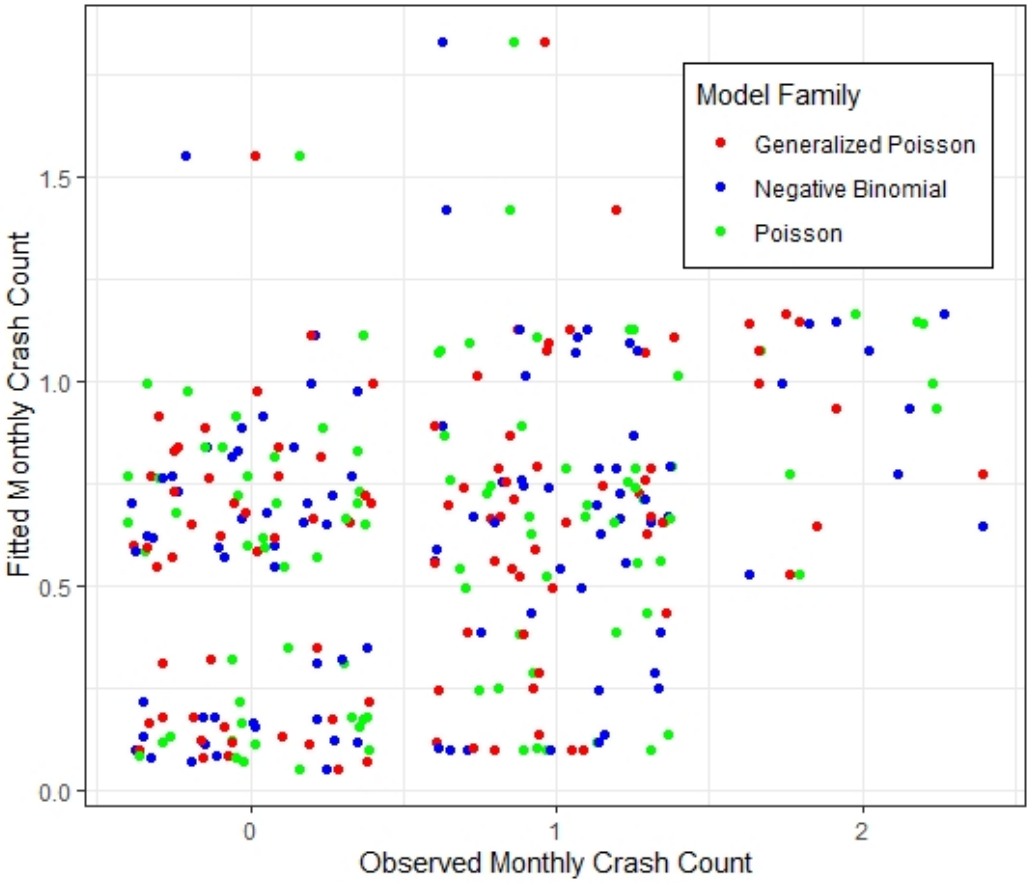


**Figure 11. Intersection Fitted Crash Counts versus Observed Crash Counts with Outliers Removed**

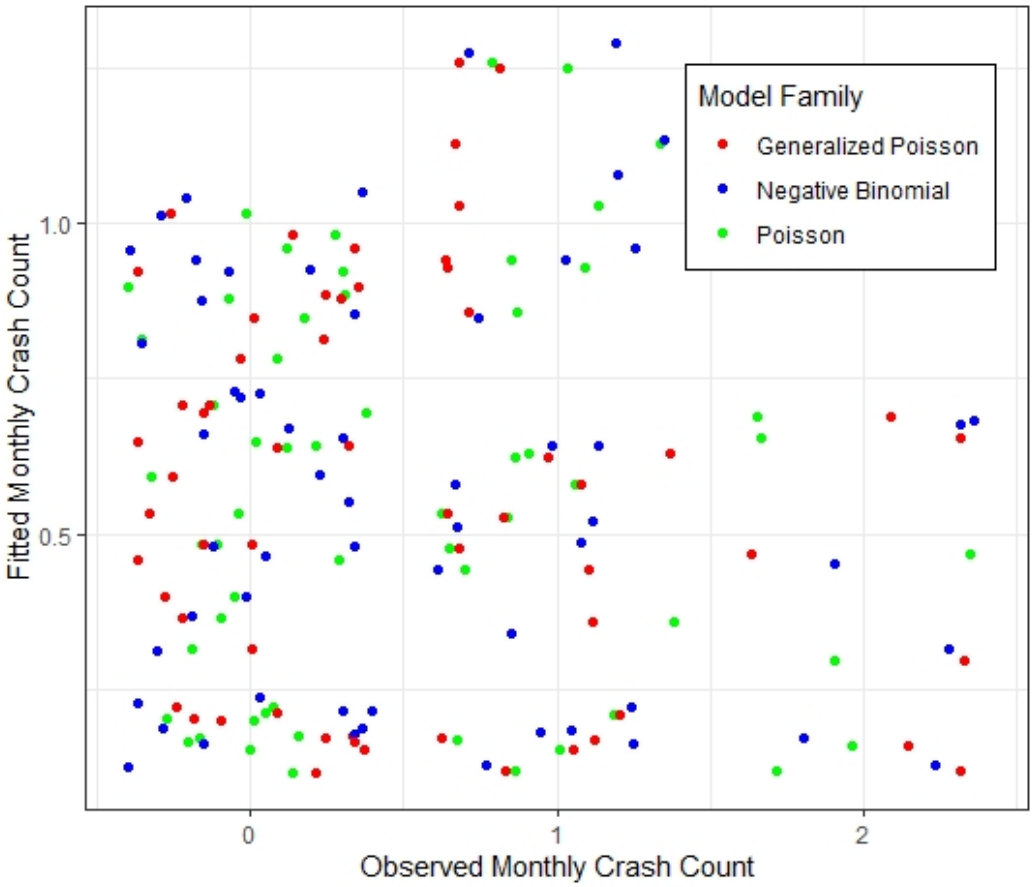


**Figure 12. Segment Fitted Crash Counts versus Observed Crash Counts with Outliers Removed**

### 3.6 Discussion

This study demonstrates the effectiveness of using harsh braking data from CVs as a surrogate safety measure. For both intersections and segments, statistically significant models may be developed from multiple model families. Models such as these may be used to predict future crash rates for the purposes of prioritizing improvements and identifying risks to the public.

The results of this study reveal the jerk threshold for intersections and segments. For intersections, the jerk threshold is -9.0 ft/s$^3$, corresponding to the regressor variable *Jerk18*. This threshold was identified to be the most effective for all three statistical model families. The jerk threshold for segments was found to be -1.0 ft/s$^3$, corresponding

to the variable *Jerk2*. A jerk threshold of -9.0 ft/s$^3$ for intersections and -1.0 ft/s$^3$ for segments indicates that intersections and segments operate differently in terms of safety. The jerk threshold is the value of jerk that differentiates ordinary events from harsh braking events. Events that do not meet the jerk threshold have little or no bearing on crash prediction. A larger absolute value of jerk threshold for intersections over segments indicates that braking must be more severe at an intersection to qualify a braking event as *harsh*. This could be due to different expectations of drivers in these differing contexts. At intersections, drivers expect to brake and are typically able to see the status of the traffic signal well in advance. Moderately hard braking, such as an event that generates a jerk value of -5.0 ft/s$^3$, is expected and therefore ordinary. Such an event in a segment context, however, would be relatively unexpected and therefore extraordinary because segments are expected to have more uniform and smooth flow. This event would therefore qualify as a harsh braking event in a segment context but not in an intersection context.

The coefficient estimates for the harsh braking event count variables were found to be negative for all statistical models generated, indicating that an increase in harsh braking events is correlated with a decrease in the crash frequency. The coefficient estimates for the jerk variable is small relative to other covariates, when the covariates are statistically significant. For the intersection analysis, the results from the Poisson regression indicate that the effect of the jerk variable is to reduce monthly crashes by a negligible amount, but the presence of a school within 1000 ft is expected to increase monthly crashes by 300%. This study was predicated upon the notion that a harsh braking event corresponds to a traffic conflict and that traffic conflicts and collisions are related.

That these models have negative estimates for the coefficients of the harsh braking event counts suggests that harsh braking events are not just indicative of a traffic conflict occurring but rather of a traffic conflict being prevented. Harsh braking events are events which might have been collisions but never were as a result of the evasive action of the drivers involved.

The statistical models presented in Tables 3.2 and 3.3 possess an excellent level of statistical significance at a level better than 0.1%. Poisson models are simpler than negative binomial models and generalized Poisson models, making them preferable if applicable. While the requirement that the mean and variance of the dependent variable be equal was approximately satisfied for the dataset used in this study, the reliability of that assumption is questionable. Therefore, negative binomial and generalized Poisson regression are recommended for crash prediction models based upon harsh braking data.

As mentioned above, the presence of schools was found to increase crash frequency within intersection influence areas. This confirms the efficacy of the use of school presence in HSM safety analysis methodology. The estimated coefficients for the *Schools* variable are positive, indicating that the presence of a school or multiple schools nearby increases the frequency with which crashes are expected to occur. The presence of schools increases pedestrian activity and the presence of young drivers which may help explain this increase.

The graphs presented in Figures 8 and 9 illustrate that these models fail to predict high crash counts while performing better at locations with lower numbers of observed crashes. While the predictive ability of these models leaves much to be desired, the statistical significance of the regressor variables in the models speaks to their overall

strength. As CV penetration rates increase, allowing models based on CV data to be trained by a fuller picture of the activity on roads, models of this form will likely become more effective. Preliminary studies such as this, using CV data in its technological infancy, set the stage for a future in which CVs become significantly more widespread and CV data captures a large portion if not a majority of roadway traffic. Figures 11 and 12, as well as the RMSE values presented in Tables 3.4 and 3.5 demonstrate that the model's predictive ability improves when outliers are removed. The RMSE values for intersections decreased from approximately 0.95 to 0.63 for intersections and from approximately 1.55 to 0.76 for segments. The decrease in RMSE indicates that the models produce more accurate crash count estimates when outliers are removed.

## 3.7 Conclusions

This study developed several statistical models which use harsh braking event counts from CV data in Salt Lake City as a regressor variable and crash counts as the dependent variable. Both intersections and segments were considered separately in this study with models derived for each. Poisson, negative binomial, and generalized Poisson models were developed which revealed the jerk threshold for intersection influence areas to be -9.0 $ft/s^3$ and the jerk threshold for segments to be -1.0 $ft/s^3$. Additionally, the presence of schools within 1,000 ft was found to be a statistically significant variable for intersection influence areas.

Crash prediction models such as these, based on harsh braking event counts, hold promise for agencies and industry as another tool for safety analysis. Agencies may investigate these models and tailor them to their jurisdictions for the purpose of adding such models to their established methodologies. Such tailored models may then be

employed as a means of conducting comparative safety analysis for the purpose of identifying crash prone locations and prioritizing improvements. Once a particular area is identified as being crash prone, further investigation into the cause of the safety hazard may commence. The location may suffer from a sight distance problem, a lack of capacity which causes a bottleneck, or some other condition. In western cities such as Salt Lake City which are experiencing large amounts of new construction, sight distance can change from year to year as new structures are built and volumes increase closer to their respective capacities as time goes on. Employing harsh braking models such as those developed in this study requires less labor investment than existing methods, allowing for more frequent and widespread analyses to identify and characterize road hazards.

Future research into SSMs that are based on harsh braking events could include investigation of regional differences in models, the use of additional regressor variables in segment-based models, and harsh positive acceleration data from CVs. Regional differences may exist pertaining to the relationship between harsh braking and collisions. It is possible that while harsh braking events may indicate a greater level of safety in Salt Lake City, they may indicate more dangerous conditions in other locations. Harsh braking events were found to be positively correlated to crashes in a previous study in Louisiana which is contrary to the findings of this study (Mousavi, 2015). While this may be due to the significant differences in the methods of data collection between these two studies, regional variations may also be a factor and ought to be investigated further. Additional regressor variables were not investigated in the segment-based models developed in this study to the degree to which they were investigated in the intersection-

based models. The inclusion of such additional regressor variables for segments ought to be investigated more fully in a future study. These variables may include speed limits, curvature parameters, lane widths, or total number of lanes, among others. Finally, harsh positive acceleration data may be obtained in the same manner in which harsh braking data was collected in this study. Harsh acceleration may be an indicator of safety or the lack thereof because it can represent erratic driving behavior or situations in which a driver is attempting to clear a potential crash location rapidly. Consideration of harsh acceleration data may be done separately from harsh braking data or in combination with harsh braking data. If attempts are successful, this would yield yet another tool for agencies and industry to employ for surrogate safety analysis.

## 3.8 Acknowledgments

The authors are grateful to the Boise State University Department of Civil Engineering for their support of this research. The authors would also like to express their gratitude to the support received from the PacTrans Region 10 University Transportation Center that made the procurement of the CV data possible.

## 3.9 Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: M. Khanal; data collection: N. Edelmann; analysis and interpretation of results: M. Khanal, N. Edelmann; draft manuscript preparation: N. Edelmann. All authors reviewed the results and approved the final version of the manuscript.

# CHAPTER 4: CONCLUSIONS

Data from connected vehicles has opened up many new possibilities for traffic analysis, including in the area of safety analysis. CV data allows for the incorporation of large amounts of real-world data into traffic analyses which would have been difficult or impossible to collect before this technology. In past research, safety analysis has been shown to be possible using traffic microsimulations. These simulations are based upon real-world data such as turning movement counts, peak hour factors, heavy vehicle percentages, and lane utilization factors, among others. This data is rudimentary, but it allows for microsimulations to generate kinematic data that may be used to compute surrogate safety measures. The strength of data generated with simulations over manually collected data is the precision and objectivity of the data, but simulation data comes at the cost of being a simulated representation of traffic rather than traffic itself. CV data allows for the precision and objectivity of simulation data along with the realism of being data from actual physical cars on the roads. Safety analysis with CV data benefits from these qualities.

Traffic safety analysis with surrogate safety measures is a field of research that has experienced development in recent years. Although some of the measures which are used today, such as proportion of stopping distance and post-encroachment time, were developed nearly fifty years ago, some newer measures and improved implementation techniques have been developed within the last twenty years. These efforts have sought to expand upon the measures to make them more descriptive of the events they represent,

such as the development of the extended surrogate safety measures. The implementation techniques have been improved first through the incorporation of simulations and more recently the use of CV data. In 2015, harsh braking data was demonstrated to be related to crash occurrence (Mousavi, 2015).

Harsh braking data as a surrogate safety measure is highly compatible with CV data, as the CV data includes information which may be used to derive counts of harsh braking events. CVs collect position and time data, among other data, which may be used to compute the value of the first derivative of acceleration, known as jerk. The jerk value for particular waypoints within the CV data can be compared to a jerk threshold value to determine if that particular waypoint qualifies as harsh braking or an ordinary event. The counts of harsh braking events may be used individually or in conjunction with other variables to develop crash prediction models, using historical crash data as training data.

Once harsh braking event models are trained with historical data, they may be applied to new circumstances, such as new construction or a modification to existing infrastructure, within a short period of time of the project opening. This period of time may be as short as a few months, in stark contrast to the years of waiting for crash data to accumulate for traditional safety analysis. CV data is collected automatically, reducing the need for data collection employees to be put in dangerous situations collecting data from the roadside. Even with the currently low penetration rates of CVs, the data produced by them is voluminous, making statistical techniques more effective for crash count prediction. As time goes on, penetration rates are expected to increase continuously until CVs make up a large portion of roadway traffic or even a majority of cars on the

road. Preliminary research, such as the one presented in this document, despite the low penetration rate, is necessary to prepare for this increase in CVs.

## 4.1 Summary of Work

A review of existing literature on surrogate safety measures was conducted and an investigation was undertaken into the use of harsh braking data from connected vehicles for statistical crash prediction models. The research work done in these two areas was used to generate two manuscripts which were submitted to be included in conferences and for potential publishing in journals associated with the conferences. The two manuscripts were presented in their entirety within this document.

The existing literature established surrogate safety measures and methods of implementation. Ten of the papers were selected for inclusion in a review paper which presented summaries of the important findings within each of the papers and discussed the value of the papers in several categories, including theoretical value, practical value, relevance, and difficulty of implementation. The manuscript generated from this work was submitted to the 2022 conference of the International Association of Journals and Conferences.

The research into the use of harsh braking data from CVs for crash prediction models involved the compilation of data for both intersections and segments within the Salt Lake City limits and the development of statistical regression models from that data. The data compiled included CV harsh braking data, CV volume data, crash data, and additional data for intersections which was explored for use as regressor variables, such as intersection geometric characteristics and nearby schools and bus stops. This additional data was not compiled for segments. Statistical regression analysis was

undertaken using these datasets with Poisson, Negative Binomial, and Generalized Poisson regression using R. The findings of these efforts are presented in a manuscript which was submitted to the 2023 Annual Meeting of the Transportation Research Board in Washington, D.C.

## 4.2 Implementation

The findings of these efforts may be applied by both researchers and practitioners. The review paper presented in Chapter 2 summarizes the important findings of a variety of papers on the subject of surrogate safety measures. This paper can serve as a first resource for practitioners looking to understand what surrogate safety measures are available to use and guide deeper research efforts as they work to implement these methods in practical applications. Researchers would benefit from the paper's illumination of what work has been accomplished so far and the summaries of the future research suggested by the authors of the papers included in the review.

The findings of the research into harsh braking may be implemented by transportation agencies looking to incorporate CV data into their safety analyses. This will likely need to wait until the CV penetration rates become higher than they are currently. Once implemented, agencies may use regression models, such as those presented in the paper presented in Chapter 3, to conduct widespread safety analyses within their jurisdictions more often than may be done currently using traditional safety analysis techniques. The results of these analyses may be used to inform decision making regarding the prioritization of safety-related improvements.

## 4.3 Recommendations for Future Work

In conducting this research, some potential topics for future research became clear. These topics include investigation of regional differences in harsh braking crash prediction models, the examination of additional regressor variables in segment-based models, and a similar study to that presented in the manuscript in Chapter 3 which uses harsh positive acceleration data rather than harsh braking data.

It is possible that there are regional differences from state to state or even city to city regarding the relationship between harsh braking event counts and crash counts. It is likely that there are at least slight variations in the estimates for the coefficients within the harsh braking models. It may even be that harsh braking events indicate a greater level of safety in Salt Lake City, the location of the study in Chapter 3, but indicate that a particular location is less safe in another city or region. Research into whether these variations exist and to characterize these variations should they be found to exist would be worthwhile, especially if agencies eventually adopt similar crash prediction models. Knowing whether and how to tailor models to a particular municipality would be helpful for transportation agencies.

The statistical significance of additional regressor variables in segment-based models may also be investigated in the future. Additional regressor variables were investigated for intersection-based models, including the numbers of approaches with left and right turn lanes, the maximum number of lanes a pedestrian would need to cross, and the numbers of bus stops and schools within 1,000 ft from the center point of the intersection. These variables were not examined for segments as they were not applicable, but there may be other similar variables which may be statistically significant

in predicting crash frequency. Some examples might be curvature parameters, speed limits, lane widths, or the total number of lanes in a segment, among other possibilities.

Using harsh positive acceleration data from CVs is another potential research topic. A similar study to that in Chapter 3 may be undertaken using harsh positive acceleration data. Harsh acceleration can be indicative of erratic driving behavior or of a driver quickly clearing a location where a collision could have occurred. Statistical regression models may be developed with harsh acceleration event counts used as a regressor variable. This variable may be used either separately or in combination with harsh braking data.

# REFERENCES

AASHTO. (2010). *Highway Safety Manual*: American Association of State Highway and Transportation Officials.

Allen, B. L., Shin, B. T., and Cooper, D.J. (1978). "Analysis of Traffic Conflicts and Collisions". *Transportation Research Record* 667, 67-74.

Astarita, V., et al. (2020). Surrogate Safety Measures from Traffic Simulation: Validation of Safety Indicators with Intersection Traffic Crash Data. *Sustainability*, *6974*(12), 1-20.

Bagdadi, O. & Várhelyi, A. (2011). Jerky driving – An indicator of accident proneness? *Accident Analysis and Prevention*, 1359-1363.

Date, S. (No date). *Time Series Analysis, Regression and Forecasting: The Generalized Poisson Regression Model*. Retrieved from https://timeseriesreasoning.com/contents/generalized-poisson-regression-model/

Gettman, D. & Head, L. (2003). *Surrogate Safety Measures from Traffic Simulation Models Final Report.* U.S. Department of Transportation Federal Highway Administration Office of Research, Development, and Technology. https://doi.org/10.3141/1840-12

Guido, G., Saccomanno, F., Vitale, A., Astarita, V., & Festa, D. (2010). "Comparing Safety Performance Measures Obtained from Video Capture Data". *Journal of Transportation Engineering 137*(7), 481-491.

He, Z., et al. (2018). Assessing Surrogate Safety Measures Using a Safety Pilot Model Deployment Dataset. *Transportation Research Record*, *2672*(38), 1-11.

Hilbe, J.M. (2011). *Negative Binomial Regression: Second Edition*: Cambridge University Press.

Hunter, M., Mathew, J.K., Li, H., & Bullock, D.M. (2021). Estimation of Connected Vehicle Penetration on US Roads in Indiana, Ohio, and Pennsylvania. *Journal of Transportation Technologies*, *11*, 597-610. https://doi.org/10.4236/jtts.2021.114037

Minderhoud, M. M. & Bovy, P.H. (2001). "Extended Time-to-Collision Measures for Road Traffic Safety Assessment". *Accident Analysis and Prevention 33*(1), 89-97.

Mousavi, S. M. (2015). *Identifying High Crash Risk Roadways through Jerk-Cluster Analysis* (Master's thesis, Louisiana State University, Baton Rouge, Louisiana, United States). Retrieved from https://digitalcommons.lsu.edu/gradschool_theses/159

NCSS. (2016a). Negative Binomial Regression. *NCSS Statistical Software*. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf

NCSS. (2016b). Poisson Regression. *NCSS Statistical Software*. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Poisson_Regression.pdf

Souza, J. Q., Sasaki, M. W., & Cunto, F. J. C. (2011, September 14-16). *Comparing Simulated Road Safety Performance to Observed Crash Frequency at Signalized Intersections.* International Conference on Road Safety and Simulation, Indiana, USA.

Tarko, A. P. (2012). Use of Crash Surrogates and Exceedance Statistics to Estimate Road Safety. *Accident Analysis and Prevention*, *45*, 230-240.

TRB. (2016). *Highway Capacity Manual, Sixth Edition: A Guide for Multimodal Mobility Analysis*: Transportation Research Board National Research Council.

Wang, C. & Stamatiadis, N. (2013). Surrogate Safety Measure for Simulation-Based Conflict Study. *Transportation Research Record*, *2386*, 72-80.

Wang, J., Huang, H., & Zeng, Q. (2017). The effect of zonal factors in estimating crash risks by transportation modes: Motor vehicle, bicycle, and pedestrian. *Accident Analysis and Prevention*, *98*, 223-231.

# APPENDIX: R CODES FOR SELECTED MODELS

## Intersection_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Tidy_Data_All_Months.csv")
allmonths <- spreadsheet[1:100,]

# Selected Model:

pois173 <- glm(Monthly_Crashes ~ Jerk18 + `Schools_within_1,000_ft` + offset(log(Monthly_CVs)),
               "poisson", data = allmonths); summary(pois173)
```

## Intersection_Negative_Binomial_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Tidy_Data_All_Months.csv")
allmonths <- spreadsheet[1:100,]

# Selected Model:

negb173 <- glm.nb(Monthly_Crashes ~ Jerk18 + `Schools_within_1,000_ft` + offset(log(Monthly_CVs)),
            data = allmonths); summary(negb173)
```

## Intersection_Generalized_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
library(VGAM)
spreadsheet <- read_csv("CV_Tidy_Data_All_Months.csv")
allmonths <- spreadsheet[1:100,]

# Selected Model:

genp18 <- vglm(Monthly_Crashes ~ Jerk18 + offset(log(Monthly_CVs)),
             genpoisson1, data = allmonths); summary(genp18)
```

## Segment_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Segment_Data.csv")
segmentdata <- spreadsheet[1:60,]
glimpse(segmentdata)
#check to see if mean of monthly crashes = stand. dev.
mean(segmentdata$Monthly_Crashes)
var(segmentdata$Monthly_Crashes)
#Mean = 0.8222, variance = 1.564

# Selected model:

pois1 <- glm(Monthly_Crashes ~ Jerk2 + offset(log(Monthly_CVs)),
             "poisson", data = segmentdata); summary(pois1)
```

## Segment_Negative_Binomial_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Segment_Data.csv")
segmentdata <- spreadsheet[1:60,]
glimpse(segmentdata)
#check to see if mean of monthly crashes = stand. dev.
mean(segmentdata$Monthly_Crashes)
var(segmentdata$Monthly_Crashes)
#Mean = 0.8222, variance = 1.564

# Selected model:

negb1 <- glm.nb(Monthly_Crashes ~ Jerk2 + offset(log(Monthly_CVs)),
            data = segmentdata); summary(negb1)
```

## Segment_Generalized_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
library(VGAM)
spreadsheet <- read_csv("CV_Segment_Data.csv")
segmentdata <- spreadsheet[1:60,]
glimpse(segmentdata)
#check to see if mean of monthly crashes = stand. dev.
mean(segmentdata$Monthly_Crashes)
var(segmentdata$Monthly_Crashes)
#Mean = 0.8222, variance = 1.564

# Selected model:

genp1 <- vglm(Monthly_Crashes ~ Jerk2 + offset(log(Monthly_CVs)),
            genpoisson1, data = segmentdata); summary(genp1)
```

## No_Outliers_Intersection_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Tidy_Data_All_Months_no_outliers.csv")
allmonths <- spreadsheet[1:100,]

# Selected Model

pois33 <- glm(Monthly_Crashes ~ Jerk1 + Approaches_with_Left_Turn_Lanes + offset(log(Monthly_CVs)),
            "poisson", data = allmonths); summary(pois33)
```

## No_Outliers_Intersection_Negative_Binomial_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Tidy_Data_All_Months_no_outliers.csv")
allmonths <- spreadsheet[1:100,]

# Selected Model

negb33 <- glm.nb(Monthly_Crashes ~ Jerk1 + Approaches_with_Left_Turn_Lanes + offset(log(Monthly_CVs)),
            data = allmonths); summary(negb33)
```

## No_Outliers_Intersection_Generalized_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
library(VGAM)
spreadsheet <- read_csv("CV_Tidy_Data_All_Months_no_outliers.csv")
allmonths <- spreadsheet[1:100,]


# Selected Model:

genp33 <- vglm(Monthly_Crashes ~ Jerk1 + Approaches_with_Left_Turn_Lanes + offset(log(Monthly_CVs)),
               genpoisson1, data = allmonths); summary(genp33)
```

## No_Outliers_Segment_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Segment_Data_no_outliers.csv")
segmentdata <- spreadsheet[1:60,]
glimpse(segmentdata)
#check to see if mean of monthly crashes = stand. dev.
mean(segmentdata$Monthly_Crashes)
var(segmentdata$Monthly_Crashes)
#Mean = 0.8222, variance = 1.564

# Selected model:

pois1 <- glm(Monthly_Crashes ~ Jerk2 + offset(log(Monthly_CVs)),
             "poisson", data = segmentdata); summary(pois1)
```

## No_Outliers_Segment_Negative_Binomial_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
spreadsheet <- read_csv("CV_Segment_Data_no_outliers.csv")
segmentdata <- spreadsheet[1:60,]
glimpse(segmentdata)
#check to see if mean of monthly crashes = stand. dev.
mean(segmentdata$Monthly_Crashes)
var(segmentdata$Monthly_Crashes)
#Mean = 0.8222, variance = 1.564

# Selected model:

negb1 <- glm.nb(Monthly_Crashes ~ Jerk2 + offset(log(Monthly_CVs)),
              data = segmentdata); summary(negb1)
```

## No_Outliers_Segment_Generalized_Poisson_Selected_Model.R

```r
library(tidyverse)
library(summarytools)
library(MASS)
library(modelr)
library(VGAM)
spreadsheet <- read_csv("CV_Segment_Data_no_outliers.csv")
segmentdata <- spreadsheet[1:60,]
glimpse(segmentdata)
#check to see if mean of monthly crashes = stand. dev.
mean(segmentdata$Monthly_Crashes)
var(segmentdata$Monthly_Crashes)
#Mean = 0.8222, variance = 1.564

# Selected model:

genp1 <- vglm(Monthly_Crashes ~ Jerk2 + offset(log(Monthly_CVs)),
              genpoisson1, data = segmentdata); summary(genp1)
```