A Data Driven Approach to Quantifying the Effect of Crashes

Final Report

(Year 2)

SOUTHEASTERN TRANSPORTATION CENTER

Obaidur Kazi, Xu Zhang, and Mei Chen

February 2018

US DEPARTMENT OF TRANSPORTATION GRANT DTRT13-G-UTC34

DISCLAIMER

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle A Data Driven Approach to Quantifying the Effect of Crashes		5. Report Date February 2018
		6. Source Organization Code
7. Author(s) Xu Zhang, Obaidur Kazi, Mei Chen		8. Source Organization Report No.
9. Performing Organization Name and Address Southeastern Transportation Center UT Center for Transportation Research 309 Conference Center Building Knoxville TN 37996-4133		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. DTRT13-G-UTC34
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary of Transportation–Research 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Year 2 Final Report: 1/2016-12/2017
		14. Sponsoring Agency Code USDOT/OST-R/STC
15. Supplementary Notes:		

Abstract
The growth of data has begun to transform the transportation research and policy, and opened a new window for analyzing the impact of crashes. Currently for the crash impact analysis, researchers tend to rely on the reported incident duration, which may not always be accurate. Further, the queue resulted from the crash could propagate to the upstream and take much longer to return to normal conditions than the traffic at the crash site. Additional complications arise when the crashes occur under congested conditions, in which case the crash-induced congestion need to be separated from the recurring congestion or other events' impacts. To address these concerns, a difference-in-speed approach is developed in this research to estimate the true crash impact duration using stationary sensor data and incident logs. The proposed method utilizes the Kalman Filter algorithm to combine real-time conditions and historical patterns to establish travelers' anticipated travel speeds under incident-free conditions and then employ the difference-in-speed approach to quantify the temporal and spatial extent of the crash impact. Subsequently, practical applications based on the model outcomes are presented. This includes the prediction of the impact duration using several statistical models and the evaluation of the impact of crashes on the travel time reliability by different periods of the day and crash types. This study illustrates the usages of data to improve the understanding of traffic crashes and their impacts on the congestion and reliability. The results of the study can aid in the incident management decision-making process for strategies that reduce the impact of the crashes.

16. Key Words	17. Distribution Statement
Heat map, space-time velocity map, incident, crash analysis, big data	Unrestricted; Document is available to the public through the National Technical Information Service; Springfield, VT.

18. Security Classif. (of this report)	19. Security Classif. (of this page)	20. No. of Pages	21. Price
Unclassified	Unclassified		…

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

The growth of data has begun to transform the transportation research and policy, and opened a new window for analyzing the impact of crashes. Currently for the crash impact analysis, researchers tend to rely on the reported incident duration, which may not always be accurate. Further, the queue resulted from the crash could propagate to the upstream and take much longer to return to normal conditions than the traffic at the crash site. Additional complications arise when the crashes occur under congested conditions, in which case the crash-induced congestion need to be separated from the recurring congestion or other events' impacts. To address these concerns, a difference-in-speed approach is developed in this research to estimate the true crash impact duration using stationary sensor data and incident logs. The proposed method utilizes the Kalman Filter algorithm to combine real-time conditions and historical patterns to establish travelers' anticipated travel speeds under incident-free conditions and then employ the difference-in-speed approach to quantify the temporal and spatial extent of the crash impact. Subsequently, practical applications based on the model outcomes are presented. This includes the prediction of the impact duration using several statistical models and the evaluation of the impact of crashes on the travel time reliability by different periods of the day and crash types. This study illustrates the usages of data to improve the understanding of traffic crashes and their impacts on the congestion and reliability. The results of the study can aid in the incident management decision-making process for strategies that reduce the impact of the crashes.

1. Introduction

1.1 Background

The growth of data has begun to transform research and policy, and transportation is no exception. As detection and communication technologies keep advancing, transportation sensors continue to generate more and more data. Abundant data is now available on the infrastructure conditions, operating characteristics, and traveler behaviors across temporal and spatial domains. This data presents an opportunity to better understand the interactions between travelers and the surrounding environment. It also opens a new window for comprehensive analysis of traffic incidents and their spatiotemporal impact on the traffic flow. This is of particular value because traffic incidents are the major causes of congestion and travel time unreliability on a transportation network, which seriously affect travel experience and cause significant economic and environmental losses.

Congestion on the roadways has been one of the major problems in the U.S. from both travelers' and operational point of view. It affects the nation's economy in many ways, such as causing thousands of unproductive hours, drastically altering the transportation schedule, causing distress to the travelers, and increasing environmental pollution. Due to the intricate interaction between uncertain demand and supply, many factors contribute to the occurrence of traffic congestion when traffic demand exceeds the physical capacity of the roadway. Those factors can be from both endogenous and exogenous sides of the transportation system, such as crashes, road work, inclement weather, special events, signal malfunctioning, and demand fluctuation, etc. Based on the cause of the congestion, it can be divided into two main groups: the recurrent and non-recurrent congestion. Recurrent congestion occurs almost every day at the same times of the day, i.e. peak periods, when the demand exceeds the capacity. On the other hand, the non-recurrent congestion occurs due to abnormal or unpredictable events from time to time that cause surge of the demand or reduction of the capacity[1].

To combat the congestion issue, many strategies have been put forward and can be categorized as follows:

- Expanding the capacity of the existing transportation system
- Increasing operational efficiency of the existing infrastructure capacity
- Reducing the unnecessary travel demand or shifting the demand during peak periods to non-peak periods

The experience over the years in the transportation field has suggested that the most efficient, intelligent and economical solution to alleviate the congestion is to increase the operational efficiency of the existing transportation infrastructure. One of the significant strategies falling under this category is the Intelligent Transportation System (ITS), which contains a broad range of design, control, and communication techniques ensuring safe, fast, smooth and economical transporting of people and goods.

One important strategy under the ITS is the Traffic Incident Management (TIM), which aims to reduce the duration and impact of incidents, and improve the safety of motorists and crash victims through the planned, systematic and coordinated use of human, institutional and technical resources [2]. As the main purpose of the TIM is to reduce the influence of traffic incidents, it requires a timely and precise estimation of traffic incident duration. By having a reliable prediction of incident duration, traffic managers could deploy appropriate resources to and around the traffic incident scene and provide travelers with real-time information to reduce incident related traffic congestion.

Several states' Department of Transportation (DOTs) have developed the TIM program to evaluate their performance in responding to the incidents, reducing the incident duration and restoring traffic operation. These efforts have proven beneficial in terms of the return on capital investment. According to the 2012 Urban Mobility Report, incident management treatments have saved about 7.2 billion dollars from the reduction of 337 million hours of delay.

To help transportation agencies in developing more efficient TIM programs, it is desirable to examine traffic incident's impact on the traffic flow, a need that has drew increasing attention from the researchers. Currently for the crash impact analysis, agencies tend to rely on the reported incident logs. Yet, there are some issues regarding the reported time frame. First, the record may not be always accurate. It is found that the recorded end time for many crashes is not reflective of the true end time of the crash impact. The end time of some crashes may even be incomplete. Second, even if the timeline of the crash is accurate, due to queue spillback, the upstream traffic may take longer to return to normal conditions than the traffic at the crash site. Additional complications arise when the crashes occur under congested conditions, in which case the crash-induced congestion need to be separated from the recurring congestion or other events' impacts. Therefore, it is imperative to capture the "true" impact of the crashes.

1.2 Existing Studies

A large number of studies have focused on examining the incident duration, which is defined as the elapsed time from the moment an incident is detected until the cause is removed from the scene [3]. Over the last few years, various methodologies and techniques have been used to model and analyze the incident duration. These models mainly establish the relationship between incident duration and different influencing factors. A set of variables significantly affecting incident duration have been identified as follows:

- Incident type and severity,
- Number and type of vehicles involved,
- Geometric characteristics,
- Temporal characteristics,
- Environmental effects, and
- Operational factors.

The most representative approaches for incident duration models are described in details as follows.

Linear regression analyses. Garib et al. developed linear regression models to estimate the magnitude and duration of freeway incident delays [3]. The author developed a multiple linear regression model based on 205 incidents over a two-month period from Oakland, California, to predict the incident duration. The duration is modeled as a function of six significant variables: number of lanes affected (X_1), number of vehicles involved (X_2), truck involvement as binary variable (X_3), natural logarithm of the police response time (X_4), time of day as binary variable (X_5) and weather conditions as binary variable (X_6). Then the final log-based regression model is given by $\text{Log (Duration)} = 0.87 + 0.027\, X_1\, X_2 + 0.2\, X_3 + 0.68\, X_4 - 0.17\, X_5 - 0.24\, X_6$

Non-parametric regression methods. The basis of the nonparametric regression is to make current decisions based on past, similar experience. Smith and Smith used a non-parametric regression model to predict the incident duration; however, the performance of the model was unsatisfactory, with an average error more than 20 min [4].

Time sequential methods. Khattak et al. developed a time sequential model by identifying ten distinct stages of the incident duration based on the availability of information [5]. Each stage had a separate truncated regression model and the model progressively added more variables. The purpose of the study is to demonstrate the methodology rather than show its performance.

Conditional probability analyses. Developing conditional probability is another use of probability in incident duration. Traffic management agencies might be interested in knowing cases like the probability of an incident lasting 30 minutes given that it has been already been active for 15 minutes. To find answers for such situations, Nam and Mannering used hazard based models (using conditional probabilities to find the likelihood that an incident will end in next short time period given its continuing duration) to develop the incident duration model [6].

Moreover, incident duration models based on probabilistic distribution analyses ([7], [8]), support vector regression [9], discrete choice models [10], Fuzzy logic models [11], Bayesian classifier [12], artificial neural networks[13] have also been frequently explored by the researchers. It should be noted that all of the developed models are often site/facility specific and require calibrations to be used at other locations/facilities.

Another set of research has primarily focused on examining the incident induced delay, which can be defined as the additional delay caused by the incident. Garib et al. developed linear regression models to estimate cumulative incident delay as a function of incident duration, traffic demand, and capacity reduction represented by number of lanes affected and number of vehicles involved [3]. To evaluate the operational effectiveness and performance of the freeway, Skabardonis et al. developed a methodology to estimate the incident induced delay by comparing the travel time (calculated from loop detector speeds) under incident and incident free condition [14]. The methodology was developed based on the assumption that incidents would affect the transportation system by increasing the travel time of the road users. Additional studies focused on examining the delay based on queuing theory and shock wave analysis[15]. However, there are some limitations underlying this group of studies such as queuing models require identification of capacity reduction and demand change in order to calculate the extent of delay, which is difficult to measure due to the stochastic nature of the incidents [14]. Meanwhile, the shockwave models estimate the delay based on wave speeds; as a result, the inaccuracy in estimating wave speeds might cause serious misinterpretation of the incident induced delay [16].

There are also a few studies using traffic simulation models for the analysis of the effect of traffic incidents and corresponding incident management strategies. Cragg and Demestsky developed a CORSIM simulation model to assess incident impact and traffic diversion strategies on freeway [17]. Zhang et al. used TSIS simulation to predict delays due to incidents on freeway [18]. Recently, a legislation has been passed in South Carolina regarding quick clearance criteria in an incident site. Fries et al. assessed the impact of quick clearance criteria deployed in South Carolina using Paramics based simulation [19]. Kabit et al. developed a VISSIM simulation model to quantify the impacts of major traffic incidents and estimate their associated cost [20]. These studies demonstrate promising results of using simulation based approaches to determine the impact of incidents. However, the needs for detailed traffic and incident data and careful calibration of simulation models limited their uses for large-scale analysis.

In addition to the congestion delay, transportation researchers have recently turned their attention toward travel time reliability. Using traffic crash and empirical traffic flow data collected from the Netherlands, Tu et al. presented an empirical travel time reliability analysis [21]. One limitation in their research was that the duration and severity of each accident were unknown, so they assumed each accident had a duration of three hours. Yu et al. used reliability analysis to assess freeway crash risks and to evaluate hazardous freeway segments [22]. The analysis integrated traffic flow parameters and real-time crash occurrence risk at the disaggregate level with weather parameters. The authors found the method provided more accurate crash predictions than that based on the logistic regression [22]. Zhong et al. used data on rural roads in Wyoming to model and predict crashes [23]. The data included accident records, traffic volume, speed, and other factors from 36 roads over a 10-year period. Negative binomial regression and Poisson regression were used to examine the causes of rural crashes. Multiple regression approaches have been used to analyze the relationship between crash rates and geometric roadway features, but multiple studies have found linear regressions are unsuitable ([24], [25]). [23] demonstrated that roads with higher speeds and traffic volumes elevated crash rates at certain higher risk locations. Wright et al. showed that incidents produced higher values in all reliability measures [26]. They also examined how incidents affect the probability of traffic congestion on freeway segments. Compared to the normal condition, they found that shoulder incidents significantly increased the probability of traffic breakdown, while incidents spreading across multiple lanes resulted in the most significant increases in travel time variability.

With the advancement of technologies, the traffic monitoring system is also equipped with modern sensory devices, which generate and archive large amount of data on a daily basis.

This data is continuous and captures the dynamics of traffic at monitored segments of the highway network. Many studies have begun to investigate those archived traffic sensor data sets and incident logs to quantify the impact of incident. Chung and Recker utilized the loop detector data to identify incident induced congestion [27]. They applied the integer programming technique to identify temporal and spatial extent of the region of the congestion caused by an accident and then estimate associated delay. Pan et al. also investigated similar data sets to estimate the impact of incidents [28]. With the availability of massive sensor data sets and incident logs, more data driven and location specific approaches should be developed to identify the spatiotemporal extent of traffic incidents and evaluate their characteristics, which significantly influence the incident induced congestion.

Few studies have looked at the interactive effects of traffic and weather factors and roadway geometry on different crash types. Yu et al. attempted to explore the use of microscopic traffic and weather indicators to differentiate between crash types and to analyze the crash type propensity at the micro-level for three major crash types — rear-end, sideswipe, and single-vehicle crashes [29]. Ahmed et al. investigated the effect of the interaction between roadway geometric features and real-time weather and traffic data on the occurrence of crashes on a mountainous freeway [30]. They found that geometric factors were significant in all seasons. Crash likelihood could double during the snowy season due to slick pavement conditions and steep grades, which collectively produced a hazardous road surface. Hojati et al. presented a framework to exhaustively mine traffic-incident data and directed subsequent analysis toward an incident delay and travel-time reliability model [31].

Although there are several models that are highly efficient, they are not applicable to other cases as they call for the use of different variables. As such, results may not be transferable across different locations. Data collection and reporting process have also been incommensurate. While the findings of previous studies will not reduce the number of crashes/incidents, they will reduce their effects and help the traffic management center to take adequate measure to minimize the impact.

1.3 Objectives

In this study, we will develop a data-driven approach to automatically and accurately quantify the spatiotemporal impact of crashes. To construct a speed profile that is more representative of the traffic condition on the day the crash occurs, the historical trends of the traffic will be adjusted by the pre-crash traffic condition. For this purpose, we will introduce the Kalman Filter algorithm to combine current traffic and historical trends to formulate the crash free normal traffic pattern. This will allow us to capture the dynamics of the impact of the crash during both recurrent and non-recurrent congested conditions. Furthermore, the required inputs of the approach are the stationary sensor data and incident logs, which are widely available and readily obtainable for transportation agencies.

2. Data Collection

2.1 Study Area

This study collects multi-source data during the 2011-2013 period on the I-65 corridor in the Louisville metropolitan area. Figure 1 shows the spatial extent of the corridor. The corridor is 5.6 miles long in the northbound direction which starts from MP-131 and ends at MP-136.6. The southbound direction is 5 miles long which starts from MP-136 and ends at MP-131. The speed limit along both directions is 55 mph.

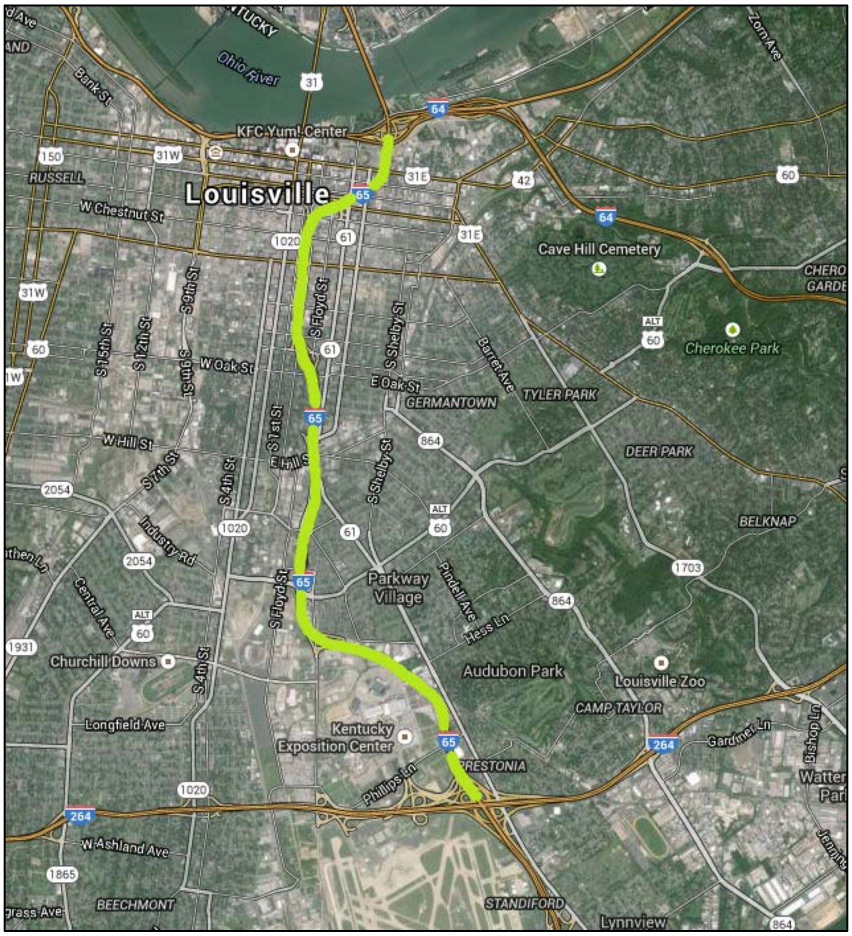


Figure 1: Study Corridor of I-65

2.2 Stationary Sensor Data

Stationary sensor data is provided by TRIMARC. TRIMARC is a regional traffic management center designed to improve the performance of the freeway system in the

Louisville metropolitan area, which extends into Southern Indiana. The TRIMARC data contains time, speed, volume, and lane occupancy information for each day. The original data is recorded at 30-second intervals and aggregated into 15-min intervals with the Kentucky Archived Data Management System. There are a total of 26 stationary sensors located on the study corridor; 15 of them are located on the I-65 N and the rest of 11 sensors are on the I-65 S. The average spacing between adjacent sensors is approximately 0.4 mile. Figure 2 illustrates the sensor locations on I-65 N.

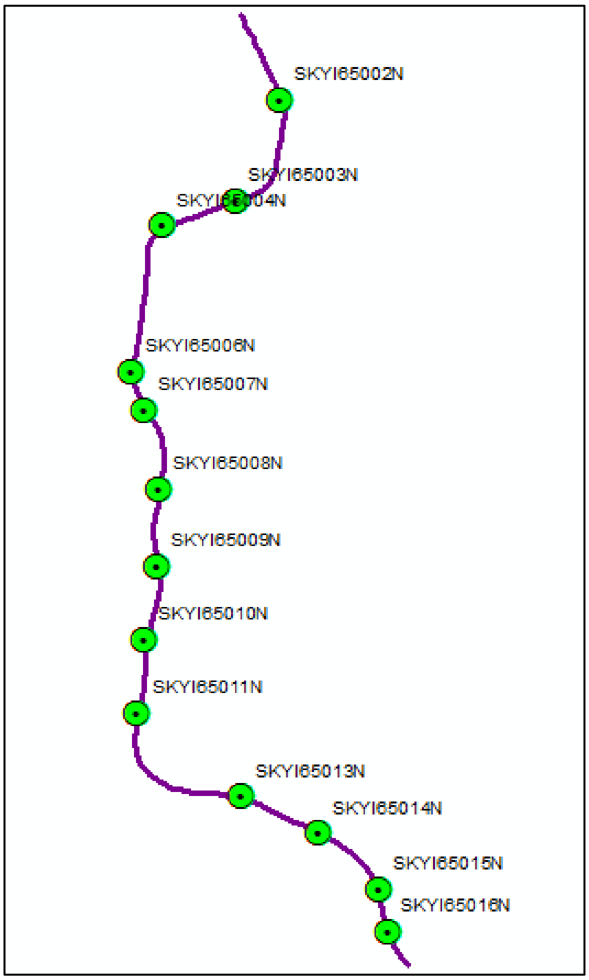


Figure 2: Location of Stationary Sensors on I-65N

2.3 Incident Data

The incident information is obtained from incident logs also provided by TRIMARC. The information available in the incident logs can be grouped into four main categories: the incident characteristics including the type of incident, the number of vehicles involved, whether injury is involved, and number of lanes blocked; the spatial characteristics such as the location of the incident, the route and direction, and the mile-marker; the temporal characteristics such as the time when the incident is notified to the TMC, the time when the incident is cleared, and total duration of the incident; and environmental characteristics including the weather conditions at the time of the incident.

2.4 Weather Data

Historical weather records are downloaded from https://www.wunderground.com/, which collects the data from weather sensors at Louisville International Airport, which is about 2 miles away from the study corridor. The data includes weather information such as temperature, wind speed, direction, visibility, weather condition, precipitation, etc. The weather data is preprocessed and integrated with the incident data to get a complete picture of the incident.

3. Identify Crash Impact Zone

In order to automatically identify the impact of the crash on the traffic flow, this study proposes a data-driven approach to analyze the stationary sensor data along with the incident log under the crash condition. The proposed methodology for identifying the crash impact zone contains five major steps:

1. Obtaining the current speed profile under crash condition
2. Identifying the crash scenario
3. Determining the background speed profile
4. Identifying the crash impact from the difference-in-speed profile
5. Determining the impacted region

3.1 Obtain Speed Profile

When a crash occurred during the uncongested condition, it is very easy and straightforward to isolate the crash impact from the normal traffic condition. However, when a crash occurred during the congested condition, it would be very difficult to separate the crash impact from the already congested traffic condition. In order to know if there is any impact of a crash, a current speed profile is first obtained for the day when the crash occurred. The current speed profile provides the ground truth measure of real time traffic conditions under the impact of both crash induced and recurrent congestion. To represent the current speed profile visually, a contour map using all current stationary sensor speed data has been developed, which is described as follows.

At first, we assume each sensor measurement represents the segment traffic condition from that sensor to the adjacent upstream stationary sensor. The current traffic speed V of the j^{th} segment at i^{th} time slice is denoted as V(i,j), where i= 1,2,3….96 (as ninety-six 15 minute time slices in a day) and j= 1,2,3…….s (s is the total number of sensors). Next, the current traffic speed measurements are coded with a continuous color scheme with red representing low speeds while green representing high speeds. The obtained contour map can also be called space time velocity map or heat map. The heat map (as shown in Figure 3) increases the visual understanding of the spatiotemporal pattern of speeds with the congested area highlighted by red.

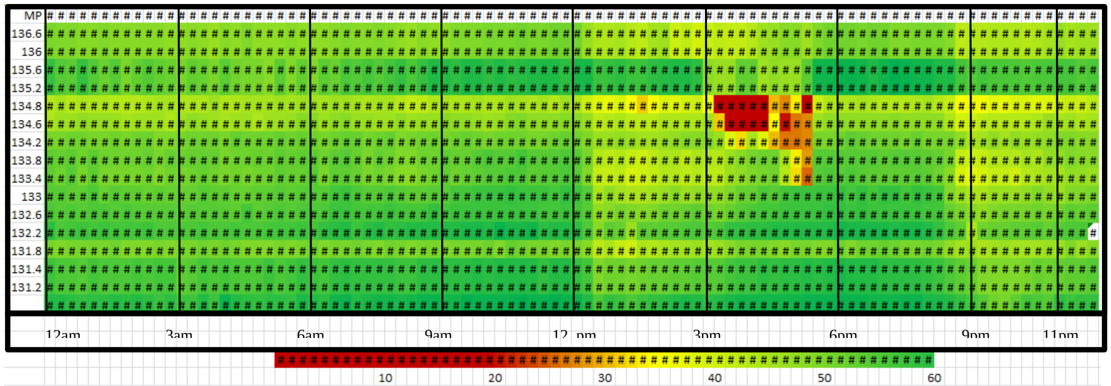


Figure 3: Heat Map

3.2 Identify Crash Scenario

In this step, the methodology will identify the crash scenarios, which can be divided into three categories:

- Type-1: Crash Induced Congestion
- Type-2: Crash without Congestion
- Type-3: Congestion Induced Crash

The crash induced congestion or Type-1 crash could be defined as the scenario when the congestion occurred as a result of the crash. Crash without congestion is the scenario when there was no congestion after the crash, which is defined in this study as a Type-2 crash. Finally, congestion induced crash is the scenario when the roadway was congested before the crash occurred, which is defined in this study as a Type-3 crash.

A search technique has been developed to identify these three types of crash scenarios. The technique searches whether there is an existence of congestion before and after the crash occurrence, knowing the start time and location of the crash from the incident log. The searching process has two windows:

- Time window
- Space window

The time window consists of four 15-min time slices. If the crash start time is T, then the four time slices will be T-1, T, T+1, T+2. The space window consists of two immediate sensors, denoted by u_1 and u_2, upstream of the crash location. The combination of each time slice and sensor is considered as one cell. Thus, real time speeds at a total eight cells are checked to see if there is any congestion or not. The congestion here is determined by the following criteria. In other words, it would be considered as congested if the current speed is below 75% of the posted speed limit.

$$\text{Congested if } \frac{\textit{Speed limit} - \textit{Current Speed}}{\textit{Speed limit}} > 25\%$$

All the time slices at u_1 will be first analyzed, followed by all the time slices at u_2 if no congestion is found at u_1.

- If congestion is found at the $(T\text{-}1)^{th}$ time period, it would be considered a Type-3 crash
- If congestion is found at the T^{th}, $(T+1)^{th}$, or $(T+2)^{th}$ time period, it would be considered a Type-1 crash
- If congestion is not found at any time slice at any location, it would be considered a Type-2 crash.

Figure 4 shows the different types of crash scenarios. Star mark represents the crash start time and the actual location of the crash. Figure 4(a), (b) & (c) represent three cases of the Type-1 crash, where the impact of the crash is present at the time of the crash, with a small time lag and at the upstream segment with a small time lag, respectively. Figure 4(d) represents the Type-2 crash, where no congestion is present after the crash. Figure 4(e) & (f) represent two cases of the Type-3 crash, where the crash happens in the middle of the recurrent congestion. With this categorization, we can identify the starting time and the location of the crash impact.

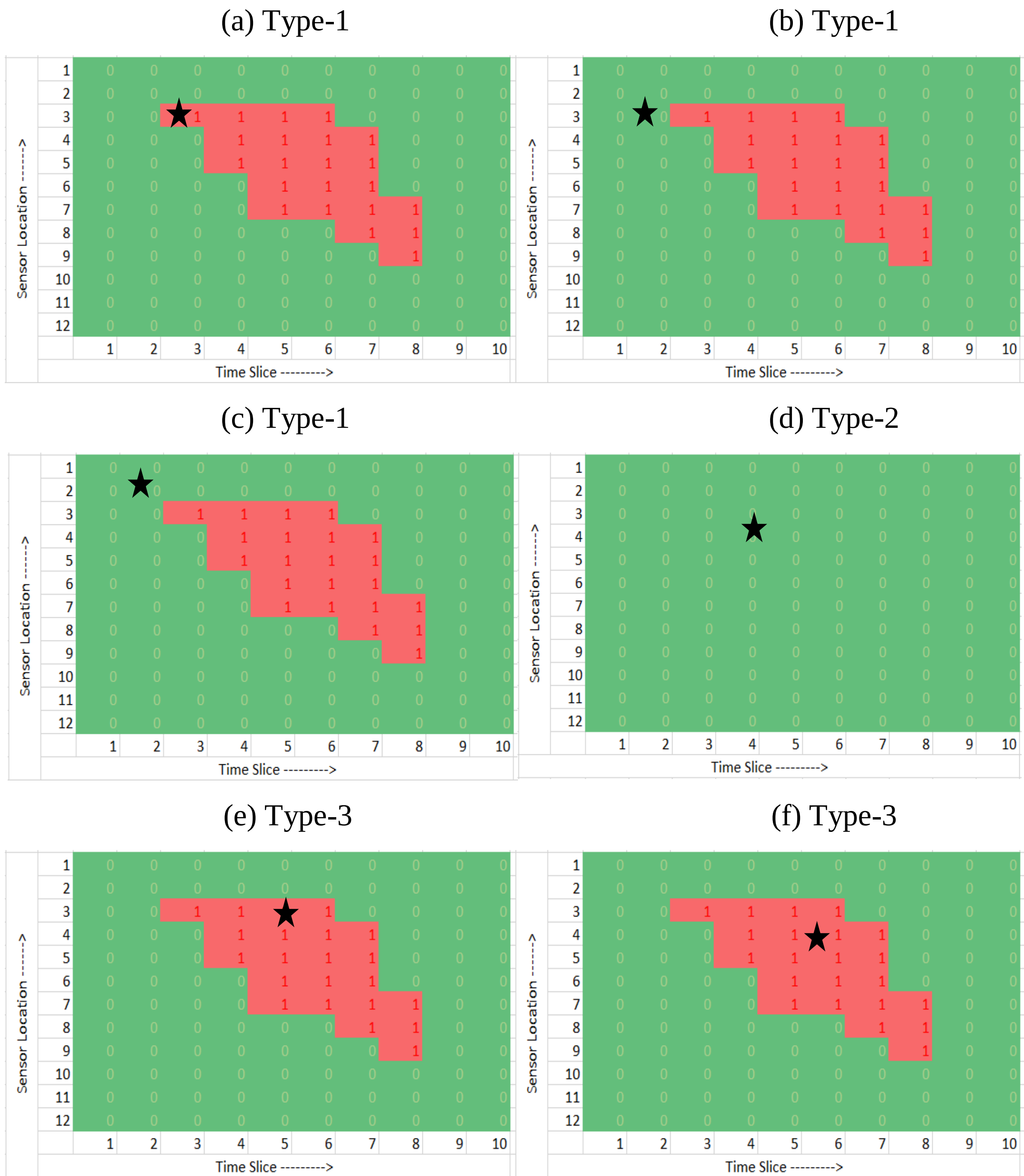


Figure 4: Crash Scenarios

3.3 Determine Background Speed Profile

A background speed profile is a profile that reflects the expected traffic condition throughout the day for a specific location. Before identifying the impact of the crash, it is very important to know the normal traffic/speed condition of that location. As we know, the recurrent congestion is the traveler's expected traffic condition as long as it occurs periodically. On the other hand, the non-recurrent congestion is often caused by unexpected incidents or inclement weather condition. Therefore, the background speed profile should reflect both recurrent congestion and free flow traffic condition as expected by the daily road users. That is why it should be constructed based on the traffic data that shares a similar traffic pattern such as at the same period of time, during weekend vs weekday etc.

Instead of using a fixed background speed profile derived from the historical data, in this research a dynamic profile is introduced under the notion that the pre-crash condition of a specific day may be different from an "average" day. To capture this variation, the dynamic background profile is created by combining both the real-time traffic condition before the crash occurs and the historical trends. To achieve this purpose, the Kalman Filter (KF) algorithm is adopted to predict the dynamic background speed profile. the KF is a powerful mathematical tool that can estimate the future states of the variables even without knowing the precise nature of the system modeled [32]. It is a recursive procedure that corrects its estimates whenever new observations become available, with the objective of minimizing the estimated error covariance. the KF has been used widely in various fields of transportation such as forecasting traffic parameters, predicting bus arrival time, to name a few[33]. As our main interest of this study is to identify the individual impact of each crash, the KF starts creating a background speed profile instantly when the crash occurs.

The KF based procedure developed in this study is designed to predict the crash-free normal speed profile based on both historical profile and the pre-crash condition. The historical average speed as well as the variance of speeds at each sensor location for every 15 min interval are used as the inputs of the state predictors in Kalman filter. The procedure predicts the speed of the next time slice with the knowledge of the speed of the previous time slice. The whole process of generating the background speed profile is explained as below.

Now assume, k denotes the pre-crash time slice, k+1 denotes the time slice at crash moment. The term x_k is the historical average speed for the particular sensor, A_k is the ratio of historical average speed of $k+1^{th}$ and k^{th} time slice. r_k is the real time speed

State Prediction : $\hat{X}_{k+1} = A_k x_k + w_k$

Observed Speed : $Z_k = \beta_k r_k + v_k$,

where,

A_k = state transition model which is applied to the previous state

β_k= 1

w_k = white noise associated with the transition process which is assumed to have zero mean and variances of Q_k

v_k = observation noise which is assumed to have zero mean and variances of R_k

The overall filtering process is the recursive prediction process. At the moment crash occurs, the formulation of background speed profile will be started using the following process:

- Step 1 : Initialize
  Set k = (T-1); T= Accident start time
- Step 2 : Initialize Observed Speed, Z_k
- Step 3 : Initialize Covariance P_k
- Step 4 : Extrapolate state variable.
  $$\hat{X}_{k+1} = A_k x_k$$
- Step 5 : Extrapolate Covariance
  $$\hat{P}_{k+1} = A_k P_k A_k^T + Q_k$$
- Step 6 : Compute Kalman Gain
  $$K_{k+1} = \hat{P}_{k+1} * \beta_k (\beta_k \hat{P}_{k+1} \beta_k^T + R_k)^{-1}$$
- Step 7 : Update State variable
  $$X_{k+1} = \hat{X}_{k+1} + K_{k+1}(Z_k - \beta_k \hat{X}_{k+1})$$
  Stop if k+1=96, otherwise go to step 8
- Step 8 : Update Covariance
  $$P_{k+1} = (1 - \beta_k K_{k+1}) \hat{P}_{k+1}$$
- Step 9 : Update Observed Speed
  $$Z_{k+1} = \beta_{k+1} X_{k+1}$$
- Step 10 : Update Time slice
  k=k+1

  go to Step 2.

The KF starts with taking the pre-crash speed as an input to predict the speed of the next time period and this process will continue for all of the upstream sensors and create a background speed profile at each sensor location of the whole corridor. We have used the predicted speed of one time slice as the pseudo-observed speed (Step-9) for the prediction of the next time slice. The main motivation to propose this strategy is the fact that our key objective is to get the normal traffic speed pattern to capture the special events (such as crash) from daily traffic. The method captures the normal traffic speed pattern effectively.

The following figure exemplifies the superiority of using a KF-based dynamic profile over the fixed one. The first part of Figure 5 shows on a specific day, the pre-crash speed is higher than the historical average speed, so the obtained background speed profile starts above the historical trend. On the other hand, the second part presents the case where on a particular day the pre-crash speed is lower than the historical average speed, so the background speed profile is initiated with a speed lower than the historical trend.

During the calculation of the historical average speed at different time slices, we have found that the traffic/speed patterns during the weekdays and weekends are significantly differen, as shown in Figure 6. So the average speed is calculated separately for weekdays and weekends. If a crash occurs on a weekday, the weekday average speed at that location will be chosen as the input of the background speed profile. Likewise, for weekend crashes, the average speed from weekends will be used as the input.

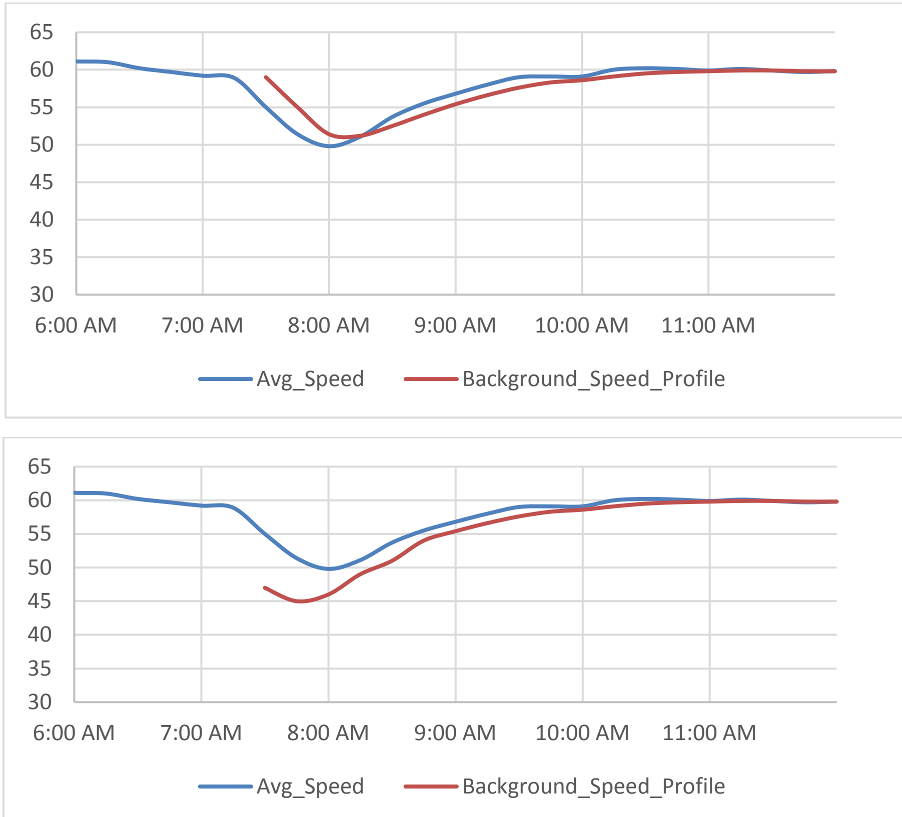


Figure 5: Capturing Dynamics of Traffic by the Background Speed Profile

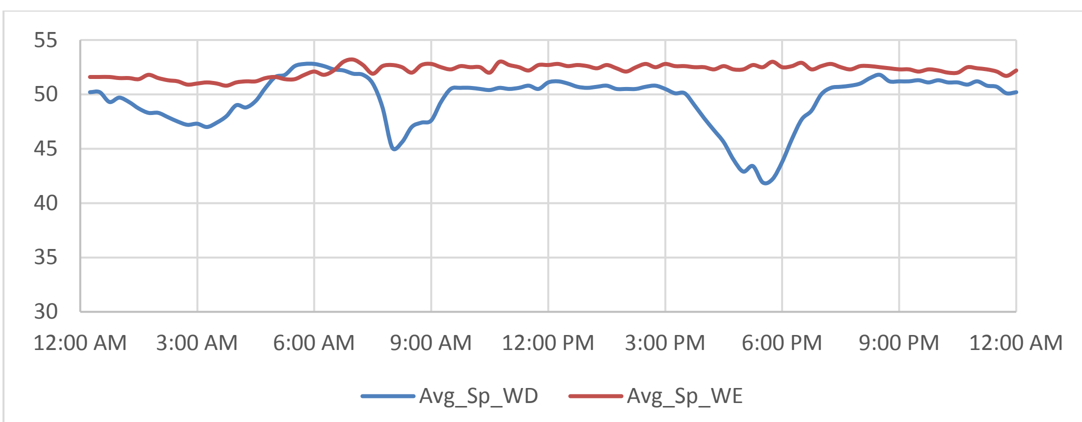


Figure 6: Average Speed Comparison- Weekday vs Weekend

3.4 Identify the Crash Impact

By superimposing the current speed profile over the background speed profile, the difference in speed profile can be established and used as the basis for estimating the crash impact. Figure 7 shows the process of separating the crash impact from the background speed profile.

The start and end time of the crash impact could be identified visually from the difference in speed profile. In order to automate the process the following formulations are used.

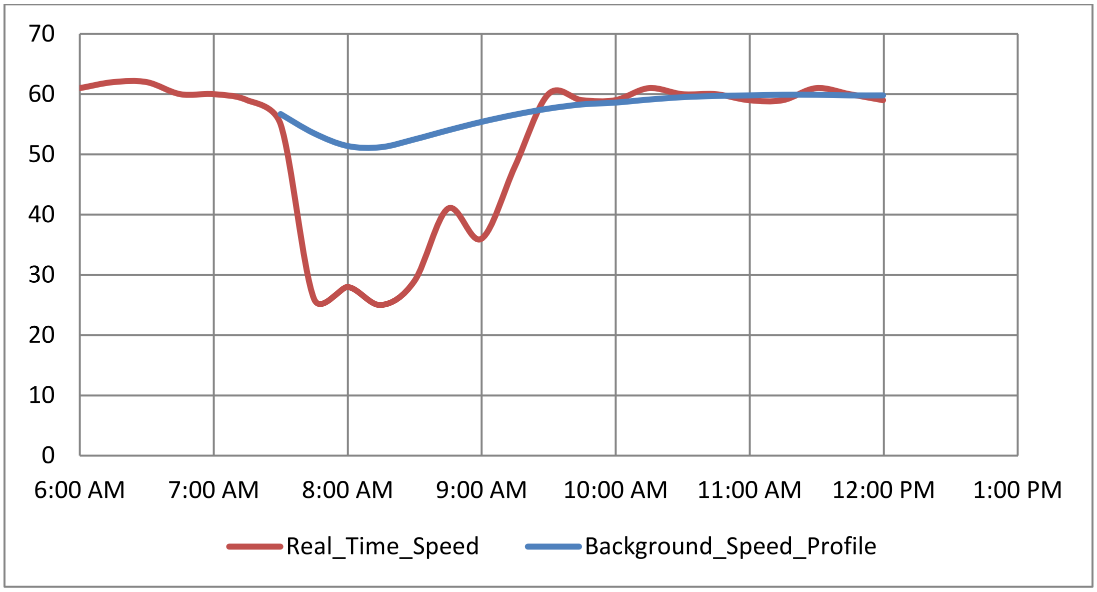


Figure 7: Real Time vs Background Speed Profile

If BSP<CSP,

DS= 0

If BSP>CSP and CSP< (75%*SL)

DS = BSP-CSP

If BSP>CSP and (75%*SL) < CSP < SL

DS = BSP-CSP, when DS ≥5 otherwise DS=0

If BSP>CSP and SL< CSP

DS = BSP-CSP, when DS ≥10, otherwise DS=0

where,

BSP = Background Speed Profile (mph)

CSP = Current Speed Profile (mph)

DS = Difference in Speed (mph)

SL = Speed Limit (mph)

The above process will continue from the crash start time to the remaining portion of the day for all the sensors upstream of the crash location. In this study, all DS values are not taken as the crash impact. Instead, a filtering process is introduced to filter the noise. In this way, the process adds an empirical tolerance value that specifies a least DS value at different level of current speeds, which must be achieved by DS to be considered as an impact of the crash. To visually represent this impact, a contour map has been created, where Y-axis represents the location of different sensors and X-axis represents different time periods. Now this map will show whether the segment is congested during the crash time. If DS is greater than zero, then it is assumed that the corresponding segment is congested. That is why the selection of DS is very important and a filtering process is used to control the noise.

Figure 8 shows an example of the contour map showing the difference in speed profile. This map identifies the area where the crash shockwave propagates. It also captures congestion that may be caused by other events that are not relevant to the current crash. Those external events will be excluded in the next step.

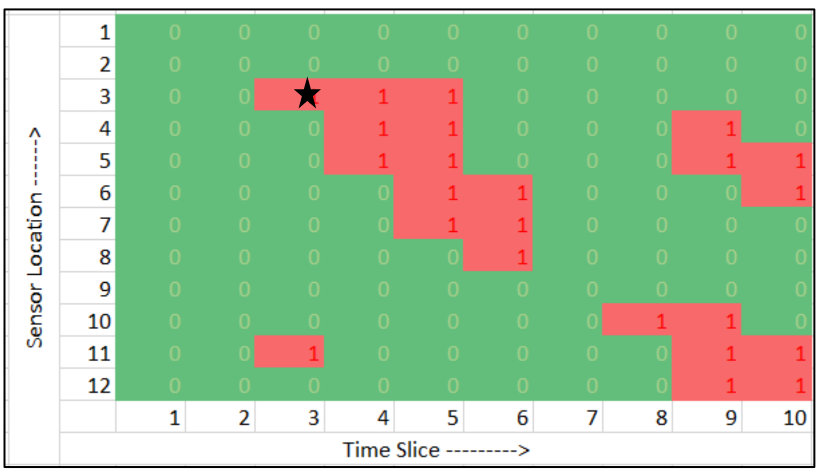


Figure 8: Contour Map of Difference in Speed Profile

3.5 Determine the Impacted Region

The final step is to identify the boundary of the spatiotemporal impact of the crash. The following conditions are considered to define the final boundary.

- After the crash, the spatiotemporal progression of the shockwave must be uninterrupted.
- The spatiotemporal boundary of the crash shockwave progression must be at upstream.

- Entire boundary of the impacted region must be contiguous

Fulfilling these conditions, the final boundary can be determined. Figure 9 shows some examples of the infeasible shapes. Figure 9 (a) represents the irregular progression of the crash shockwave which is a clear violation of our assumption. According to our consideration the spatiotemporal progression of the shockwave must be uninterrupted and it should advance at upstream in a cascading format. So the marked irregular portions should be ignored in the final region. Also, there should not be any holes as shown in Figure 9 (b) in the impact region as it contradicts the uninterrupted progression concept. These holes are created when current speed profile crosses the threshold of the background speed profile towards no-congested stage for a few moments, then again returns to the congested condition. This occurs due to the highly stochastic nature of the traffic. So when such hole is found that is surrounded by the crash impacted region, then the hole is considered as a part of the final impacted region. This inclusion will not affect the calculation of total delay due to crash. Figure 9 (c) & (d) are the examples of violating the assumption that the entire boundary of the impacted region must be contiguous. Here we observe the presence of some external events which are not related to the crash, those events are excluded from the final impact region.

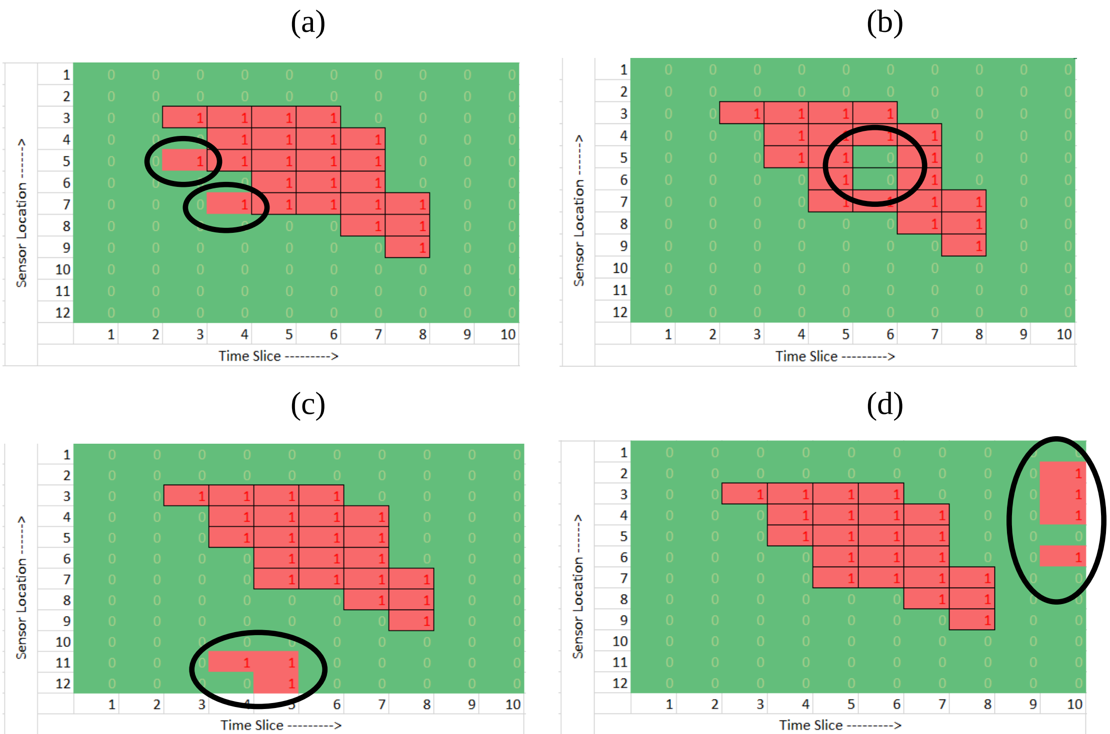


Figure 9: Infeasible Shapes

After resolving the infeasible shape, we have found the final boundary of the crash impact zone, as shown in Figure 10. From the final boundary, the temporal and spatial length of the crash impact can be accordingly determined knowing the horizontal and vertical length of the boundary respectively. Spatial length represents how far a crash shockwave spreads and temporal length represents how much time a crash disturbs the normal flow of traffic. As each cell represents 15-min time period, the calculated impact duration would be multiples of 15 minutes. Since the spatial aspect of the crash impact is an important issue, the proposed method provides a way to address it. Since the spatial length shows the number of upstream sensors affected by the crash and we know the distance between two sensors is approximately 0.4 miles, the spatial extent of the crash can be calculated from this information.

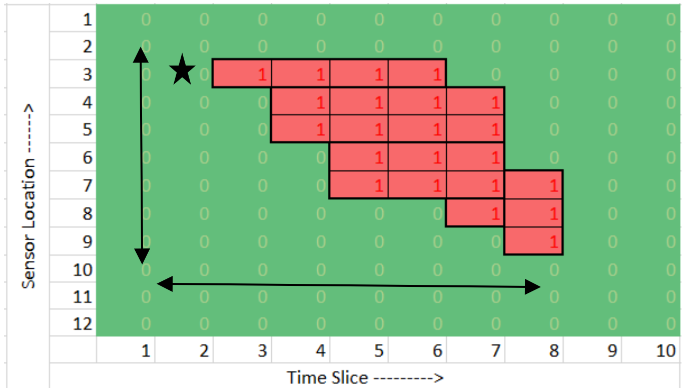


Figure 10: Final Impact Zone

The success in identifying the crash impact in the dynamic traffic environment depends on an accurate representation of background crash-free traffic condition at the time of the crash. The proposed method handles this issue by using the KF algorithm and developing the expected normal condition that travelers would anticipate by combining both the ongoing traffic condition and the historical trends. Instead of using a fixed background profile, here we are able to obtain a dynamic background profile using the pre-crash speed as inputs to capture the dynamics of daily traffic.

4. Estimate Duration of Crash Impact

One of the main objectives of the TIM program is to reduce the impact of the incident/crash on normal traffic flow. The most effective way to do this is to clear the incident scene as quickly as possible. The ability of quickly estimating the impact duration can help highway authorities effectively allocate their emergency resources to minimize the negative effect of incidents. Additionally, documenting the impact end time and understanding their properties will allow better crash management strategies in the future. That is why in this study an investigation is conducted to understand the properties of the impact duration and factors affecting them.

The impact duration of a crash can be defined as the time elapsed from the beginning of the disruption of the normal traffic condition since the occurrence of the crash to the time when the normal traffic condition is resumed. The total impact duration can be divided into following subcomponents:

- Detection time (time required to detect the presence of a crash)
- Response time (time from the incident response team being notified of the crash to their arrival at the crash scene after being notified)
- Clearance time (time required to clear the crash)
- Traffic recovery time (time required to resume the normal traffic condition after the crash being cleared)

In our study, the impact durations are only calculated for Type-1 and Type-3 crashes. Type-2 crashes are not included in the impact duration model since they have no impact on the traffic. There is a wide range of methods that could be used to predict the impact duration. In this study three methods are explored for the impact duration prediction:

- Multiple linear Regression
- Logistic Regression
- Quantile Regression

4.1 Multiple Regression Model

The first impact duration prediction model is developed based on widely used multiple linear regression and impact duration dataset. Through the trial and error process, the following impact duration prediction model with the best goodness of fit is obtained:

$$\ln(Y) = 4.928 - 0.029X_1 - 0.119X_2$$

Alternatively, the model can be reformulated as:

$$Y = e^{(4.928 - 0.029X_1 - 0.119X_2)},$$

where:

Y = Impact duration in minutes;

X_1= Post crash space mean speed on the first upstream segment with unit of mph;

X_2= Binary variable for weather condition such as rain (Yes =0, No=1).

The model can predict 32% (R-squared) variation of the impact duration in a natural logarithmic format as a function of two independent variables (post-crash speed and weather condition). Other variables tested here are found to be insignificant. It is worth mentioning that the p-value for the corresponding F-statistics of the overall model is less than 0.05 (F-statistics = 90.37) and the p-value for the parameter of each independent variable is less than 0.05. Accordingly, there is enough statistical evidence to support the linear relationship between the natural logarithm of impact duration and each independent variable.

The final dependent variable takes the natural logarithmic format because the normal probability plot indicates that the impact duration does not follow the normal distribution. An effort has been made to transform the duration into different functions, but none of them passes the normality test. Of those transformed forms, the natural log format shows a better result than others. Since the model will be used as point estimators for the impact duration while not for determining confidence intervals, the impact of non-normality would not be that significant [34].

Based on the final model, the coefficients of the independent variables are negative. This indicates that the impact duration is negatively correlated with other two independent variables. More specifically, the duration is expected to be shorter with higher post-crash speed. The result also indicates that the crash impact is more severe under the rainy condition than that under the non-rainy condition. Other variables, including lane blockage, injury and other factors are found to be insignificant.

Predicting the impact duration rather than the crash duration can help provide more complete information that is pertinent to travelers' concerns. The developed procedure offers a new way to more realistically predict the impact duration for crashes, instead of relying on the given incident duration in incident logs. The regression model is relatively simple and not data intensive; therefore, it can be quickly deployed by transportation agencies in the real-time application. The predicted impact duration can also help the freeway management authority have an idea for what control strategies should be implemented to minimize the traffic congestion and thus improve the freeway performance.

The R-squared value (0.32) for this multiple linear regression model is not very high. One possible reason would be that the impact duration is measured as a multiple of 15 minutes, but predicted duration is in continuous number format which increases the difference between predicted and observed values.

4.2 Logistic Regression Model

Next, we proceed to using the logistic regression for impact duration modeling. At first each independent variable is fitted with the response variable to check its significance. It is found that the post-crash space mean speed of the first upstream segment and the weather condition are significant in predicting the incident duration, which is similar to the finding of the multiple linear regression.

The results of the logistic regression are summarized in the Table 1, which consists of the regression constants and the regression coefficients of the predictor variables (Post-crash speed, weather condition). The regression constants corresponding to the different response categories (θ_1, θ_2, θ_3, θ_4, θ_5, θ_6) are calculated by assuming default factor levels for all the predictor variables. The associated p-values for the regression constants and coefficients are also provided. For ascertaining the significance of the predictor variables, a 5% significance level is assumed throughout this research. Hence a particular predictor variable would be considered statistically significant, if the corresponding p-value is less than 0.05.

Table 1: Logistic Regression Results

Predictor Variable	Factors	Regression-Coefficient(β)	p-value
Const(A) -- θ_1		-5.084	<0.0001*
Const(B) -- θ_2		-3.814	<0.0001*
Const(C) -- θ_3		-2.26	<0.0001*
Const(D) -- θ_4		-1.21	<0.0001*
Const(E) -- θ_5		-0.484	0.0232*
Const(F) -- θ_6		0.605	0.0129*
Post-crash Speed		0.0749	<0.0001*
Weather Condition	Default: No Rain		n/a
	Rain (Y)	-0.3363	0.0011*

Based on the regression constants and the coefficients obtained by the regression analysis, the probabilities of different responses (impact duration class) can be calculated for different predictor variable scenarios. For each particular predictor variable situation, the response probabilities can be calculated by using the corresponding regression constant (θ, depending on the response probability) and appropriate regression coefficients (β, depending on the predictor variable).

The effect of each predictor variable on the impact duration can be determined by changing the regression coefficient of that variable only while holding all the other variable coefficients unchanged. A positive regression coefficient indicates a reduction in the impact duration due to the corresponding factor. On the other hand, a negative regression coefficient indicates an increase in the impact duration time. Post-crash speed has a positive coefficient, meaning higher post-crash speeds correspond to shorter impact duration. On the other hand, weather condition has a negative coefficient which indicates impact duration will increase comparing to the default no rain condition.

An example of calculating the probability of each response of the impact duration based on logistic regression results is presented in the following section. According to logistic regression result, we have found two significant independent variables. Between them, one variable is categorical (weather condition) and the other is continuous (post-crash speed) in nature. Since a continuous variable is present in the model, there is no absolute base condition for this model. Absolute base condition is achieved when all independent variables become categorical. Therefore, the response probabilities are calculated assuming a random speed value and weather condition (rain or no rain).

Now sample calculation is shown for speed with 10 mph and no rain condition. Based on the logistic regression result, the probability of the impact duration falling in the category A(0-15 min) is

$$\frac{e^{\theta_1+\beta x\prime}}{1+e^{\theta_1+\beta x\prime}} = \frac{e^{-5.084+0.0749*10}}{1+e^{-5.084+0.0749*10}} = 0.01$$

The probability of impact duration being B(16-30min) can be calculated as

$$\frac{e^{\theta_2+\beta x\prime}}{1+e^{\theta_2+\beta x\prime}} - \frac{e^{\theta_1+\beta x'}}{1+e^{\theta_1+\beta x'}} = \frac{e^{-3.814+0.0749*10}}{1+e^{-3.814+0.0749*10}} - \frac{e^{-5.084+0.0749*10}}{1+e^{-5.084+0.0749*10}} = 0.03$$

Similarly, we can know the probability of impact duration being C(31-60 min), D(61-90 min), E(91-120 min), F(121-180 min), and G(>180 min) is 0.14, 0.21, 0.18, 0.23, and 0.2, respectively.

Now if the speed is still 10 mph but the weather condition is rainy, the probability of the impact duration falling in the category A(0-15 min) becomes

$$\frac{e^{\theta_1+\beta x\prime}}{1+e^{\theta_1+\beta x\prime}} = \frac{e^{-5.084+0.0749*10-0.3363}}{1+e^{-5.084+0.0749*10-0.3363}} = 0.009$$

The probability of impact duration being B(16-30 min) now is

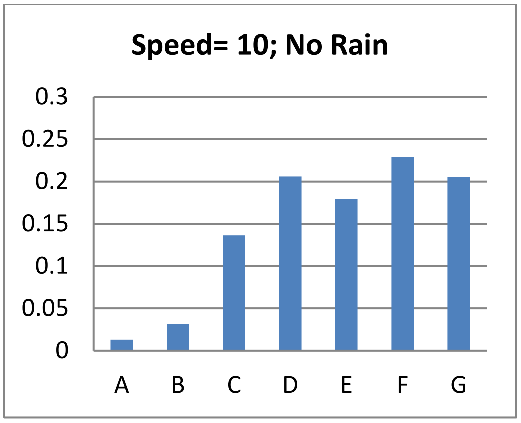
$$\frac{e^{\theta_2+\beta x\prime}}{1+e^{\theta_2+\beta x\prime}}-\frac{e^{\theta_1+\beta x'}}{1+e^{\theta_1+\beta x'}}=\frac{e^{-3.814+0.0749*10-0.3363}}{1+e^{-3.814+0.0749*10-0.3363}}-\frac{e^{-5.084+0.0749*10-0.3363}}{1+e^{-5.084+0.0749*10-0.3363}}$$
$$=0.02$$

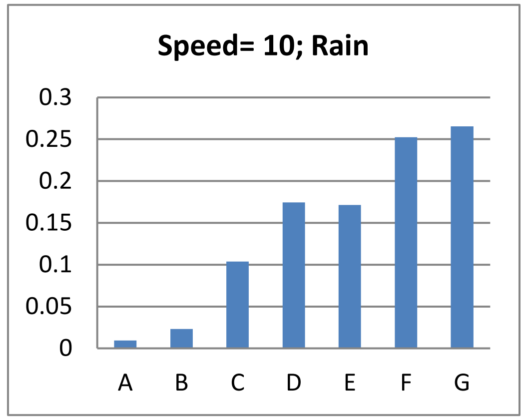
In the meantime, the probability of impact duration being C(31-60 min), D(61-90 min), E(91-120 min), F(121-180 min), and G(>180 min) is 0.1, 0.17, 0.17, 0.25, and 0.26.

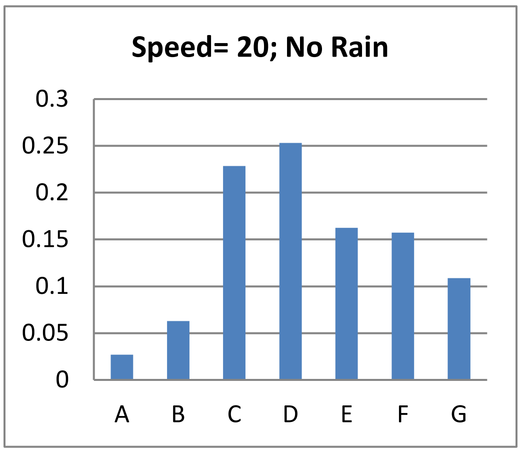
Based on the above probability values, we can see that at 10 mph post-crash speed and under non-rainy condition, the impact duration has the highest probability to be within the "F" class with the impact duration lasting between 120 and 180 minutes. On the other hand, at the same speed level with the rainy condition, the impact duration has the highest probability to lie in the "G" class which means the impact would last more than 180 minutes. As a result, the impact of rain on the crash impact duration is apparent.

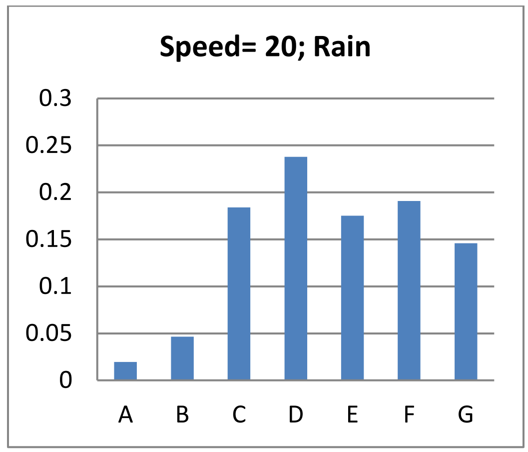
To further investigate the impact of post-crash speeds, individual probability of responses is calculated at different post-crash speed level (10,20,30,40,50,60 mph) for both rainy and no rain condition. the results are presented in Figure 11.

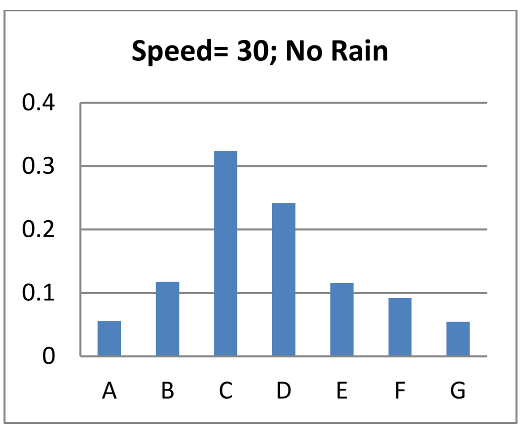
From Figure 11, it can be observed that the probability of lower impact duration would increase for both rainy and non-rainy conditions as post-crash speeds increase. It should be noted that the proposed model provides a framework to make reasonable judgement about the impact duration due to crash. Moreover, the model can predict the impact duration by using the post-crash speed at the upstream segment and the weather condition (rainy or not), both of which can be obtained easily from real-time sensor data and weather information. The information derived from the model would be very helpful for improving freeway incident management and decision making.

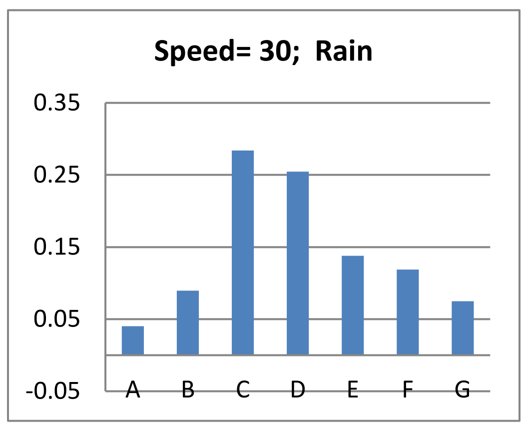

Speed= 10; No Rain
0.3
0.25
0.2
0.15
0.1
0.05
0
A B C D E F G


Speed= 10; Rain
0.3
0.25
0.2
0.15
0.1
0.05
0
A B C D E F G


Speed= 20; No Rain
0.3
0.25
0.2
0.15
0.1
0.05
0
A B C D E F G


Speed= 20; Rain
0.3
0.25
0.2
0.15
0.1
0.05
0
A B C D E F G


Speed= 30; No Rain
0.4
0.3
0.2
0.1
0
A B C D E F G


Speed= 30; Rain
0.35
0.25
0.15
0.05
-0.05
A B C D E F G

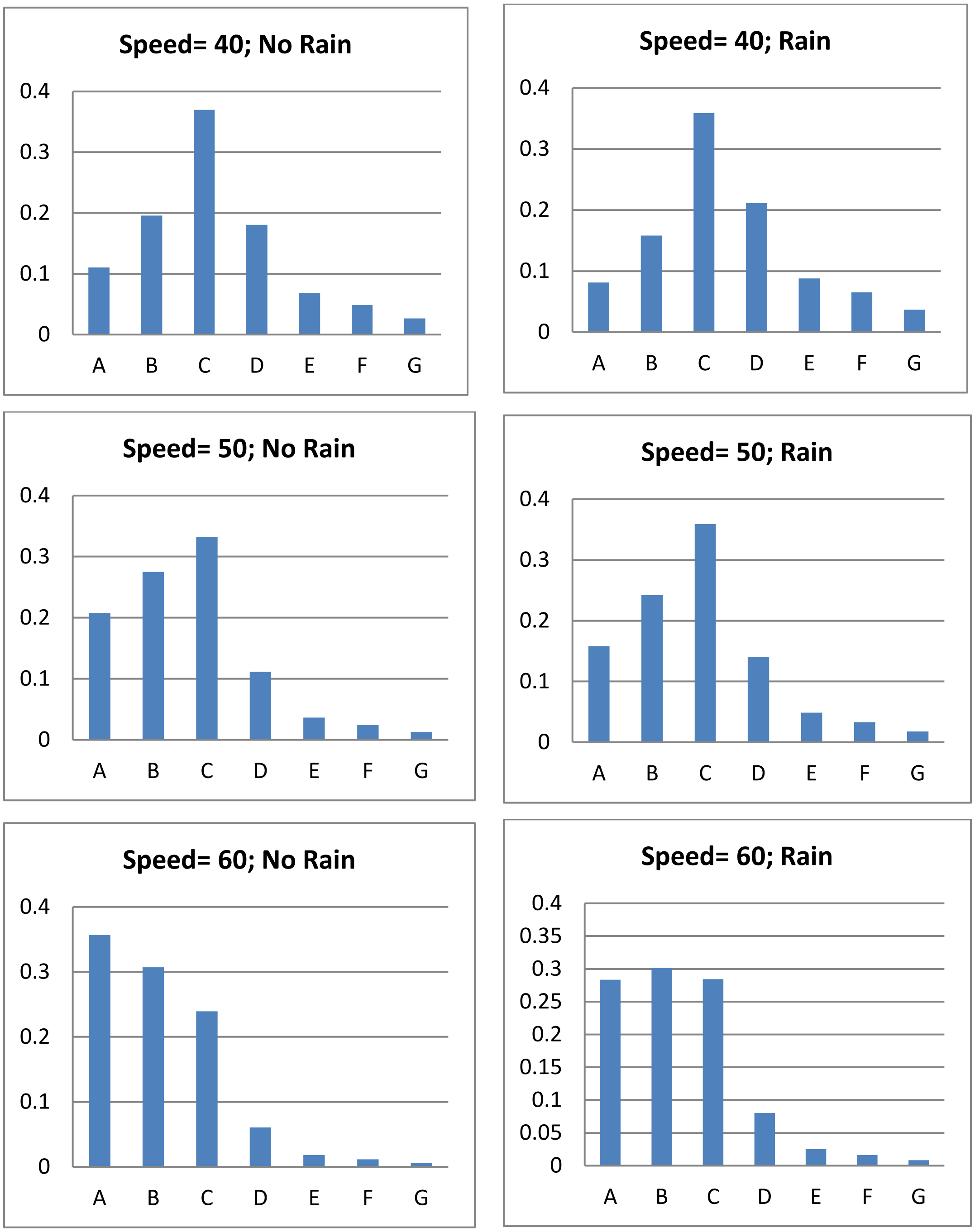


Figure 11: Response Probabilities for Impact Duration at Different Speed and Weather Conditions

4.3 Quantile Regression Model

Quantile regression is a way of estimating functional relationships between the variables for all portions of a probability distribution [35]. It estimates the conditional quantiles of a dependent variable distribution in the linear model that provides a more complete picture of causal relationships between variables. The advantage of quantile regression compared to the ordinary least square (OLS) regression is that the quantile regression estimates are more robust against the outliers.

Now suppose for a random variable Y, the cumulative distribution function is F(y), where $(y) = P(Y \leq y)$, then the τ^{th} quantile or percentile function of Y would be $Q(\tau) = F^{-1}(\tau)$, where $0 \leq \tau \leq 1$.

Q(τ) can be mathematically formulated as $Q(\tau) = \beta_0(\tau) + X_1\beta_1(\tau) + X_2\beta_2(\tau) \ldots. +X_n\beta_n(\tau)$, where $\beta_0(\tau)$ is the intercept and $\beta_1(\tau)$, $\beta_2(\tau)$...... $\beta_n(\tau)$ represent the coefficients of the explanatory variables at the τ^{th} quantile or percentile. Then, the equation can be estimated by solving the following minimization problem:

$\arg min \sum_{i=1}^{m} \rho_\tau \{y_i - (\beta_0(\tau) + X_1\beta_1(\tau) + X_2\beta_2(\tau) \ldots. +X_n\beta_n(\tau))\}$,

where ρ_τ is the loss function and can be defined by

$$\rho_\tau(\alpha) = \begin{cases} (\tau - 1).\alpha\ ; & \alpha < 0 \\ \tau.\alpha\ ; & \alpha \geq 0 \end{cases}$$

where $\alpha = y_i - (\beta_0(\tau) + X_1\beta_1(\tau) + X_2\beta_2(\tau) \ldots. +X_n\beta_n(\tau))$

Above equations can be reformulated into a standard linear programming problem, which can be easily solved with the simplex method.

Using the above described methodology, a quantile regression model is developed to predict the duration of the impact of a crash. From the experience of OLS method, it has been recognized that the resulting estimates of various effects on the conditional mean of impact duration were not indicative of the size and nature of these effects on the upper tail of the impact duration distribution. A more complete picture of the variable effects can be presented by estimating a family of quantile regression functions.

The impact of different significant independent variables on the whole distribution of the duration is shown in Figure 12. The x-axis represents the percentiles of interest, ranging from the 5^{th} to the 95^{th} percentile. The y-axis represents the independent variable effect in minutes. The solid red line represents the conditional mean outputted by the OLS method, while the dashed red lines show the conventional 95 percent confidence interval of the mean. Meanwhile, the dash dotted black line represents the percentile values and the shaded gray area shows a 95 percent pointwise confidence band for the quantile regression estimates.

It is found that three independent variables, including the space mean speed of second upstream sensor at the time when crash occurs, injury, and weather condition, significantly affect the impact duration. Figure 12 shows the summary of the quantile regression results where we have three independent variables and an intercept. For each of the four coefficients, we plot 19 distinct quartile regression estimates for τ ranging from the 5^{th} to the 95^{th} percentile. These point estimates can be interpreted as the impact of one unit change of the variable on the duration, as other variables remain unchanged.

The intercept of the model is the estimated value when all independent variables are zero; it can be interpreted as the estimated quantile function for the duration distribution of a crash event with no injury, no rain, and zero space mean speed of two upstream stations at the crash moment. For example, at the 50^{th} percentile the intercept was 104.73 minutes, while at the 75^{th} percentile the intercept becomes 160.88 minutes.

Now we will discuss the effect of the independent variables on the impact duration. The second panel of Figure 12 shows how the change in speed affects the duration at different percentiles. For example, according to the 50^{th} percentile model, per unit (mph) increase of speed will decrease 1.85 minutes duration time. In other words, 10mph speed gain will more likely reduce the duration by 18.5 minutes; on the other hand, 10mph speed reduction will add 18.5 more minutes to the 50^{th} percentile impact duration. Similarly, according to the 75^{th} percentile model, 10mph speed gain will deduct the 75^{th} percentile duration by about 28 minutes and vice versa.

For other two variables including rain and injury, it is found that at lower quartiles, rain and injury are not significant variables to predict the duration. Rain becomes a significant variable at the 40th percentile and remains significant at the upper tail of the duration distribution. Injury is found to be a significant variable at and over 75th percentile and associated positive coefficient shows that crashes with injury are more likely having higher durations.

At the higher quantiles both injury and rain are significant, indicating crashes with higher durations are more likely to occur with the presence of rain and injury. Based on the location of OLS line and quantile regression lines, it is observed that OLS overestimates the effect of rain and injury at lower quantiles while underestimates the effect of those variables on the higher quantiles. It is understandable as the OLS model only addresses the average relationship between the variables; as a result, it is unable to account for the influence of the variables on different percentiles. Hence, a more complete picture of the effect of the predictors can be offered by the quantile regression. For example, the 90th percentile duration of crashes with rain is 36.37 minutes higher than crashes with no rain. According to the OLS model, the average duration is only 17.54 minutes higher, thus the OLS model underestimates the effect of rain for longer durations.

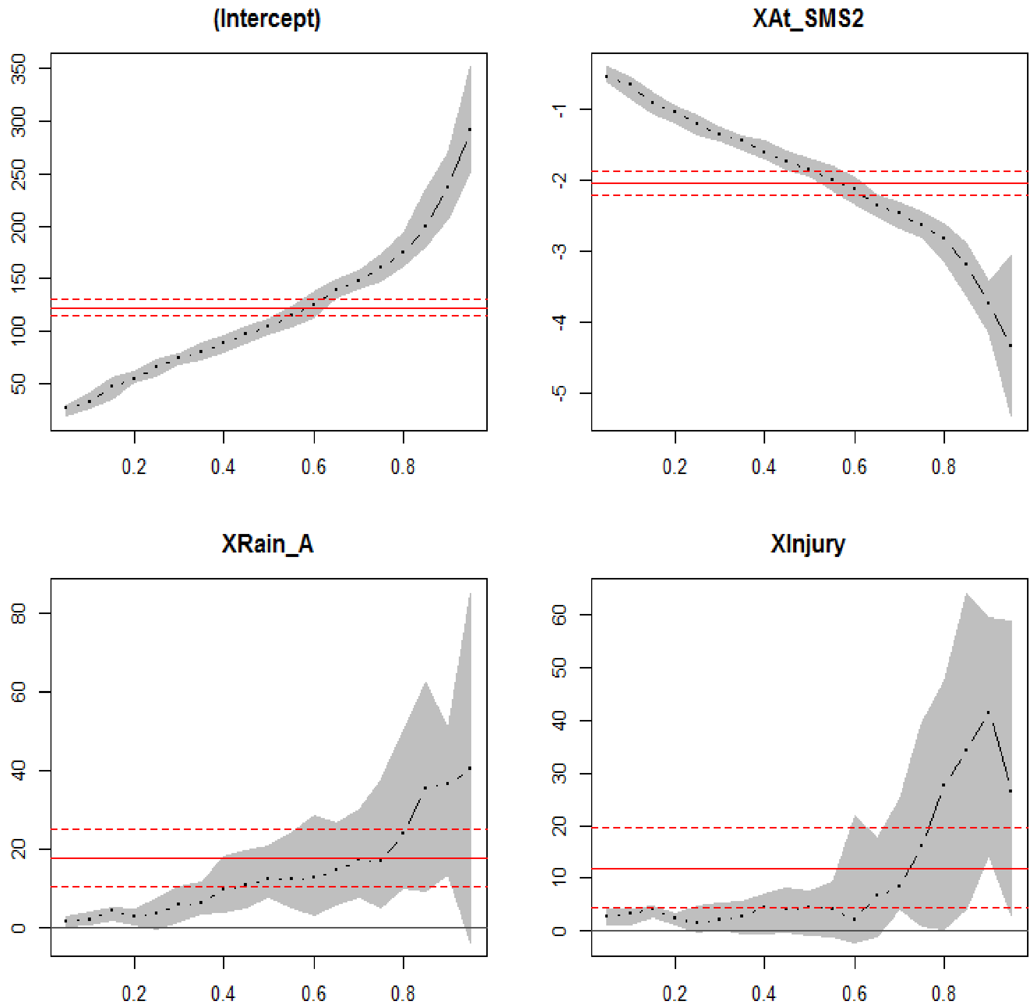


Figure 12: Quantile Regression Result

Table 2 provides the quantile regression results for predicting the impact duration. It includes OLS, 25th percentile, 50th percentile, 75th percentile and 95th percentile regression results. The table contains four sections. the first section shows the resulting duration at different quantiles when space mean speed of two upstream stations at the time when crash occurs is 30 mph without injury and rain; the second section shows the resulting duration for the same speed, without injury but under rainy condition; the third section shows the result for injury involved crashes with same speed but without rain; and the fourth section shows resulting duration with same speed with injury and rainy condition. This table demonstrates how the resulting duration changes at different quantiles with the change of explanatory variables.

Table 2: Impact Duration Prediction

Variable	X	OLS(Mean)		25th		Median		75th		95th	
		β	β*X	β	β*X	β	β*X	β	β*X	β	β*X
constant		122.58	122.58	65.54	65.54	104.73	104.73	160.88	160.88	290.77	290.77
At_SMS2	30	-2.04	-61.2	-1.22	-36.6	-1.85	-55.5	-2.81	-84.3	-4.35	-130.5
Rain	0	17.54	0	3.16	0	12.05	0	16.59	0	40.4	0
Injury	0	11.95	0	1.46	0	4.39	0	16.06	0	26.41	0
Total Duration			61.38		28.94		49.23		76.58		160.27

Variable	X	OLS(Mean)		25th		Median		75th		95th	
		β	β*X	β	β*X	β	β*X	β	β*X	β	β*X
constant		122.58	122.58	65.54	65.54	104.73	104.73	160.88	160.88	290.77	290.77
At_SMS2	30	-2.04	-61.2	-1.22	-36.6	-1.85	-55.5	-2.81	-84.3	-4.35	-130.5
Rain	1	17.54	17.54	3.16	3.16	12.05	12.05	16.59	16.59	40.4	40.4
Injury	0	11.95	0	1.46	0	4.39	0	16.06	0	26.41	0
Total Duration			78.92		32.1		61.28		93.17		200.67

Variable	X	OLS(Mean)		25th		Median		75th		95th	
		β	β*X	β	β*X	β	β*X	β	β*X	β	β*X
constant		122.58	122.58	65.54	65.54	104.73	104.73	160.88	160.88	290.77	290.77
At_SMS2	30	-2.04	-61.2	-1.22	-36.6	-1.85	-55.5	-2.81	-84.3	-4.35	-130.5
Rain	0	17.54	0	3.16	0	12.05	0	16.59	0	40.4	0
Injury	1	11.95	11.95	1.46	1.46	4.39	4.39	16.06	16.06	26.41	26.41
Total Duration			73.33		30.4		53.62		92.64		186.68

Variable	X	OLS(Mean)		25th		Median		75th		95th	
		β	β*X	β	β*X	β	β*X	β	β*X	β	β*X
constant		122.58	122.58	65.54	65.54	104.73	104.73	160.88	160.88	290.77	290.77
At_SMS2	30	-2.04	-61.2	-1.22	-36.6	-1.85	-55.5	-2.81	-84.3	-4.35	-130.5
Rain	1	17.54	17.54	3.16	3.16	12.05	12.05	16.59	16.59	40.4	40.4
Injury	1	11.95	11.95	1.46	1.46	4.39	4.39	16.06	16.06	26.41	26.41
Total Duration			90.87		33.56		65.67		109.23		227.08

Table 3 provides the prediction error that is the absolute difference between predicted and measured duration for different quantile regressions. As our measured impact duration is calculated at 15 minute increments, a prediction error less than 15 minutes is selected to compare the performance of different quantile models. It is observed that, at the 25th

percentile, about 55.5% of crashes have been predicted with less than 15 minutes prediction error; at the 50th and the 75th percentile about 49% and 39% of crashes have been predicted with less than 15 minutes prediction error respectively. Therefore, the 25th percentile model can be used to calculate the least projected duration and the 75th percentile model can be used to determine the maximum projected duration. The 95th percentile model might overestimate the actual impact, so the 75th percentile model is chosen to project the maximum impact duration in this study.

Table 3: Comparison of Percentage of Samples at Different Prediction Tolerances

Prediction Error (min)	OLS	25th	50th	75th	95th
<=5	13.1	39	29.8	20.4	1.9
<=10	25.1	49.9	40.1	32	2.9
<=15	**39.8**	**55.5**	**49.2**	**39.2**	**3.9**
<=30	64.5	66.4	70.4	59.3	11.3
<=60	85.8	82.9	86.7	80.5	37.4

From the incident management point of view, crashes with longer durations are more critical. Quantile regression analysis enables the agency to identify the causes and measure the effect of contributing variables for those longer durations. Such information will help in making effective plans to reduce the longer durations. Moreover, accurate prediction of impact duration of crashes will help traffic management centers better prepared to reduce the crash induced congestion by implementing strategies such as route diversion, informing travelers in advance about crashes and their impact durations etc.

4.4 Impact on Travel Time Reliability

Analyzing travel time reliability enables us to understand how different factors affect the travel experience of the users. In this study, travel rate (seconds/mile) is chosen as the measure of effectiveness to understand how crashes affect the travel reliability. Travel rate can be defined as the time required for traveling per unit distance. This analysis can guide the agencies toward improvements in the operation of road networks. Conducting reliability analysis consists of the following steps:

- Select the region or facilities of interest and study period
- Compile the travel rate data for each facility

- Identify what types of nonrecurring events (peak/off-peak crash, different types of crash etc.) are present in the data
- Develop cumulative distribution functions (CDFs) of the travel rate (TR) for each combination of nonrecurring events

In this study, we analyze the impact of crashes on I-65 NB and SB corridor. First, we calculate the daily travel rate (sec/mile) based on TRIMARC speed data with 15-minute intervals on the study segments in the year 2011-2013. Travel rates are then separated into two groups based on whether they are in peak period or off-peak period. Here, the morning and afternoon peak periods are considered as between 6 AM and 9 AM and 4 PM and 7 PM, respectively.

Our analysis focuses on the impact of the crashes on the reliability. The impact duration determined by the previously developed five step process is used here to select the time slices impacted by crashes. Based on the time periods and crash impact information, we have four different groups: peak crash, peak no crash, off-peak crash, off-peak, no crash.

The results for those four groups by direction of the corridor are shown in Figure 13. Based on the figure, the travel rate increases significantly due to the impact of crashes, compared to the non-crash condition. In addition, the travel rates are the highest for the peak crash group for both directions, due to the overlapping effect of recurrent congestion and crash-induced congestion. Figure 14 and Figure 15 depict the CDF curves of travel rates under different conditions, which can give a more complete picture in terms of the whole travel rate distribution.

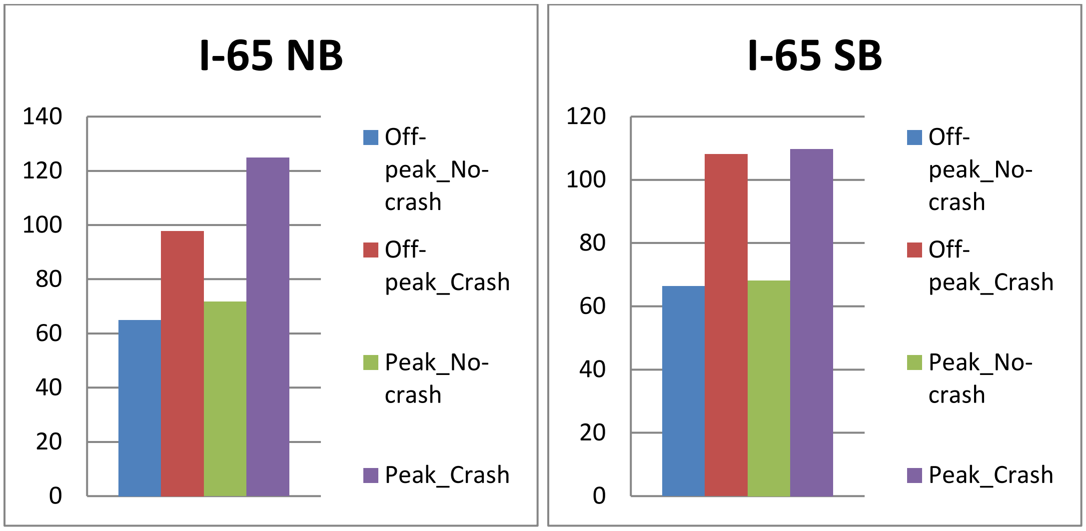


Figure 13: Average Travel Rate

To explain those figures, let's consider a travel rate of 80 sec/mile. According to Figure 14, travelers could travel at 80 sec/mile or less for 96 percent of time during the off-peak period when there is no crash on I-65N. In comparison, the value would be 84 percent for travelers traveling during the peak period under no crash condition. Now when a crash occurs, only 41 and 23 percent of the time travelers can travel at or below 80 sec/mile during the off-peak and peak period, respectively, indicating the significant effect of crashes on the traffic flow. Similar trends can also be observed on I-65S as shown in Figure 15.

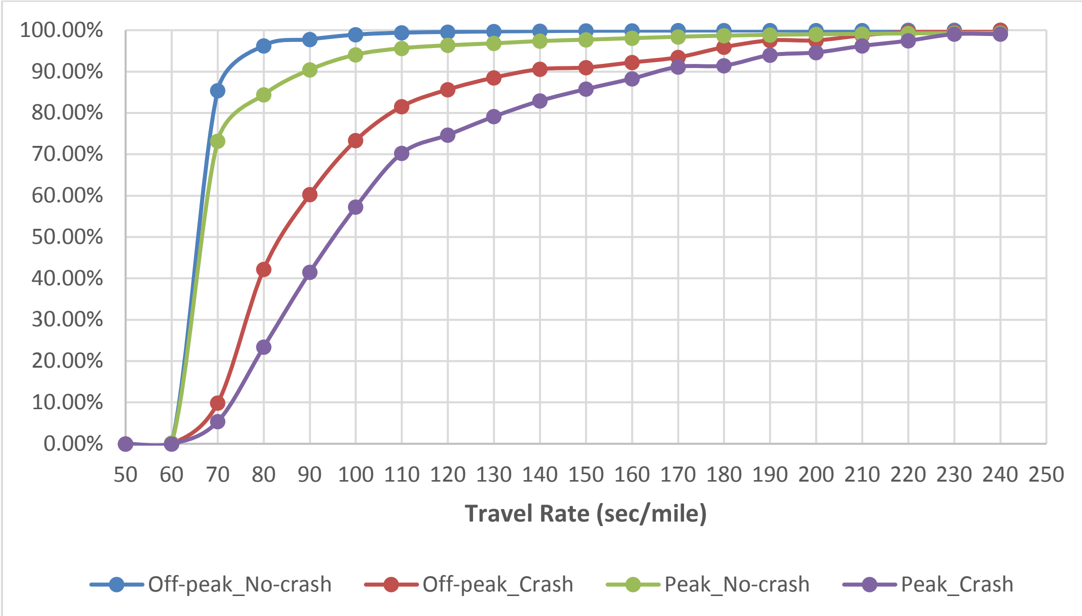


Figure 14: CDF of Travel Rates for I-65N

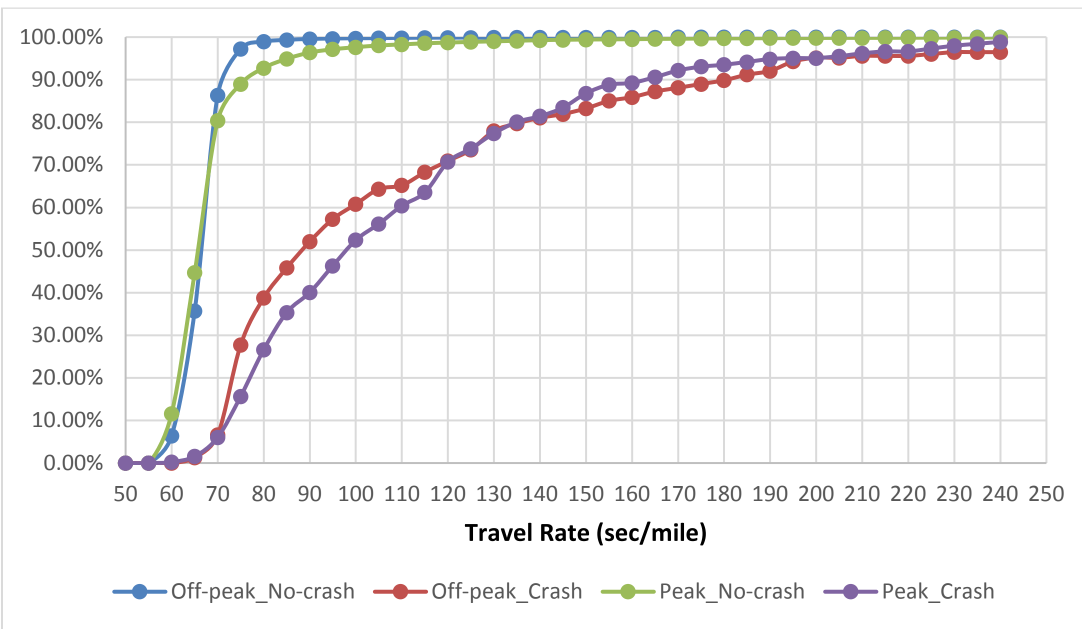


Figure 15: CDF of Travel Rates for I-65S

The benefit of using the actual impact duration over the reported incident duration from the incident log is that it provides a more accurate estimation of travel rate and shows the true situation the travelers experience. Similar analysis with CDF curves is also conducted using the reported incident duration to show the difference between two estimations. Figure 16 demonstrates the comparison between reported and measured duration in calculating CDF of travel rates on I-65N. From Figure 16 (where R=Reported, M=Measured, P=Peak, Cr=Crash), it is clear that CDF of travel rates using reported incident durations underestimates the true effect of the crashes. For example, considering the measured impact duration along I-65N during off-peak under crash condition, only 42% of the time in this period can have 80 sec/mile travel time or less. If the reported incident duration is used, it would be 65% of the time, which is much higher than the measured case. Similar trends are observed during the peak crash group. When using the measured duration only 22% of the time could have the travel rate at or below 80 sec/mile, compared to 41.7% when using the reported duration. During the non-crash situations, the two estimations are almost same. Our main objective is to identify the true impact of the crashes on the travel rates, and that is why actual impact duration found by the five step crash identification process is suggested for estimating the CDF of travel rate as it provides more accurate estimations of the impact of the crashes on the travel experience.

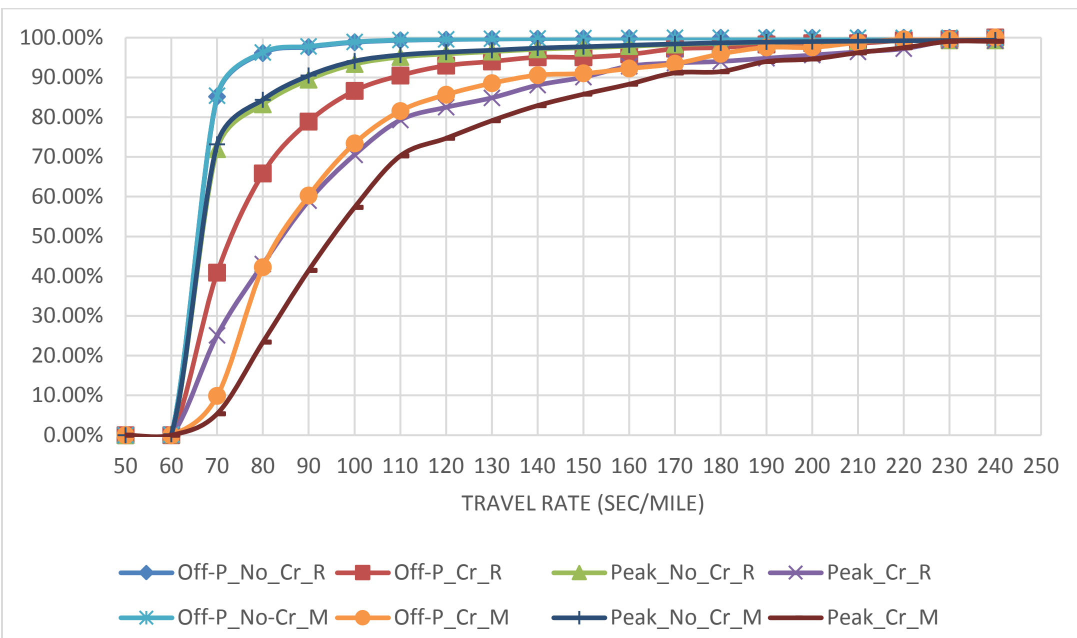


Figure 16: CDF of Travel Rates on I-65N for Reported vs Measured Duration

Next, we analyze the travel rate information for each route under four scenarios to determine the impacts of the crashes on route travel times. These four scenarios are listed as follows.

- No crashes: normal condition with no crashes
- Crash with single lane blocked: crashes with a single lane blocked scenario with and without the shoulder blocked
- Crash with multiple lanes blocked: crashes with multiple lanes blocked scenario with and without the shoulder blocked
- Crash with only the shoulder blocked: crashes that block only the shoulder.

The travel rate CDF for four scenarios are derived to show their different impacts. Taking a travel rate of 90 sec/mile as a baseline for I-65N, then based on Figure 17, about 97% of the time travelers can travel at or below 90 sec/mile when there is no crash. However, when a crash occurs and it blocks the shoulder, it would be only 51% at or below this travel rate. As lane blockage has more apparent impact on the traffic, the results are also revealed in the CDF comparisons. When one lane is blocked by the crash, only 40% of the time travelers can go at or below 90 sec/mile. In addition, when multiple lanes are blocked, there would be only 34% of the time the traffic can move at the same travel rate. The observations on I-65s are similar as shown in Figure 18.

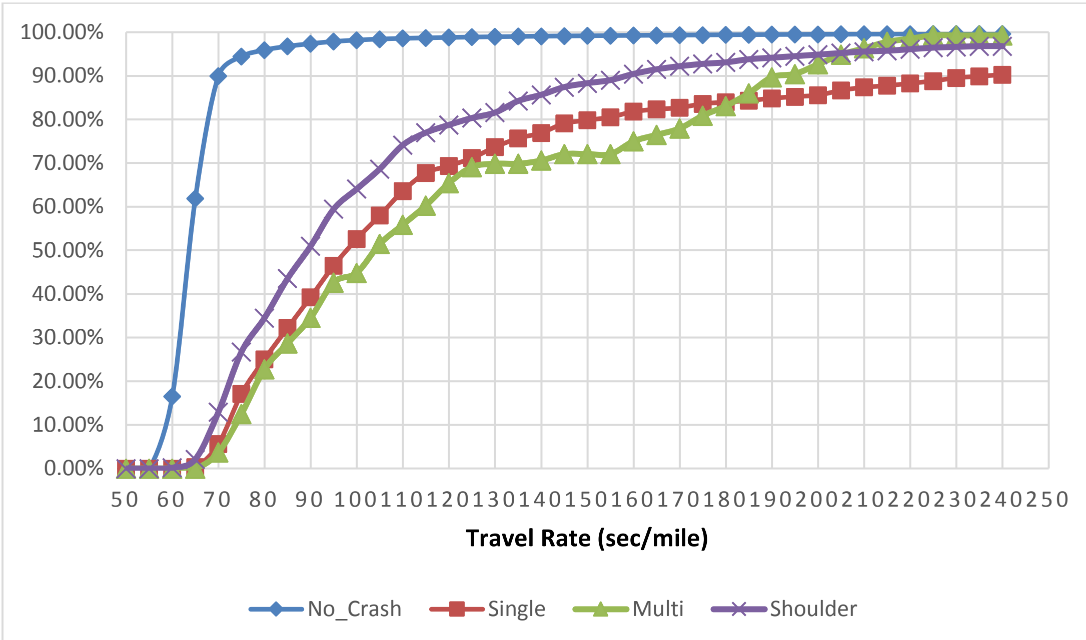


Figure 17: CDF of Travel Rates for Crash Types at I-65N

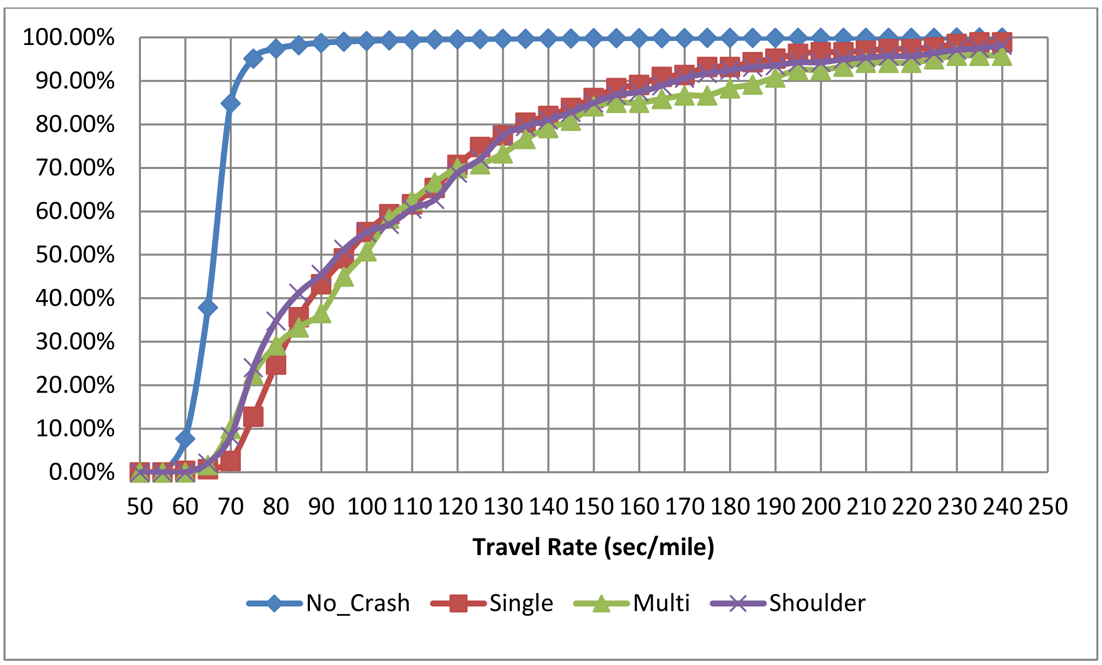


Figure 18: CDF of Travel Rates for Crash Types at I-65S

For travel time reliability analysis, our main objective is to build a graphical framework to show the true effect of crashes on the travel experience. That objective is fulfilled with detailed analysis by scenarios. The CDF of travel rates offers a better way to understand the effect of crashes on the traffic and quantify such effect in a easily understandable format. The analysis provides some insights that would helpful for the traffic management agency decision-making in mitigating the impact of the crashes.

5. Conclusion

Crash induced congestion is one of the major causes of the traffic delay and unreliability. This study develops a methodology for identification of spatiotemporal impact of individual crash based on the stationary sensor data and presents several practical applications of this methodology. The methodology defines the crash impact as the reduction of traffic speeds experienced by the travelers under crash conditions. The speed reduction is determined with respect to the expected traffic speeds based on travelers' past experience. The method involves the development of a background speed profile by using Kalman Filter algorithm and then superimposing the current traffic speed under crash conditions onto it. The resulting difference between two profiles shows the reduction of speed due to the crash and indicates the crash induced congestion. Through the methodology, we can capture the dynamics of the impact of the crash during recurrent and non-recurrent congested conditions.

Integration of the incident data with the stationary sensor data provides the traffic management center a data driven approach to measure the crash induced congestion and its temporal and spatial extent. The case study has demonstrated the performance of the methodology to automatically identify the impact zone of each individual crash. The use of a simple yet informative heat map enhances our visual understanding of the spatiotemporal impact of crashes.

After impact durations are determined, three different models based on the multiple linear regression, logistic regression, and quantile regression are developed to predict the impact duration of a crash for real-time application. The linear model can predict 32% variation of impact durations based on two independent variables, i.e. post-crash speed and weather condition. The logistic model can predict the probability of impact durations falling within different duration ranges based on the same independent variables. Moreover, the quantile regression provides a more detailed analysis of the impact duration from the perspective of the distribution percentiles. Instead of giving a single output, it provides a range of values for the best possible to the worst possible scenario of a crash impact. It is worth mentioning that these models only require the post-crash speed at the upstream segment and the weather condition, both of which can be obtained easily from real-time sensor data and weather information. Therefore, the models can be quickly adapted by transportation management centers and the results can help them better allocate response and management resources and inform travelers of the traffic condition ahead.

The developed methodology also helps refine the analysis on the impact of crashes on the travel time reliability based on the more accurate impact analysis results. To visually see the impact of the crashes, the CDF plots are employed, which can show the overall extent of the travel rates throughout the analysis period. Based on the comparison of travel rate CDF curves, the effects of different types of crashes are observed. Among those crash types, crashes that result in multiple lane blockage affect the traffic most significantly. Compared to the normal condition, in which over 97% of the time travelers can travel at a relatively high travel rate, they can travel in the same level of condition with only 34% of the time when multi-lane crashes occur. The findings in this study can help to shape crash management strategies for different types of crashes at different periods.

REFERENCES

1. Lomax, T., D. Schrank, S. Turner, and R. Margiotta., *Selecting travel reliability measures*. 2013: Texas Transportation Institute.
2. Neudorff, L. G., Randall, J. E., Reiss, and R. Gordon, *Freeway management and operations handbook*. 2003.
3. Garib, A., A. Radwan, and H. Al-Deek, *Estimating magnitude and duration of incident delays.* Journal of Transportation Engineering, 1997. **123**(6): p. 459-466.
4. Smith, K.W. and B.L. Smith, *Forecasting the clearance time of freeway accidents*. 2001: Research Report, STL-2001-01.
5. Khattak, A.J., J.L. Schofer, and M.-h. Wang, *A simple time sequential procedure for predicting freeway incident duration.* Journal of Intelligent Transportation Systems, 1995. **2**(2): p. 113-138.
6. Nam, D. and F. Mannering, *An exploratory hazard-based analysis of highway incident duration.* Transportation Research Part A: Policy and Practice, 2000. **34**(2): p. 85-102.
7. Giuliano, G., *Incident characteristics, frequency, and duration on a high volume urban freeway.* Transportation Research Part A: General, 1989. **23**(5): p. 387-396.
8. Golob, T.F., W.W. Recker, and J.D. Leonard, *An analysis of the severity and incident duration of truck-involved freeway accidents.* Accident Analysis & Prevention, 1987. **19**(5): p. 375-395.
9. Wu, W., S.Y. Chen, and C.J. Zheng, *Traffic incident duration prediction based on support vector regression.* Proceedings of the ICCTP, 2011. **2011**: p. 2412-2421.
10. Lin, P.W., N. Zou, and G.L. Chang. *Integration of a discrete choice model and a rule-based system for estimation of incident duration: a case study in Maryland*. in *CD-ROM of Proceedings of the 83rd TRB Annual Meeting, Washington, DC*. 2004.
11. KIM, H.J. and C. Hoi-Kyun, *A comparative analysis of incident service time on urban freeways.* IATSS research, 2001. **25**(1): p. 62-72.
12. Boyles, S., D. Fajardo, and S.T. Waller. *A naive Bayesian classifier for incident duration prediction*. in *86th Annual Meeting of the Transportation Research Board, Washington, DC*. 2007.
13. Wang, W., H. Chen, and M.C. Bell, *Vehicle breakdown duration modelling.* Journal of Transportation and Statistics, 2005. **8**(1): p. 75-84.
14. Skabardonis, A., K. Petty, H. Noeimi, D. Rydzewski, and P. Varaiya, *I-880 field experiment: Data-base development and incident delay estimation procedures.* Transportation Research Record: Journal of the Transportation Research Board, 1996(1554): p. 204-212.
15. Wang, Y., M.E. Hallenbeck, and P. Cheevarunothai, *Quantifying incident-induced travel delays on freeways using traffic sensor data*. 2008, Transportation Northwest, University of Washington.
16. Messer, C.J., C.L. Dudek, and J.D. Friebele, *Method for predicting travel time and other operational measures in real-time during freeway incident conditions.* Highway Research Record, 1973(461).

17. Cragg, C.A. and M.J. Demetsky, *Simulation Analysis of Route Diversion Strategies for Freeway Incident Management*. 1995.
18. Zhang, H., F. Ni, and S. Yang. *Prediction of freeway traffic incident delay based on simulation*. *15th International IEEE Conference on Intelligent Transportation Systems*. 2012. IEEE.
19. Fries, R., C. Hamlin, M. Chowdhury, Y. Ma, and K. Ozbay, *Operational impacts of incident quick clearance legislation: a simulation analysis.* Journal of advanced transportation, 2012. **46**(1): p. 1-11.
20. Kabit, M.R.B., P. Charles, L. Ferreira, and I. Kim , *Modelling major traffic incident impacts and estimation of their associated costs.* Transportation planning and technology, 2014. **37**(4): p. 373-390.
21. Tu, H., H. Li, H. Van Lint, and H. van Zuylen, *Modeling travel time reliability of freeways using risk assessment techniques.* Transportation Research Part A: Policy and Practice, 2012. **46**(10): p. 1528-1540.
22. Yu, R., Q. Shi, and M. Abdel-Aty, *Feasibility of incorporating reliability analysis in traffic safety investigation.* Transportation Research Record: Journal of the Transportation Research Board, 2013(2386): p. 35-41.
23. Zhong, C., V.Sisiopiku, K. Ksaibati, and T. Zhong, *Crash prediction on rural roads*. in *Proceeding of the 3rd International Conference on Road Safety and Simulation, Indianapolis, Indiana*. 2011.
24. Miaou, S.P. and H. Lum, *Modeling vehicle accidents and highway geometric design relationships.* Accident Analysis & Prevention, 1993. **25**(6): p. 689-709.
25. Okamoto, H. and M. Koshi, *A method to cope with the random errors of observed accident rates in regression analysis.* Accident Analysis & Prevention, 1989. **21**(4): p. 317-332.
26. Wright, B., Y. Zou, and Y. Wang, *Impact of Traffic Incidents on Reliability of Freeway Travel Times.* Transportation Research Record: Journal of the Transportation Research Board, 2015(2484): p. 90-98.
27. Chung, Y. and W.W. Recker, *A methodological approach for estimating temporal and spatial extent of delays caused by freeway accidents.* IEEE Transactions on Intelligent Transportation Systems, 2012. **13**(3): p. 1454-1461.
28. Pan, B., U. Demiryurek, C. Gupta, and C. Shahabi, *Forecasting spatiotemporal impact of traffic incidents for next-generation navigation systems.* Knowledge and Information Systems, 2015. **45**(1): p. 75-104.
29. Yu, R., M.A. Abdel-Aty, M.M. Ahmed, and X. Wang, *Crash-type propensity analysis with bayesian models using microscopic traffic and weather data*. in *Transportation Research Board 92nd Annual Meeting*. 2013.
30. Ahmed, M., M. Abdel-Aty, and R. Yu, *Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data.* Transportation research record: Journal of the Transportation Research Board, 2012(2280): p. 51-59.
31. Hojati, A.T., L. Ferreira, P. Charles, and M.R. bin Kabit, M.R., *Analysing freeway traffic-incident duration using an Australian data set.* Road & Transport Research: A Journal of Australian and New Zealand Research and Practice, 2012. **21**(2): p. 19.

32. Kalman, R.E., *A new approach to linear filtering and prediction problems.* Journal of basic Engineering, 1960. **82**(1): p. 35-45.
33. Chen, M., X. Liu, and J. Xia, *Dynamic prediction method with schedule recovery impact for bus arrival time.* Transportation Research Record: Journal of the Transportation Research Board, 2005(1923): p. 208-217.
34. Mason, R.L., R.F. Gunst, and J.L. Hess, *Statistical design and analysis of experiments: with applications to engineering and science*. Vol. 474. 2003: John Wiley & Sons.
35. Koenker, R., *Quantile regression*. 2005: Cambridge University Press.