2020

Work zone crash prediction model and characteristics analysis

Clint Kassmeyer
Iowa State University

Work zone crash prediction model and characteristics analysis

by

Clint Kassmeyer

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Civil Engineering (Transportation Engineering)

Program of Study Committee:
Omar Smadi, Major Professor
Jennifer Shane
Christopher Day

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

AADT	Average Annual Daily Traffic
AIC	Akaike Information Criterion
DOT	Department of Transportation
GPS	Global Positioning System
NB	Negative Binomial
PDF	Portable Document Format
SPF	Safety Performance Function
XD Segment	Extreme Definition Segment

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Omar Smadi, who throughout my time as a graduate student helped direct me through my degree, my thesis writing, research and other graduate student questions. In addition, I appreciated the opportunity he gave me to do my research assistantship, with him as my supervisor. I would also like to thank my committee members, Dr. Jennifer Shane, and Dr. Christopher Day, for their guidance and support throughout the course of this research.

In addition, I would like to thank Neal Hawkins and Skylar Knickerbocker for the opportunity to pursue a career in Civil Engineering and my master's degree at Iowa State as well as my appreciation for all the time spent helping me with projects and questions. I would also like to thank the Center for Transportation Research and Education (CTRE) and Iowa State University for providing me with a research assistantship, not only as a means to grow my knowledge and experience in the transportation field, but also provide a boost financially.

I would also like to thank all other faculty and staff at Iowa State University and the Institute for Transportation for all the opportunities that they provided me. Specifically, Intrans for allowing me to work as a receptionist, undergrad research student and graduate research student over the last seven years.

In addition, I would also like to thank my friends and colleagues. I want to also offer my appreciation to those who were willing to participate in my observations, without whom, this thesis would not have been possible.

Finally, I would like to thank my family. Who have supported my throughout my whole life. Without their sacrifices, I would not have never had this opportunity.

ABSTRACT

This research’s main goal is to improve the safety of work zones. It is commonly thought that work zones have a negative impact on the number of vehicle crashes. In addition to the high cost for maintaining roadway and new roadway construction, safety is one of the highest concerns for local DOTs. This research uses Iowa Crash, INRIX and work zone plan data to predict and analyze key work zone crash characteristics. The work zone crash prediction model was constructed using a negative binomial regression model in combination with a random forest importance plot and an exhaustive search engine. The resulting final data included 511 crashes throughout 32 work zones from 2017-2018.

The resulting final model provides a relationship that accurately predicts the number of work zone crashes in Iowa. The equation can predict accurately for data within the boundaries of the variables used in the study in Iowa work zones. The equation had poor accuracy when predicting low risk work zone crashes, or work zones with low crash values. In addition to prediction, the research analyzed the affect each variable in the final model had on the number of crashes. Work zone length was extremely impactful on work zone crashes. While DOT district, divided roadways and AADT were also impactful. Additional research is recommended as work zone data was limited as well as a large proportion was missing. In the near future a similar study is recommended when work zone dates and lengths are more accurately reported. In addition, more work zone data should be used either from other state DOTs or more years of data.

CHAPTER 1. INTRODUCTION

Problem Statement

Road construction is a necessity to keep traffic safe and comfortable for drivers. The addition of work zones is required to keep roads in functional condition. This can be a concern for drivers, contractors and local DOTs as work zones can influence driving behaviors negatively, especially in high volume areas, metropolitan areas, or areas with difficult sight lines. The paramount concerns that come with the addition of a work zone is the negative affects to safety, traffic flow and delay. Lane closures on high volume roadways can lead to large amounts of congestion specifically for commuters in the AM and PM peak time periods.

Lane closures will inherently cause more congestion on roadways. This congestion leads to increased cost on the user including time delay cost, fuel combustion cost, traffic accident cost, time loss cost and environment costs. According to the database of City Traffic Running Status and Statistic Yearbook 2010 in Beijing, the social cost of traffic congestion was 58 billion Yuan RMB *(15)*. This value translates to roughly 8 billion USD. A large percentage of this cost is due to time delay. A recent study completed by INRIX in 2018 determined that the congestion cost in the US was $87 billion (*11*). This study estimated that the cost to each driver was 97 hours or an estimated $1,348 a year (*11*). Due to this significant cost, work zone lane closures are under pressure to not increase congestion in massive quantities. Additionally, congestion will also inherently lead to an increase in the number of vehicle crashes. Balancing these challenges while continuing to keep work zones safe for both workers and drivers is a difficult task for local DOTs to complete.

Improving work zone safety is the top priority of this study. According to the Centers for Disease Control and Prevention, the number of motor vehicle deaths that have occurred from

1982 to 2017 is 27,037, averaging 773 per year *(4)*. The highest number of work zone fatalities occurred in 2002, at 1,186 people *(4)*. As work zone technology and data analytics have improved for work zones there was a steady declined in fatalities until 2014. In 2014, the average number of work zone fatalities was 591 throughout the United States. From 2015 to 2017 the number of fatalities increased to 772 on average *(4)*. The recent increase in work zone related fatalities is one contributing factor that has prompted this research study.

As mentioned above, vehicle crashes in work zones add an additional cost to work zones. In addition to the already expensive $5-20 million per mile cost to add freeway lanes work zones have a large safety cost as well *(18)*. According to Mohan et al in 2002, 30% of work zone accidents involve construction workers *(17)*. Out of those 30%, 27,000 are first-aid injuries and 26,000 are lost-time injuries per year amounting to an annual cost of $2.46 billion *(17)*. In addition to costs paid by the workers and DOTs, there is a large cost paid by the driver and insurance agencies. They found that roughly $6.2 billion per year is the estimated cost of work zone crashes, with the average cost per incident being $7,676 *(17)*. The study found that each year motorists suffer approximately 700 fatalities, 40,000 injuries, and 52,000 property-damage-only accidents *(17)*.

Work zone safety studies have been attempting to predict the effects work zones have on crashes since the 1970's. More sophisticated techniques and software technologies being used today help eliminate some of the inconsistencies from older studies. This study identifies key contributing variables to crashes in work zones using the most up to date data available in Iowa. Variables tested in this study cover a substantially larger variety, including severity, road conditions, road location, time of day, time of year, driver behavior and much more by using crash INRIX and work zone data sets.

Objectives

The first objective of this research is to create safer work zones for both workers and drivers. As previously mentioned, safer work zones also have a positive impact on work zone costs. Creating an accurate regression model for predicting the number of crashes will accomplish this goal by analyzing key factors that have a significant negative impact to work zone crashes. This information could be used by local DOTs to add additional crash prevention features to vulnerable work zones. For example, if this study finds that a high speed limit results in a large increase in vehicle crashes, a DOT may consider adding extra speed reduction features to decrease drivers speeding tendencies. In addition, the number of predicted crashes could be used in each specific work zone while in the planning stages. Allowing for safety features to be implemented on work zones with a high number of predicted crashes before construction begins.

A secondary objective is to determine an adequate modeling approach to predicting crashes, in this case for work zones specifically. Work zone data, speed data and crash data are being collected at much higher rates and at better quality than they have been in previous years. Providing a base line for future studies is paramount for future crash research.

Another objective is to find gaps in current data collection methods. Whether data is missing, not available or not accurate data it can be noted in this study for future improvements. As mentioned above, data collection will only improve and become more accessible in the future and any noted data limitations could have a large impact in accuracies of future studies.

Organization

This thesis is divided into 6 Chapters: introduction, literature review, data, methodology, analysis and conclusion. In Chapter one, Introduction, the negative impacts that work zones have on drivers, and local DOT's are presented. The effect that active work zones have on safety, congestion and costs are noted. This chapter also includes the objectives of the thesis, mainly on improving the safety of work zones, but also briefly focusing on a basis for future work.

The literature review provides information on previous studies. The study identifies key issues that arise in work zones and other roadways involving crashes and congestion effects. The chapter also provides a background of what variables should be tested in the regression model. Studies that provide background and validation for the model selection process are also included in this section to justify their uses in the methodology and analysis sections.

Chapter 3, data, gives the reader a general knowledge of each data set used in the study as well as what variables are included from each source. This information can be found in each data set specific section. In addition, it includes a description of the process for gathering and combining the data sets in the data filtering section.

Chapter 4, methodology, investigates statistical parameters such as the mean, standard deviation, median, maximum and minimum values for numerical variables in the final data set. As well as elaborates on the reasoning for choosing a negative binomial regression model and the statistical format for such a model. The last section in the methodology chapter denotes the model selection process and elaborates more on how each model had a set of limitations to work around.

The analysis chapter provides insight on the final negative binomial regression model selected. First of which is providing coefficients, standard error, z values, probabilities and significance levels in a table. After the final regression model is presented, each variable was analyzed to determine the impact that they had on the number of crashes predicted.

Finally, the conclusion chapter, key factors are noted in the first section in bullet form. Each bullet summarizes the variable as well as the impact that it had on the number of predicted crashes in the final regression model. The second section goes in depth on the limitations of the study, primarily focusing on limitations to the data sets, but also including limitations on computational power. The last section of the chapter discusses what work is being done and should be done in the future to improve upon this study. In addition, some recommendations are made to help improve work zone crash prediction modeling based on the limitations mentioned.

CHAPTER 2. LITERATURE REVIEW

Information in the literature review chapter was split into three separate sections to keep the information more organized. Due to this study being focused on creating a work zone crash predicting model, studies included in this section address variables that were found to be key factors in crash prediction in their respective studies. Some studies below include variables that were impactful crash predictors in non-work zone environments, therefore were tested to analyze if the same result would be true in a work zone environment. The first section, crash severity studies, concentrates on work zone crash studies that focus on the types of injuries occurring. Topics include the analysis of non-injury versus injury, non-fatal versus fatal, crash manner and more in crash situations. The environmental variable studies section focuses on studies that include variables related to either roadway conditions, traffic behavior or road orientation. The last section is statistical modeling studies. This section aided in the model selection process and general understanding of each method.

Crash Severity Studies

Data collection for vehicle crashes has drastically increased in the last 5 years and as more data becomes available, better studies help to understand how to mitigate them. Accurate crash data sets help identify the impact that work zone crash severity has, which is an essential relationship to analyze. The studies below investigate a variety of states, roadway features, environments, and others and their relationships involving the severity of crashes.

A study published in 2018 by Ullman et al analyzed crash characteristics and countermeasures and modeled the results (*19*). Using a Virginia DOT crash database to the similar to what was used in this research, they determined crashes caused by work zones solely by the responding officer's report. After determining work zone related crashes, they filtered out vehicle crashes outside of the work zone area. They used a work zone database to get access to variables like traffic volumes, speed limits, and work zone geometry. Using the previously mentioned data, they created a cross-sectional statistical model (*19*). Factors were included for the segment before and during work zone activity. One of the major conclusions was work zones were a significant factor to crashes, due to the queues and congestion that resulted from them (*19*). Another major outcome found from the study was that work vehicles entering and exiting work zones cause a large increase in rear-end collisions and sideswipe crashes (*19*). Lastly, they found that urban environments caused an increase in crashes in which sight distance challenges, obstructions due to equipment being too close to the road and driver confusion were attributed to the crashes (*19)*.

A study conducted by Khattak & Council in 2002 analyzed the effects work zones have on injury and non-injury crashes (*12*). Data was collected pre-work zone and during work zone times in 36 work zones in California. Their data included crash frequency, crash severity, AADT, urban and rural information as well as work zone duration, length and location. They investigated the crash rates before and during work zone activity. Using a negative binomial model, they found that increasing work zone duration, length and AADT significantly increased the frequency of injury and non-injury crashes (*12)*.They found that on limited-access highways work zone crashes were 21.5% higher than pre-work zone crashes (*12*). Another conclusion was that the increase in non-injury crashes was lesser than those with injuries. The crash rate increased 23.5% for non-injury and 17.5% for crashes with injuries (*12*).

In Virginia from 1996 through 1999, Garber and Zhao conducted a study analyzing trends in work zone crash locations, severity and crash manner (*9*). They believed that work zones added potentially hazardous conditions for both drivers and workers. Therefore, they conducted the study using police crash records to analyze five areas of a work zone: advanced warning, transition, longitudinal buffer, activity, and termination. They found that the activity area of the work zone was the most likely area for a collision, while the termination area was the safest (*9*). When it comes to safety, they found that property damage only incidents were most likely and fatal crashes were the least likely to occur (*9*). Finally, the study determined that rear-end collisions were most common in the advanced warning area, where 83% of crashes in that zone were rear-end collisions (*9*). Rear-end collisions were also a majority of all crashes in all observed work zones at 52% of crashes (*9*).

According to a study by Akepati and Dissanayake in 2011, work zone crashes accounted for 9,900 fatalities in the United States in the last 10 years (*1*). They conducted a study using data from smart work zones in Iowa, Kansas, Missouri, Nebraska and Wisconsin to address potential hazards in work zones. Crash data was collected from 2002-2006. In order to test for a relationship between the number of vehicle crashes and other variables they used a chi-square test. They found that the largest percentage of accidents involving work zones occur where the actual work goes on, otherwise known as the activity zone (*1*). They also concluded that lane closure work zones are the work zones with the largest number of crashes and of those crashes rear end collisions are the most common (*1*). This aligns with the study above by Garber and Zhao. One result they found was that most of the crashes that occur in work zone were during daylight, with no adverse weather conditions, implying that weather is not a key contributing circumstance to crashes (*1*).

Daniel, Dixon and Jared took a different approach to analyzing work zone safety in a Georgia study in 2000. Their study examined the differences between fatal and non-fatal work zone crashes to further expand knowledge of work zone safety and apply it to preventative measures (*8*). Crash data was collected for two years for work zone and non-work zone crashes. Three work zone locations were used throughout the study. One result they found is that construction work zones result in more fatal crashes than maintenance work zones (*8*). They found that a higher proportion of fatal crashes occurred in dark conditions as opposed to non-work zone locations (*8*). Fatal crashes in work zones also were more likely to involve trucks in work zone areas as opposed to non-work zone areas (*8*). In addition, fatal crashes were more likely to involve another vehicle than non-work-zone fatal crashes (*8*).

Another common approach in work zone safety analysis utilizes crash severity indexes (CSI). A CSI is a numerical value between zero and one that estimates values for certain variables, it is interpreted as the likelihood of having fatalities when a crash occurs. A study done by Li and Bai in 2008 analyzed several variables in work zone highway crashes using a CSI modeling approach. Crash data was collected in Kansas in 2004 to validate the model. The CSI models were selected using a three-step process were first the variables were examined and significant risk factors were selected *(14)*. Second a logistic regression technique was utilized to determine CSI values for each selected variable *(14)*. Lastly, the model was validated using recent crash data *(14)*. Their resulting model contained variables: Light condition, Vehicle Type, No. of Lanes, Speed Limit, Area Information, Inoperative Traffic Control, Flagger, Stop Sign/Signal, Age, Alcohol/Drug, Impairment, Disregarded Traffic, Speeding and Following to Close *(14)*. These variables created a model that predicted the risk of fatal crashes occurring, with higher values suggesting work zones have higher risk of fatalities. They found that the CSI model performed well, though they recommended that future work incorporate other crash severity levels *(14)*. In addition, the model validation only had 18 fatal crash cases, which they noted may not be a large enough sample size for future studies.

Roadway Factors Studies

Work zones also add potentially negative impacts involving collisions. Factors like merging and lane closures can increase the likelihood of rear-end collisions. Lack of visibility due to construction crews or other obstructions could result in distractions and general confusion for the driver and possibly cause accidents. In this section roadway factors like AADT, number of lanes and speed limit among others were analyzed in their respective studies.

A Study in 2002 by Chambless et al analyzed the differences in vehicle crashes in work zones and non-work zone environments (*5*). The study was conducted in three states: Alabama, Michigan and Tennessee from 1994-1998. The study used an Information Mining for Producing Accident Countermeasure Technology (IMPACT) model. Where the IMPACT model compared crashes in work zones to a control subset, in this case crashes outside of work zones. They found that 63% of work zone crashes occur on interstate, U.S. and state roads compared to only 37% of non-work zone crashes occurring in the same roadway types (*5*). 48% of work zone crashes occurred in areas with speed limits between 55 and 45 mph, while only 34% of non-work zone crashes occurred in those speed limit ranges (*5*). 27% of crashes in a work zone listed the primary contributing crash circumstance as misjudging stopping distance and following too close while it was only reported for 15% of crashes not in work zones (*5*).

One of the most prevalent variables used to construct a model to predict crash frequency on a road section is the AADT. The total volume of traffic on the roadway is the limiting factor on the number of crashes that could occur. In 2016, Chen and Xie conducted a study to attempt to model the effects that AADT has on predicting multiple vehicle crashes at signalized intersections *(6)*. In the study they applied generalized additive models (GAMs) as well as Piecewise linear negative binomial (PLNB) regression models to fit crash data. While they found that their model could be improved by using a nonlinear function form, they found that AADT in conjunction with other joint safety effects were most important *(6)*. In addition, they found that additional research should be done on the minor to major approach AADT ratio in the near future as they had varying results *(6)*.

The shape safety performance functions or SPF is best described by a sigmoid reflecting a dose-response type of relationship between safety and traffic demand on urban freeways (*13*). This suggests that safety and congestion hold a relationship were safety deteriorates with the reduction in level of service on the roadway. As a general rule of thumb, it is believed that additional capacity granted by adding more lanes increases the safety of the roadway. In 2008, a study was conducted by Kononov, Bailey, and Allery to determine if adding additional lanes did in fact add additional safety benefits to urban multilane highways (*13*). The study was conducted using data from Colorado, California and Texas. The modeling process was created using traffic operations parameters described in the Highway Capacity Manual. They used neural networks to determine the relationship between safety and exposure (*13*). They found that as AADT increased, the slope of the SPF became steeper, insinuating that crashes increase at a faster rate than those with a lower number of lanes (*13*).

Using the Florida crash records from 2002 to 2004 Harb et al conducted a study to uncover work zone freeway crash characteristics. In the study, conditional logistic regression in conjunction with stratified sampling and multiple logistics regression models were utilized to determine key crash traits *(10)*. The main goal of the study was to analyze crash characteristics to help improve countermeasures that limit work zone hazards. According to the model results, they found that the key factors associated with work zone crash were: roadway geometry, weather condition, age, gender, lighting condition, residence code and driving under the influence of alcohol and/or drugs *(10)*.

Statistical Modeling Studies

Two of the main goals for a statistical analysis are simplicity and accuracy. Creating a model with a relatively low number of variables while still being an accurate predictive model for work zone crashes was one of the main challenges of the study. Multiple approaches were attempted in the study, each of which was researched to see if it would be an effective way to create the final regression model.

One of the first approaches was to investigate the use of lasso, ridge or elastic net modeling. These are explained further in Chapter 4, methodology, but are all relatively similar approaches to reduce the number of variables in a model. A study done in 2012 by Zou and Hastie found that an elastic net proved to be more accurate than a lasso model for leukemia studies (*20*). The data set used consisted of 40 predictor variables and the number of observations was varied from 100-400 over 50 simulations *(20)*. They concluded that elastic net encouraged a grouping effect, where strongly correlated variables tend to be all in or all out of the model *(20)*. They determined that elastic net performs much better than lasso when the number of predictors is greater than the number of observations. While in contrast lasso performed better when the number of observations was much greater than the number of predictors. Another study in 2009 by Cho et al used an elastic net to detect disease-causing genes. They determined that using both lasso and ridge penalties helped eliminate issues with multicollinearity in the model (*7*). In both studies they found that elastic net modeling outperformed the lasso model.

Another model used in the study is the exhaustive search engine. This model selection process uses both forward and backward selection methods to determine the model with the best

AIC values between the intercept model and the full model. A study done by Capelli R. et al in 2019 used an exhaustive search method to determine the ligand binding pathways of a benzene molecule *(3)*. The resulting model predicted all previously identified ligand binding pathways as well as it identified 3 pathways not yet found *(3)*. One of the biggest benefits for the study was that the computational cost for running the exhaustive search was much less than that of other options.

One difficulty faced with trying to simplify the model was having many statistically significant variables to choose from. As one of the main goals of the project was to end with a relatively simple model to use this was a major concern. According to Archer, K. and Kimes, R using Random Forest Importance plots are best for when there are many variables in the data set *(2)*. Their study incorporated random forests to predict the phenotype calls using gene expression data *(2)*. In the data set in their study they had a large number of genes (predictors) and a small number of observations. They found that the random forest was attractive for studies with goals to produce an accurate classifier and to provide insight regarding the discriminative ability of individual predictor variables *(2)*.

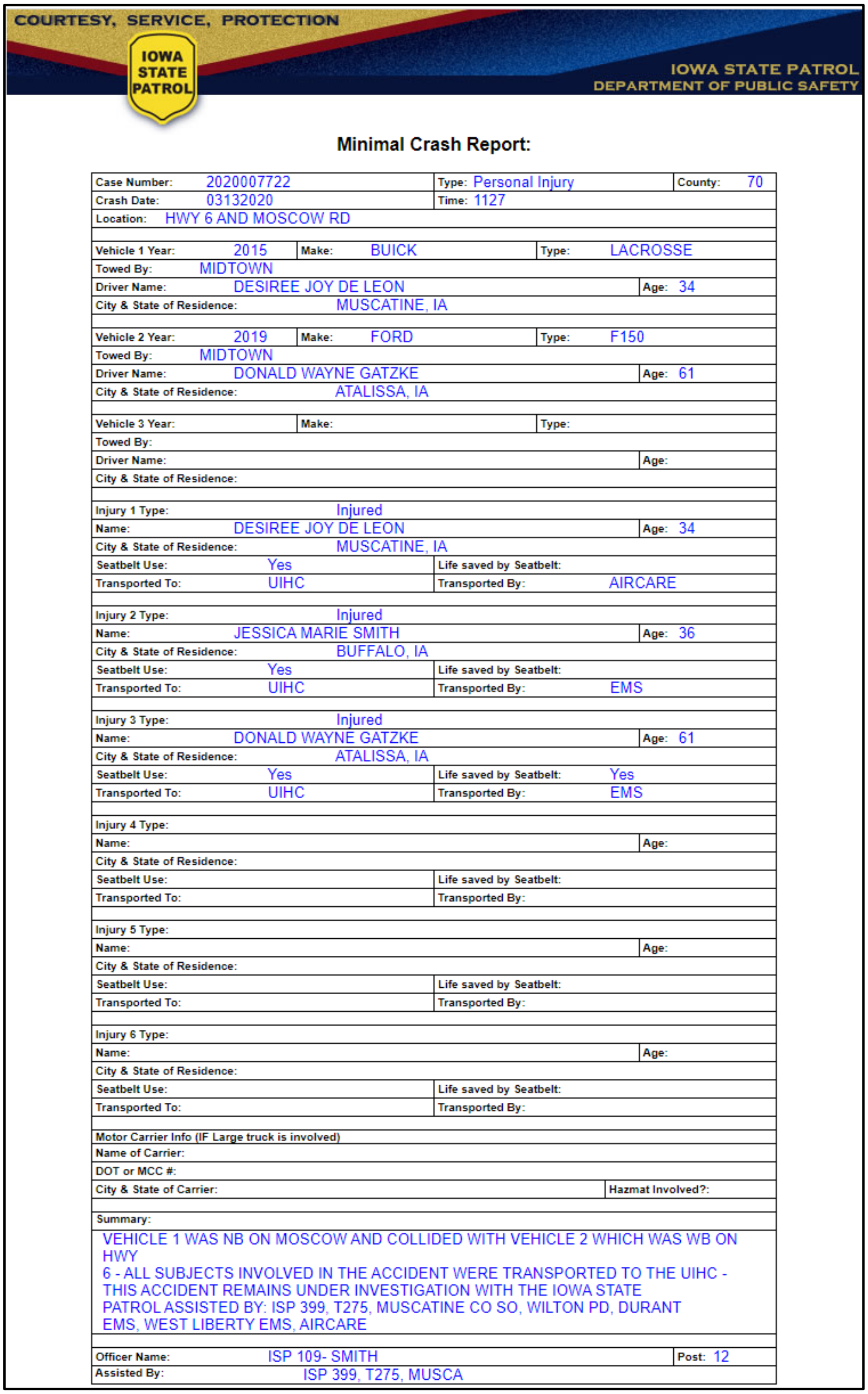
CHAPTER 3. DATA

This study aims to investigate the effects that work zones have on roadway safety in Iowa. The Iowa DOT supplied three data sets, crash data, INRIX data, and work zone plan data, which were used to accomplish this goal. Crash data and INRIX data were accessed directly from the Iowa DOT. While work zone plan data was supplied by the Iowa DOT, it had to be manually collected from work zone pdf files. All three data sets were combined into one data set that included active work zones during the time periods of 2017-2018. Work zones were considered active for the entire year as dates were not accurately or readily available in the work zone pdfs. The resulting final data set included 511 vehicle crashes in 32 work zones throughout the two years included.

Crash Data

The crash data set originated from the Iowa DOT database. The data set includes crash characteristic variables recorded when a collision occurs in Iowa based on the crash reporting requirements. Variables include those regarding reason and location of the crash, while others denote day conditions such as lighting, road conditions, weather conditions and much more. The crash data set also includes a variable called Wz.Related, which denotes if the crash occurred due to the work zone. This variable can be biased as it is recorded by the responding officer and may not always be the primary cause of the accident. In order to include all crashes in the work zone area, crashes were determined to be work zone related based solely off the longitude and latitude of the crash compared to that of the work zone.

An example of an Iowa Crash Report is shown in Figure 1. Description of the vehicles and persons involved in the crash as well as the location and general description of the incident are reported. All information is provided by the responding officer. This data is automatically sent to a dataset that records all crashes in the state by the Case Number. The report shown is a minimal report and does not include all variables that are in the data set.



COURTESY, SERVICE, PROTECTION

IOWA STATE PATROL

IOWA STATE PATROL
DEPARTMENT OF PUBLIC SAFETY

Minimal Crash Report:

Case Number: 2020007722		Type: Personal Injury	County: 70
Crash Date: 03132020		Time: 1127	
Location: HWY 6 AND MOSCOW RD			

Vehicle 1 Year: 2015	Make: BUICK	Type: LACROSSE	
Towed By: MIDTOWN			
Driver Name: DESIREE JOY DE LEON			Age: 34
City & State of Residence: MUSCATINE, IA			

Vehicle 2 Year: 2019	Make: FORD	Type: F150	
Towed By: MIDTOWN			
Driver Name: DONALD WAYNE GATZKE			Age: 61
City & State of Residence: ATALISSA, IA			

Vehicle 3 Year:	Make:	Type:	
Towed By:			
Driver Name:			Age:
City & State of Residence:			

Injury 1 Type: Injured		
Name: DESIREE JOY DE LEON		Age: 34
City & State of Residence: MUSCATINE, IA		
Seatbelt Use: Yes	Life saved by Seatbelt:	
Transported To: UIHC	Transported By: AIRCARE	

Injury 2 Type: Injured		
Name: JESSICA MARIE SMITH		Age: 36
City & State of Residence: BUFFALO, IA		
Seatbelt Use: Yes	Life saved by Seatbelt:	
Transported To: UIHC	Transported By: EMS	

Injury 3 Type: Injured		
Name: DONALD WAYNE GATZKE		Age: 61
City & State of Residence: ATALISSA, IA		
Seatbelt Use: Yes	Life saved by Seatbelt: Yes	
Transported To: UIHC	Transported By: EMS	

Injury 4 Type:		
Name:		Age:
City & State of Residence:		
Seatbelt Use:	Life saved by Seatbelt:	
Transported To:	Transported By:	

Injury 5 Type:		
Name:		Age:
City & State of Residence:		
Seatbelt Use:	Life saved by Seatbelt:	
Transported To:	Transported By:	

Injury 6 Type:		
Name:		Age:
City & State of Residence:		
Seatbelt Use:	Life saved by Seatbelt:	
Transported To:	Transported By:	

Motor Carrier Info (IF Large truck is involved)	
Name of Carrier:	
DOT or MCC #:	
City & State of Carrier:	Hazmat Involved?:

Summary:

VEHICLE 1 WAS NB ON MOSCOW AND COLLIDED WITH VEHICLE 2 WHICH WAS WB ON HWY
6 - ALL SUBJECTS INVOLVED IN THE ACCIDENT WERE TRANSPORTED TO THE UIHC -
THIS ACCIDENT REMAINS UNDER INVESTIGATION WITH THE IOWA STATE
PATROL ASSISTED BY: ISP 399, T275, MUSCATINE CO SO, WILTON PD, DURANT
EMS, WEST LIBERTY EMS, AIRCARE

Officer Name: ISP 109- SMITH	Post: 12
Assisted By: ISP 399, T275, MUSCA	

Figure 1. Minimal Crash Report *(16)*

Crash data contains a total of 2360 crashes with 84 variables reported in all active work zones from 2017 and 2018. Those obtained from the crash data set are shown in Table 1. Some variables are similar, either reported in numerical or string values. Variables include driver characteristics, road characteristics, environment characteristics, area information, time and date information, work zone information and crash characteristics. Each variable was considered, but not all variables in the original data set applied the subject of the study.

Table 1 Crash Variables

Variable	**Description**
Bearing	Direction of travel
Captured	Date of incident recorded
Cardinal	Cardinal direction of vehicles
Casenumber	Case number – Iowa DOT
Citybr	Base records city number
Cityname	City that incident occurred
Coroadrte	County road route
Country	Country that incident occurred
County Name	County that incident occurred, name format
County	County that incident occurred, numeric format
Crash Date	Date of crash
Crash Day	Day of week of crash
Crash Key	Crash key – SAVER internal unique identifier
Crash Time	Time of crash (hh:mm)
Crash Year	Year of crash
Crashmonth	Month of crash
Crcomanner	Manner of Crash
Cseverity	Severity of crash
Csurfcond	Surface condition during crash
Cvltwpid	Civil township ID
Darkness	Darkness hours on day of crash
Daylight	Daylight hours on day of crash
Dayofmonth	Day of month of incident
Dnrdstrct	Iowa Department of Natural Resources district
Dnrwlddepr	Iowa DNR Wildlife Depredation Program Area

Table 1 Continued

Variable	**Description**
Dotdstrct	DOT district in which crash occurred
Drugalcrel	Drug or alcohol related
Econtcirc	Contributing circumstances - environment
Fatalities	Number of fatalities
Firstharm	First harmful event
Injuries	Number of injuries
Intclass	Intersection class
Ispdstrct	Iowa State Patrol District
Lanedir	Location tool lane Direction
Latitude	Latitude marker of crash
Longititude	Longitude marker of crash
Lecasenumber	Law enforcement case number
Lighting	Derived lighting conditions
Litdesc	Location tool literal description
Literal	Derived literal description
Locfstharm	Location of first harmful event
Loctoolv	Location tool version
Majinjury	Number of major injuries
Majorcause	Major cause of incident
Mininjury	Number of minor injuries
Next XD Segment	Next XD Segment
Overunder	Crash occurred on overpass or underpass
Possinjury	Possible injury occured
Previous XD Segment	Previous XD Segment
Propdmg	Property damage amount
Ramp	Crash occurred on a ramp
Rcontcirc	Contributing circumstance - Roadway
Road Name	Name of roadway
Road Number	Name of roadway
Road Class	Class of roadway
Road Type	Type of roadway junction/feature
Route	Route
Ruralurban	Crash is located in Rural or Urban area
Schdst101	School District
Season	Season of crash
State	State that crash occurred in
System	Road System
Systemconc	Concatenated system

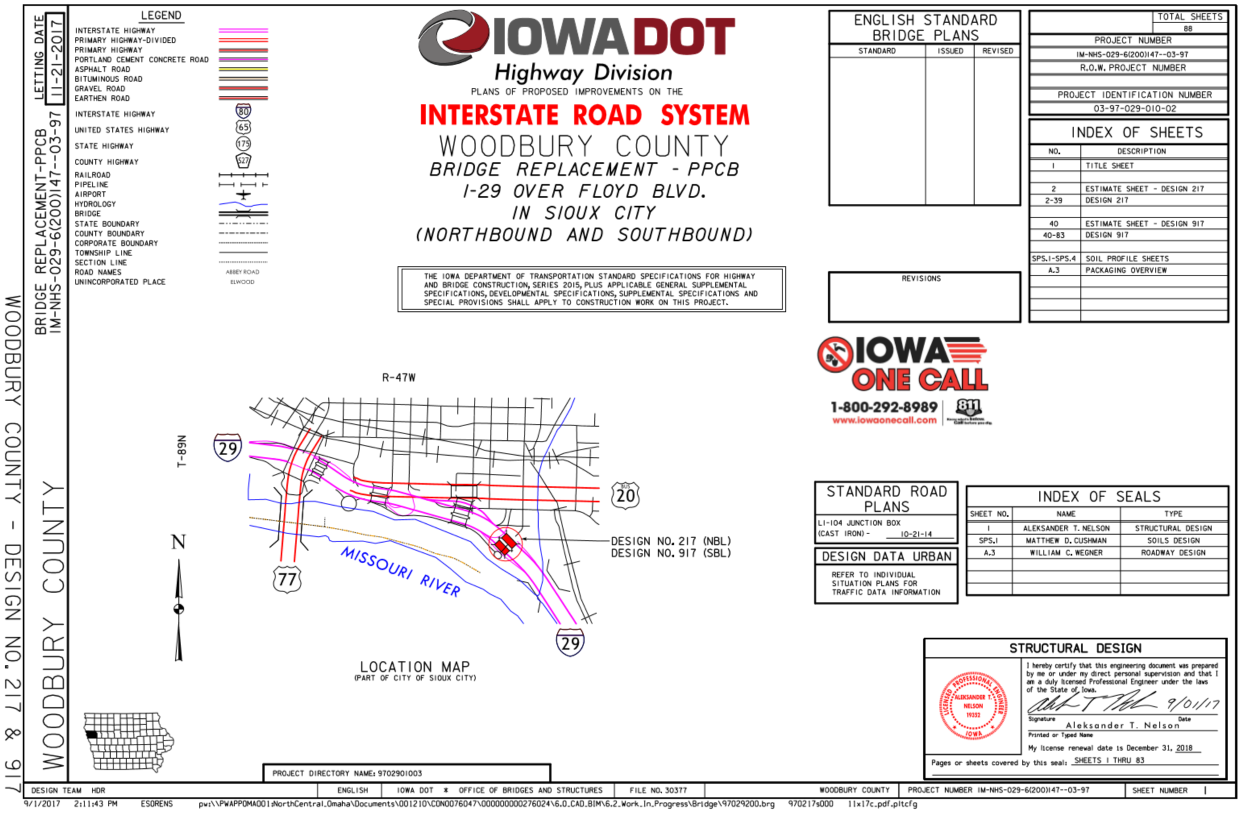
Table 1 Continued

Variable	**Description**
Systemstr	Route in string format with system
Timebin1	Hour of Crash
Timebin30	Minute of Crash
Timebin	Time of day in 2 hour bins
Timeofday	Times split into sections of the day of crash
Timestr	Time string format
Toccupant	Total number of occupants in vehicle
Twnrngsesct	Township, range, section
Unkinjury	Number of unknown injuries
Urbanarea	FHWA urban area code
Vehicles	Number of Vehicles involved
Weather1	Type of weather during crash
Weather2	Secondary weather during crash
WorkZone	Identifier for work zone
Workers	Workers/Law enforcement present
Wz Actvty	Activity in work zone
Wz Loc	Location of crash in work zone
Wz Relate	Crash Reported as work zone related
Wz Type	Type of work zone
XD Seg ID	XD Segment of roadway where crash occurred
XD Group	XD Segment Group

Work Zone Plan Data

The Iowa DOT has a google drive where work zone plans, and other information pertaining to the work zones are accessible. A large majority of work zone plans were not on the drive or available to use, filtering the number of work zones included vastly. Every work zone doesn’t have a pdf available for the plan and those without plans available had to be removed from the study. The drive also had some information included outside of the pdfs that provides interesting information, such as project duration and project cost as well as others. Each pdf was manually looked at and variables were recorded that seemed relevant. Each work zone plan was vastly different depending on location and contractor which resulted in missing data. The original number of work zones considered was approximately 150 work zones but was limited by the number of work zone plans available.

Figure 2 is an example of what a work zone plan may look like, note that documentation in each pdf varies from roughly 30 to 200 pages and not all pdfs follow the exact same format and layout. Figure 2 below shows what the first page of a work zone plan pdf looks like. As mentioned there is some variation, but for the most part, the first page was the most consistent piece of information across all work zone plans. Some notable information provided on this page is: The location of the work zone, the AADT for the most recent years in the work zone, general location and direction of work zone, direction of travel, county and contractor/consulting team.



WOODBURY COUNTY - DESIGN NO. 217 & 917

WOODBURY COUNTY

BRIDGE REPLACEMENT-PPCB
IM-NHS-029-6(200)147--03-97

LETTING DATE
11-21-2017

LEGEND

INTERSTATE HIGHWAY	═
PRIMARY HIGHWAY-DIVIDED	═
PRIMARY HIGHWAY	═
PORTLAND CEMENT CONCRETE ROAD	═
ASPHALT ROAD	═
BITUMINOUS ROAD	═
GRAVEL ROAD	═
EARTHEN ROAD	═
INTERSTATE HIGHWAY	80
UNITED STATES HIGHWAY	65
STATE HIGHWAY	175
COUNTY HIGHWAY	S27
RAILROAD	┼┼┼
PIPELINE	├┤
AIRPORT	✈
HYDROLOGY	～
BRIDGE	═
STATE BOUNDARY	-··-
COUNTY BOUNDARY	--
CORPORATE BOUNDARY	·····
TOWNSHIP LINE	──
SECTION LINE	·····
ROAD NAMES	ABBEY ROAD
UNINCORPORATED PLACE	ELWOOD

IOWA DOT
Highway Division
PLANS OF PROPOSED IMPROVEMENTS ON THE
INTERSTATE ROAD SYSTEM
WOODBURY COUNTY
BRIDGE REPLACEMENT - PPCB
I-29 OVER FLOYD BLVD.
IN SIOUX CITY
(NORTHBOUND AND SOUTHBOUND)

THE IOWA DEPARTMENT OF TRANSPORTATION STANDARD SPECIFICATIONS FOR HIGHWAY AND BRIDGE CONSTRUCTION, SERIES 2015, PLUS APPLICABLE GENERAL SUPPLEMENTAL SPECIFICATIONS, DEVELOPMENTAL SPECIFICATIONS, SUPPLEMENTAL SPECIFICATIONS AND SPECIAL PROVISIONS SHALL APPLY TO CONSTRUCTION WORK ON THIS PROJECT.

ENGLISH STANDARD BRIDGE PLANS

STANDARD	ISSUED	REVISED

REVISIONS

	TOTAL SHEETS
	88

PROJECT NUMBER
IM-NHS-029-6(200)147--03-97
R.O.W. PROJECT NUMBER

PROJECT IDENTIFICATION NUMBER
03-97-029-010-02

INDEX OF SHEETS

NO.	DESCRIPTION
1	TITLE SHEET
2	ESTIMATE SHEET - DESIGN 217
2-39	DESIGN 217
40	ESTIMATE SHEET - DESIGN 917
40-83	DESIGN 917
SPS.1-SPS.4	SOIL PROFILE SHEETS
A.3	PACKAGING OVERVIEW

IOWA ONE CALL
1-800-292-8989
www.iowaonecall.com
811

LOCATION MAP
(PART OF CITY OF SIOUX CITY)

STANDARD ROAD PLANS

LI-104 JUNCTION BOX
(CAST IRON) - 10-21-14

DESIGN DATA URBAN

REFER TO INDIVIDUAL
SITUATION PLANS FOR
TRAFFIC DATA INFORMATION

INDEX OF SEALS

SHEET NO.	NAME	TYPE
1	ALEKSANDER T. NELSON	STRUCTURAL DESIGN
SPS.1	MATTHEW D. CUSHMAN	SOILS DESIGN
A.3	WILLIAM C. WEGNER	ROADWAY DESIGN

STRUCTURAL DESIGN

LICENSED PROFESSIONAL ENGINEER
ALEKSANDER T. NELSON
19352
IOWA

I hereby certify that this engineering document was prepared by me or under my direct personal supervision and that I am a duly licensed Professional Engineer under the laws of the State of Iowa.

9/01/17
Signature Date
Aleksander T. Nelson
Printed or Typed Name

My license renewal date is December 31, 2018

Pages or sheets covered by this seal: SHEETS 1 THRU 83

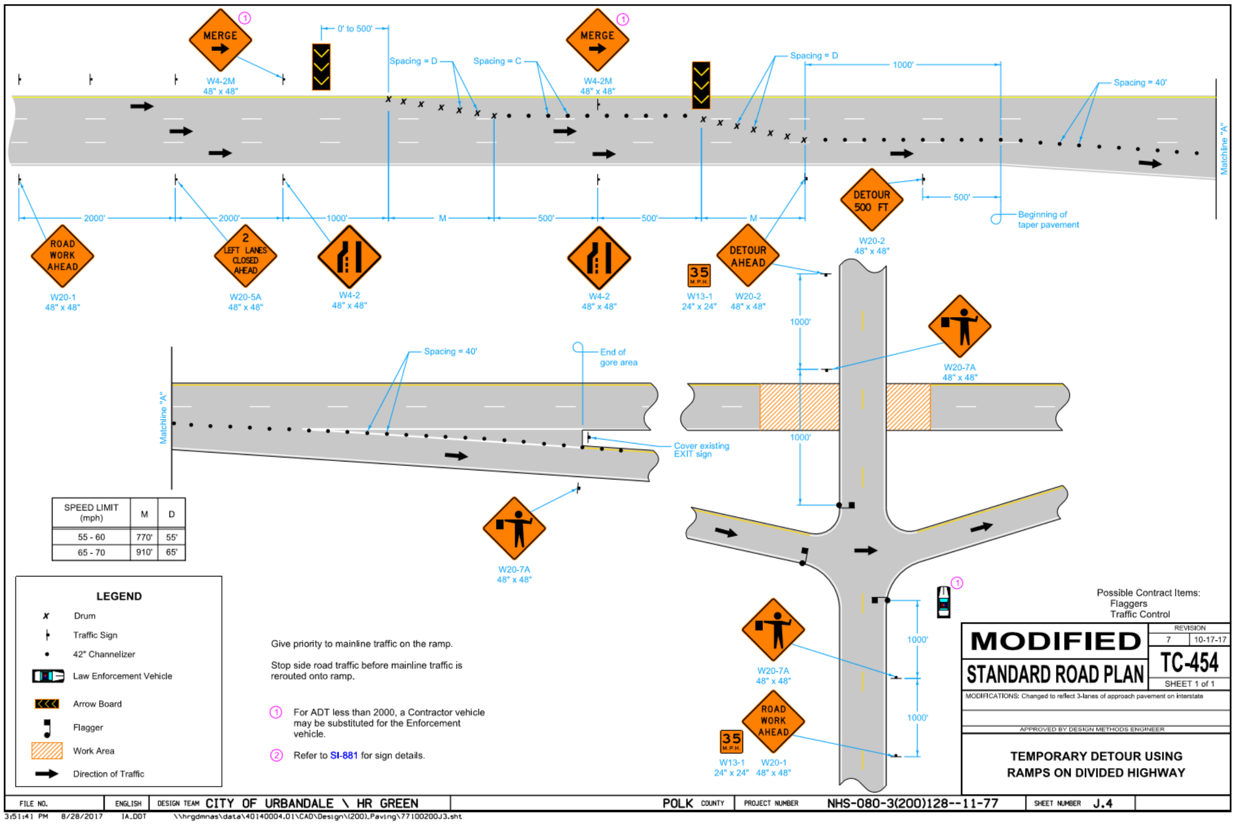
PROJECT DIRECTORY NAME: 9702901003

DESIGN TEAM HDR	ENGLISH	IOWA DOT * OFFICE OF BRIDGES AND STRUCTURES	FILE NO. 30377	WOODBURY COUNTY	PROJECT NUMBER IM-NHS-029-6(200)147--03-97	SHEET NUMBER 1

9/1/2017 2:11:43 PM ESORENS pw:\\PWAPPOMA001:NorthCentral_Omaha\Documents\001210\CON0076047\000000000276024\6.0_CAD_BIM\6.2_Work_In_Progress\Bridge\97029200.brg 970217s000 11x17c.pdf.pltcfg

Figure 2. Work Zone Plan First Page

Figure 3 shows a diagram of a lane closure on a work zone plan. A combination of this diagram as well as others was available in most work zone plans. As seen in the figure some signage as well as distances are shown in one direction of travel. This work zone pdf is a good example of the data set used in the analysis. Not all work zone plans included a diagram comparable to Figure 3. Those without diagrams were removed from the study due to missing data.



SPEED LIMIT (mph)	M	D
55 - 60	770'	55'
65 - 70	910'	65'

Give priority to mainline traffic on the ramp.

Stop side road traffic before mainline traffic is rerouted onto ramp.

1 For ADT less than 2000, a Contractor vehicle may be substituted for the Enforcement vehicle.

2 Refer to SI-881 for sign details.

Figure 3. Work Zone Plan Diagram

After gathering all the pdf work zone plans that were available the data consisted of 55 work zones with 19 variables. Each work zone with plans available had variables recorded (See Table 2). Some of the variables were not considered due to a large percentage of missing data. These variables were project cost, advanced warning distance, and number of work zone signs. Each of these variables had over 25% missing data. Therefore, removing them from the analysis and using more work zones was preferred. Below in Table 2 shows all the variables that were considered in the regression modeling. Variables have a brief description in the far-right hand side of the table under the description tab, including units when necessary.

Table 2. Work Zone Plan Data

Variable	**Levels**	**Description**
Year	Numerical	Year work zone is active
Project.Duration	Numerical	Number of days the work zone is scheduled to be open
Activity	0= Construction 1= Bridge 2= Pavement 3= Traffic Signs	Description of general type of work being done
AADT	Numerical	Average annual daily traffic for the work zone area (veh)
Percent.Trucks	Numerical	Percentage of trucks of total traffic volume in work zone area (%)
Speed.Limit	Numerical	Speed limit in work zone area denoted by signs (mph)
Distance	Numerical	Total mileage of work zone area in ArcGIS used for filtering work zones form all other roadways (mi)
Lanes	Numerical	Number of lanes on roadway (does not account for lane closure)
Signal	0= No 1=Yes	Denotes whether there is a signalized intersection in the work zone area
Curvature	0=No 1=Yes	Denotes curvature in the roadway or use of crossover when one direction of travel is shutdown
Divided	0=No 1=Yes	Divided versus undivided roadway
Crash_Count	Numerical	Predicted value, Total number of crashes in the work zone

INRIX Data

Since 2014, the Iowa DOT has partnered with INRIX to collect data for traffic analytics for the state of Iowa. INRIX data consists of probe data collected in real time from mobile phones, connected cars, trucks, delivery vans, and other fleet vehicles with GPS locator devices recording the average speed of vehicles in specific roadway sections *(11)*. The sensors used to collect INRIX data collected the average speed of vehicles on a section of roadway every minute. This data can be used to generate alerts and maps as well as calculate traffic travel times *(11)*. Each individual work zone makes up a differing number of XD (eXtreme Definition) segments that the INRIX sensors record data on. An XD segment is a segment that covers more miles of a roadway than TMC (Traffic Message Channel) segments, generally with greater granularity and with the ability to adapt more quickly to changes in the road network and the addition of new roads and new markets *(11)*. Data is automatically collected through the apps and purchased by the DOT to analyze driver behavior for safety and traffic flow projects. Data is collected for all major state and US roadways in Iowa

The data makes up 288 points of data for every XD segment every day (105,120 data points per XD segment yearly). In this study computational power was a significant limitation, therefore in order to reduce the total number of data points the INRIX data was reduced to hourly. Variables included in this study are shown below in Table 3. The main reason for including this data set was to compare the average speed of drivers with the speed limit to determine if higher speeds result in more vehicle crashes. Data was collected throughout the entire year as the exact dates which work zones were active was unknown in this study. Variables have a brief description in the far-right hand side of the table under the description tab, including units when necessary.

Table 3 INRIX Data

Variable	Description
XD_Segment	XD segment of roadway
Timebin1	Hour of day
Average.Speed	Average speed of vehicles aggregated every 1 hour (mph)

Data Filtering

The crash data set is a shapefile (.shp) that includes every crash recorded in Iowa as points using longitude and latitude fields. The shapefile format is a geospatial vector data format for geographic information system software. In order to include only crashes in work zones an ArcGIS file containing line segments for each work zone was utilized. ArcGIS is a geographical information software that incorporates maps and other geographic information that is maintained by Esri. One containing 2017 work zones and another containing 2018 work zones. Originally, the ArcGIS file consisted of line segments for each individual work zone in their respective year. The first method employed to merge the crash locations and work zones was to split the work zone line segments from the ArcGIS file into a series of points. Then using longitude and latitude markers for the points a radius was constructed to capture only crashes at a set distance from the work zone. The result was a substantial number of tiny circles that would incapsulate the work zone roadway and all crashes inside the circumference of the circles would be included in the study. Unfortunately, with this approach there was additional crashes included outside the work zone. These crashes occurred on side roads or roads in close proximity to that of the work zone. Determining the correct radius size was a tedious task, as well as it largely impacted the number of crashes included at each work zone.

Fortunately, in early 2019 Tableau created a feature that allowed for points to be filtered by an intersecting line. This allowed all crashes directly on the roadway to be automatically be filtered. Therefore, a buffer of 250 feet was added to the roadways in the ArcGIS file to include collisions off road. Crash data consisted of approximately 100,000 data points for all work zones in Iowa after including the corresponding XD segments. After completing this step all the crashes in work zones in Iowa were all in one data set.

Next the Iowa work zone crash data set needed to be combined with both the INRIX and work zone plan data sets. They were combined by using the two common fields in the data sets, year and work zone. For each data set an inner join was used to keep all matching data from each data set. This ensures that only the work zones that are in the work zone plan data would be used in the regression model. This final data set included over 1 million rows of data including 40 variables. The Crash_Count variable was created using the unique count values for each crash key identifier variable and each year. The resulting final data included 511 crashes throughout 32 work zones from 2017-2018.

CHAPTER 4. METHODOLOGY

Variable Analysis

After combining all data sets, the first step was to analyze the numerical variables in the combined data set. Table 4 is the descriptive statistics for the data set used in this study. The table includes all numerical variables used in the final data set. Displaying the mean, standard deviation, median, minimum and maximum values for the data. Note that if the final regression model is used to predict the number of work zone crashes predictions are most accurate when all variables are somewhere between the minimum and maximum values provided in the table below.

Table 4. Descriptive Statistics

N = 939,958 observations in 32 Work Zones						
Variable	**Definition**	**Mean**	**Std. Dev**	**Median**	**Min Value**	**Max Value**
Crash_Count	Number of work zone crashes	37.23	21.30	33.00	1	75
AADT	Average annual daily volume	3888.99	29167.3	38600	3700	98500
Average.Speed	Vehicle speed, aggregated at 1 hour	63.23	8.11	65.7	2.00	88.12
Distance	Work zone length (mi)	7.76	3.71	6.56	0.82	15.83
Lanes	Number of lanes	2.49	0.81	2.00	1	5
Percent.Trucks	Percentage of trucks of total volume	19.11	10.07	17.00	2.00	35.00
Project.Duration	Estimated days active	144.49	102.21	115.00	7	540
Speed.Limit	Speed limit (mph)	52.49	7.68	55	35	65
Vehicles	Number of vehicles involved in crash	1.62	0.64	2.00	1	4

As previously mentioned before the Average.Speed and Speed.Limit variables seem to be significantly different. With Average.Speed being higher than the Speed.Limit both in averages and in median values. Another concern is that the predictor variable Crash_Count appears to be inflated due to a majority of work zones having extremely high or low values for Crash_Count. In the data there are more work zones that have low crash values and high values, but not many work zones have values in or around the mean and median values.

The resulting plot, Figure 4 shows the crash distribution across all work zones in the study. Each work zone was either used in 2017 or 2018. No work zone overlapped both years. As observed in the figure there is a substantial amount of work zones that have extremely low crash numbers, crash values less than 10, while only a few have extremely high values, those work zones with more than 30 crashes. Out of the 32 total work zones used in this analysis, 56.25% of the work zones have less than 10 crashes while only 15.625% have over 30 crashes.

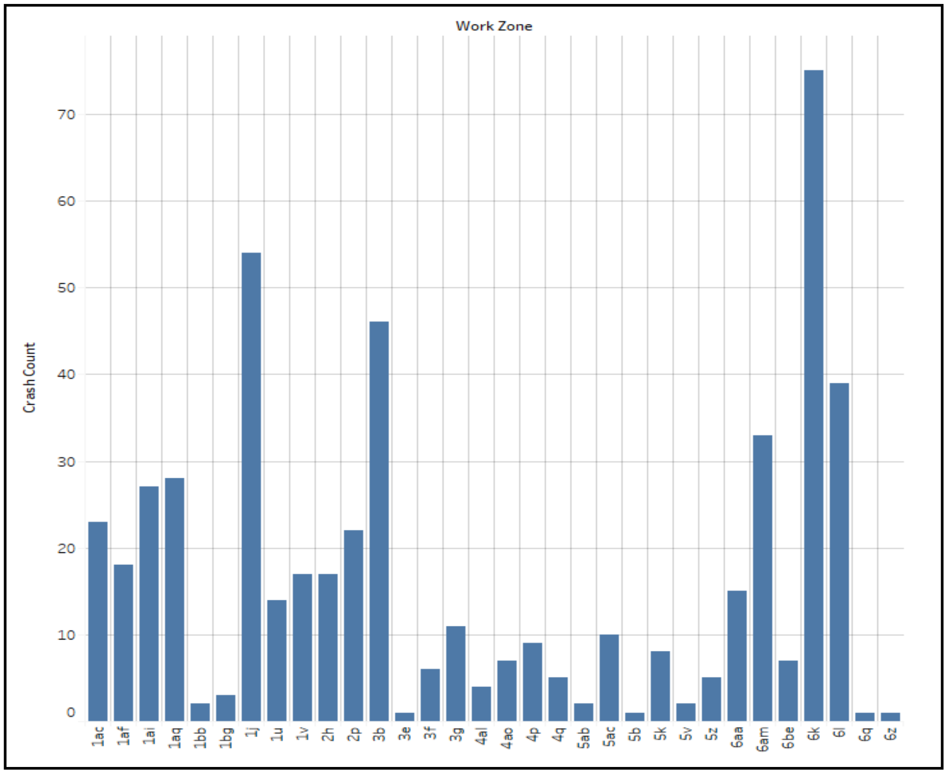


Figure 4. Work Zone Crash Distribution

Each numerical variable was analyzed for correlation with the predicted variables as well as other independent variables to consider for multicollinearity. Out of the 8 total descriptive variables above in Table 4, five were transformed to improve the correlation to the number of crashes. Figure 5 illustrates the correlation between the variables from the data set. Note that some variables are replaced with transformations that yielded better correlation. Only simple transforms were considered, such as natural logarithm, reciprocal, quadratic, square root and cube root. Out of the transforms considered natural logarithm, reciprocal and quadratic were the only transforms that improved correlation by a significant amount. A natural logarithm transform improved correlation for AADT and Lane". While Average.Speed and Vehicles improved using a reciprocal transformation and "Distance" improved using a quadratic transform. These transforms can be seen below in Figure 5. The largest correlations between independent variables and Crash_Count are with Distance^2, Speed.Limit, Lnaadt and lnLanes at roughly .50.

A few variables show signs of multicollinearity, while the most important takeaway from the study is to predict the number of crashes a few variables with multicollinearity are noted. Independent variables with 0.35 correlation were analyzed using interaction terms in the regression modeling. Lnaadt also has a high correlation with Speed.Limit, due to most high-volume traffic areas also being high speed areas, like interstates and US highways. Another source of independent variables being correlated is between the variables Distance and Percent.Trucks which is the highest in the data set at 0.72. The high correlation between these variables appears to due to work zones with long distances have a higher percentage of trucks. Initially, this would imply that large work zones are more likely to be on high speed roadways though because Distance and Speed.Limit have low negative correlation it does not appear to be true. In general, there is no extremely high correlations for this data set with the highest being between Distance^2 and Percent Trucks.

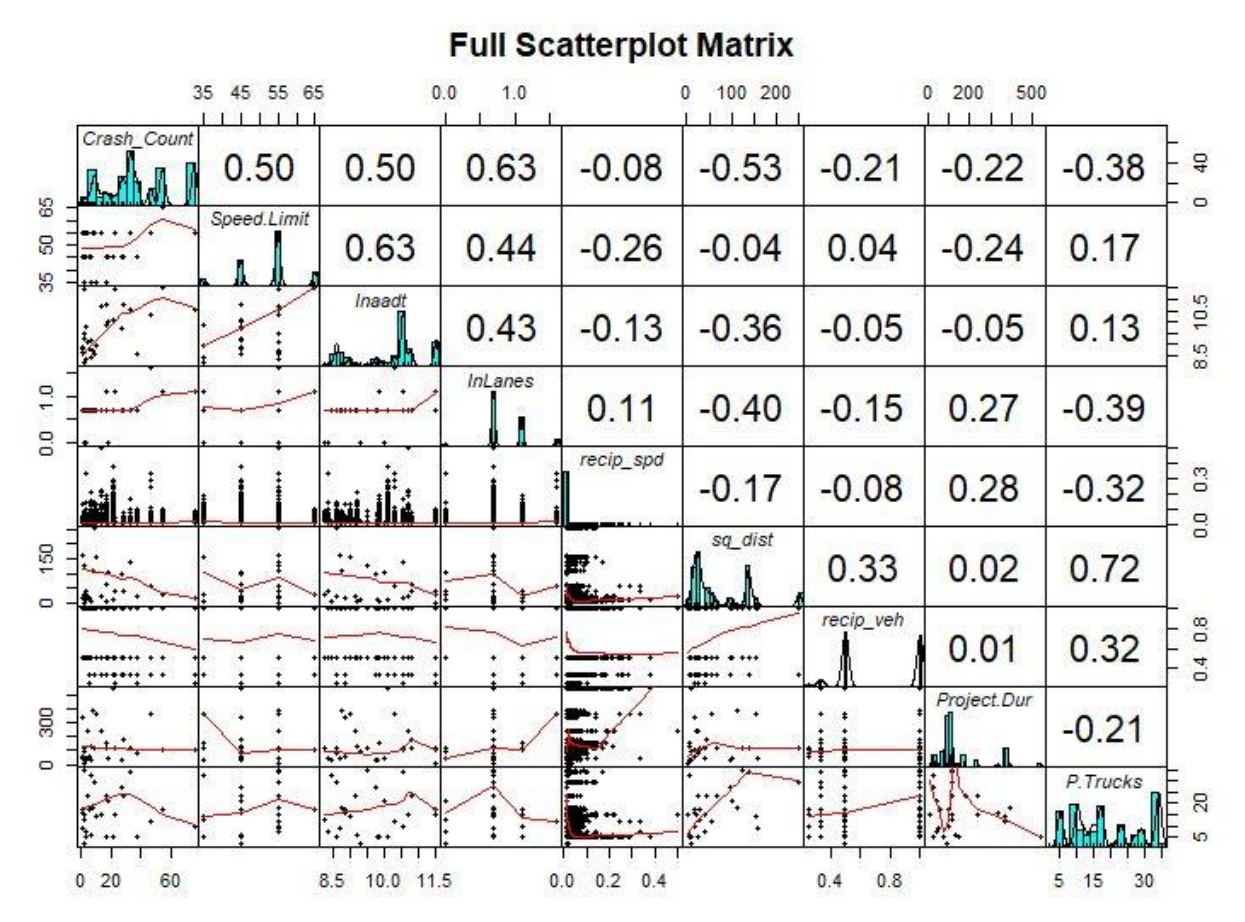


Figure 5. Correlation Matrix and Scatterplot

Statistical Modeling

Choosing the correct analysis to model the data was an important step in this study. The first iterations of modeling were accomplished using a basic linear regression model. Using the 4 linear regression modeling assumptions: Linearity, Homoscedasticity, Independence and Normality it was determined that the data would not be represented well with a linear model. Since the crash and INRIX data sets were both count data sets both Negative Binomial (NB) regression and Poisson regression were considered. Research was done to determine which model better applied to the data. NB regression applied better to the data as the variance is greater than the mean. The NB regression includes an extra parameter that is used to model the over-dispersion. NB was a popular method used in previous studies mentioned above in the Literature Review Chapter.

The NB regression is used in this paper to estimate the crash frequency and the impacting factors. The model can be written as;

$$Y_i \sim NB(\mu_i, \alpha) \tag{1}$$

where Y_i is the number of crashes in Iowa in a work zone $i, (i = \ 1, \dots, n)$, μ_i stands for the mean crash frequency, and α is the over-dispersion parameter. It is assumed that μ_i is a function of explanatory variables such that:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_k + \beta_{k+1} Ln(Exposure_i) \dots) \tag{2}$$

where x_{ij} represent the jth variable in event i. $\beta_0, \beta_1, \dots, \beta_{k+1}$ is a vector of regression parameters. Since the number of incidents is count data, to make it comparable between different events, the $Ln(AADT_i)$ is devised as the offset variable in the model. Where the offset is the maximum number of crashes that could have possibly occurred. Which is limited by the total number of vehicles driving in the work zone section. Significant variables at a level of $\alpha = 0.05$ were kept in all modeling stages of the study.

Model Selection Process

The first attempt at a NB model used an exhaustive search beginning with the null model. The null model included only the y-intercept and the offset variable Lnaadt. The full model included every applicable variable in the data set. Utilizing both forward and backward elimination the exhaustive search program finds the best model AIC (Akaike information criterion) possible between the intercept and full models. Choosing AIC discourages the model from overfitting as the number of variables in the data set is large and we likely will increase the goodness of fit by increasing the number of variables. The resulting model contained 15 variables out of a starting 45, resulting in a significant reduction of variables. Though several of the remaining variables were categorical with many categories resulting in a model with 45 possible variables and categories. While the model reduced many variables, other methods were attempted to determine a simpler model solution.

The next method to applied produce a simpler model was a lasso or elastic net regression. Ridge regression was never considered since the model doesn't remove variables, but only decreases the less significant variables to an extremely small value. The first model attempted was the lasso model as it was simpler as alpha = 1. The resulting model didn't to reduce as many variables as the previous exhaustive search model accomplished ending with 33 variables out of the total 45. One of the difficulties with this model was that it was computationally heavy with our large data set. The model could never reach the minimum error values before reaching the iteration limit and increasing the iteration limit simply made the code crash or never complete running after weeks. This result was the same for trying other alpha values as the elastic network regression attempts to do. This was also challenging for finding the best alpha value as cross validation was also an extremely taxing code with the large data set. After trying these two methods with no great success the last method was attempted.

The final methods attempted were the same as before, being exhaustive search and elastic net regression, but applied after simplifying the starting variables by using the random forest importance plot shown below in Figure 6. The plot lists the most significant variables in descending order. Using both plots can help simplify variable selection. Selecting a cutoff point is dependent on the user, but most of the time there is a noticeable “jump” or “drop-off” in values on either plot that are good quality points to choose. In this study the cutoff point was chosen at the 11th variable on the left plot as there appears to be a substantial drop off there. As well as the top 11 variables from the right were included, most of which were the same, but appear in different orders.

After the 10 variable model was decided on the variable transforms previously mentioned were applied as well as a few interaction terms. Four interaction terms were applied all of which had the highest correlation in the correlation matrix above (See Figure 5). These interactions include lnLanes*Speed.Limit, Percent.Trucks*Distance, Lnaadt*lnlanes, and Lnaadt*Speed.Limit. The resulting variables were again tested in a Lasso and Elastic net with similar issues as before where the model could not reach its minimum error values and timed out at the iteration limit. Therefore, the model was run in an exhaustive search model again resulting in the final model.

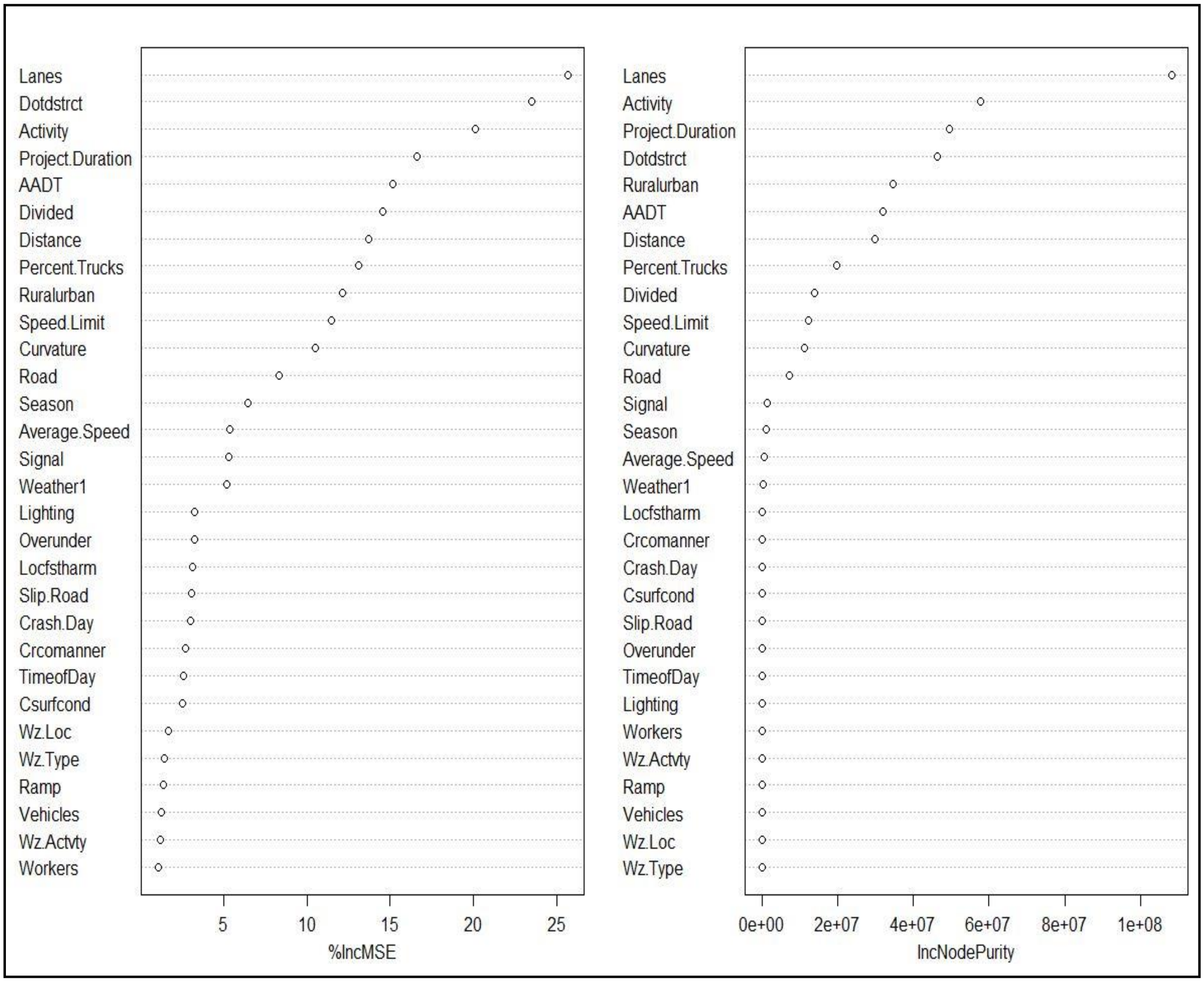


Figure 6. Random Forest Importance Plot

CHAPTER 5. ANALYSIS

Final Model Results

As seen in Table 1, there was a large number of variables included in the crash data set. Due to computational power some of the variables were not considered in the final modeling section. Table 5 shows all variables from the crash data set that were included in the final model selection process. Out of the original 84 variables from the crash data set, 15 were included in the final modeling process. All variables in the table are categorical. Variables with an extreme number of categories were not included in the model. One example is a variable that included the roadway the crash occurred on. As there were many categories with no obvious way of simplifying the variable was excluded. Variables have a brief description in the far-right hand side of the table under the description tab. There is no missing data from this data set. Levels not included in the table below include: Other, Unknown and Not Reported.

After the model selection process was completed the final NB model was created. Table 6 shows the variables kept in the final model. All variables in the final regression model were significant at an $\alpha = 0.001$ level. To reduce the number of variables, an exhaustive search and random forest importance plot were used as mentioned above. The model includes 12 variables of which 7 are numerical and 5 are categorical, as well as 4 interaction terms.

Table 5. Final Crash Data Variables

Variable	Levels	Description
Locfstharm	1= On Roadway, 2= Shoulder, 3= Median, 4= Roadside, 5= Gore, 6= Outside Trafficway, 7= Parking Lane/Zone, 8= Continuous Left Turn, 9= Separator	Location of first harmful event
Csurfcond	1= Dry, 2= Wet, 3= Ice/Frost, 4= Snow, 5= Slush, 6= Mud/Dirt, 7= Water, 8= Sand, 9= Oil, 10= Gravel	Surface condition during crash
Lighting	1= Daylight, 2= Darkness, 3= Dawn, 4= Dusk	Derived lighting conditions
Overunder	1= Overpass, 2= Underpass	Crash occurred on overpass or underpass
Ruralurban	R= Rural, U= Urban	Crash is located in Rural or Urban area
Workers	1= Workers Only, 2= No Workers, 3= Workers and Officer, 4= Law Enforcement Only, 5= No One Present	Workers/Law enforcement present
Wz.Activity	1= Construction, 2=Maintenance, 3= Utility	Activity in work zone
Wz.Loc	1= Before Work Zone Warning, 2= Advance Warning Area, 3= Within Work Zone, 4= Moving Work	Location of crash in work zone
Wz.Type	1= Lane Closure, 2= Lane Switch/Crossover, 3= Work on Shoulder/Median, 4= Moving Work	Type of work zone
Crash.Day	1= Sunday, 2= Monday, 3= Tuesday, 4= Wednesday, 5= Thursday, 6= Friday, 7= Saturday	Day of week of crash
Season	0= Winter, 1= Spring, 2= Summer, 3= Fall	Season of crash
TimeofDay	0= Early Morning, 1= AM Peak, 2= Mid-Day, 3= PM Peak, 4= Nighttime	Times split into sections of the day of crash
Weather1	1= Clear, 2= Cloudy, 3= Fog/Smoke/Smog, 4= Freezing Rain, 5= Rain, 6= Sleet/Hail, 7= Snow, 8= Blowing Snow, 9= Severe Wind, 10= Blowing Sand/Soil/Dirt	Type of weather during crash
Ramp	1= Ramp, 2= Mainline	Crash occurred on ramp or mainline
DotDstrict	6	DOT district in which crash occurred, level coincides with District

Table 6. Summary Results of Final Model

Variable	Coeff.	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	3.008	0.036	89.54	2.00E-16	***
Lnaadt	0.2579	0.0033	79.03	2.00E-16	***
Dotdstrct2	1.918	0.0027	702.23	2.00E-16	***
Dotdstrct3	-1.428	0.0013	-1136.41	2.00E-16	***
Dotdstrct4	1.757	0.0028	634.38	2.00E-16	***
Dotdstrct5	-0.4030	0.0012	-340.66	2.00E-16	***
Dotdstrct6	0.9664	0.0013	757.88	2.00E-16	***
LnLanes	-0.6327	0.0018	-35.60	2.00E-16	***
Divided1	0.7981	0.0013	593.55	2.00E-16	***
Percent.Trucks	-0.005278	0.00016	-33.69	2.00E-16	***
Activity1	-0.008353	0.0012	-6.78	1.22E-16	***
Activity2	0.08692	0.0016	54.57	2.00E-16	***
Activity3	0.04287	0.0017	24.75	2.00E-16	***
Speed.Limit	-0.2637	0.00075	-352.57	2.00E-16	***
Project.Dur	-0.001062	0.0000040	-262.99	2.00E-16	***
Distance	0.9251	0.0010	905.85	2.00E-16	***
Sq_dist	-0.04613	0.000055	-845.18	2.00E-16	***
Curvature1	0.3070	0.00094	327.01	2.00E-16	***
RuralurbanU	-0.002187	0.00023	-9.43	2.00E-16	***
lnLanes:Speed.Limit	0.1629	0.00022	745.97	2.00E-16	***
Percent.Trucks:Distance	-0.003159	0.000020	-161.91	2.00E-16	***
Lnaadt:Lnlanes	-0.6007	0.0021	-281.31	2.00E-16	***
Lnaadt:Speed.Limit	0.01052	0.000069	151.83	2.00E-16	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Null deviance: 119587367 on 939957 degrees of freedom					
Residual deviance: 1784835 on 939935 degrees of freedom					
AIC: 52330256					

The result was a model with 23 terms, including the y-intercept, all categories for each categorical term and interaction terms. The null model has a residual deviance of 119,587,367, while the final model has a residual deviance of 1,784,835 with a loss of 22 degrees of freedom. Deviance is a measure of goodness of fit, where high numbers of deviance implies a bad fit. This indicates that the final model has a significantly better fit than the base model. The residual deviance is higher than previous models used in the study. This is likely due to overfitting to the data set by including additional variables. The AIC value from the suggested model is higher than the more complex model. This suggests that the final model will perform better on other work zone data in the future. The final model proposed contains a significantly smaller number of variables. This allows for easier application to real world data in comparison to the previous models tested, which was one of the main goals of the study.

One of the key variables that did not make the final model was the Average.Speed variable gathered from INRIX data. This variable inflated the total number of data points in the study. This result is likely due to the inaccuracies in time of operation in each work zone. As of when the study was conducted the exact dates of operation was not known, only an estimated number of days that the project was operational was known information. The data used for average speed was the speed at each XD segment throughout the entire year. Due to this data restriction it is likely that throughout most of the year the work zone was not operational and therefore speeds were much higher than the speed limit posted during work zone operation. This variable should be analyzed in future work when work zone dates of operation are more accurately reported.

Another variable of interest that was included in the final model was the DOT district that the work zone was located. Originally, it was hypothesized that district 1 and district 6 would be the most impactful on the predicted number of crashes due to the large population density in these areas. Based on the final model a work zone in district 2, district 4 and district 6 are predicted to have 6.8, 5.8, and 2.6 times more crashes than a work zone in district 1 assuming all other factors are held constant. While districts 3 and 5 reduced the predicted number of crashes by 4.17 and 1.5 times respectively while holding all other factors constant.

Surprisingly, the type of activity the work zone was conducting did not have a significant impact on the number of crashes predicted. The activity only effected the by around +/- 1.00 times the number of crashes when compared to the default category “construction”. Where paving operations and traffic signals work increased the number of crashes by 1.1 and 1.05 times respectively and Bridge construction decreased the number of crashes by a factor of 1.01 times while all other factors were held constant. In addition, it was unexpected that paving operations would increase the number of crashes while bridge construction would decrease them, as compared to construction work zones. The original was that due to a bridge roadway already having a restricted amount of space, the addition of a work zone would lead to more congestion and slow speeds resulting in more rear end collisions.

While the previously mentioned variables were significant, they are not the most important variables for crash predictions in this study. The first impactful variable analyzed was the AADT of the road section. It was hypothesized that the more vehicles traveling on a roadway daily, the more likely that a crash is to occur in that section. Due to this logic it would be reasonable for AADT to be one of the most impactful variables in our data. The AADT for this data set ranged from 3700 to 98,500 vehicles. As this coefficient is a positive value in the model it will increase the number of crashes for any value in this range. When applied to the NB model above the number of crashes in a work zone increased by 1.99 times when the AADT proceeded from the minimum value 3700 vehicles to the maximum value 98,500 vehicles, all other factors were held constant. The results of this study matched that of previously mentioned study by Chen and Xie in 2016.

The variable with the highest correlation to the predicted crashes as mentioned before was the number of lanes. The minimum value for number of lanes was 1 in the data set and the maximum number of lanes was 5. An expected result was as the number of lanes increased so would the number of vehicles. This would lead to more vehicle crashes as mentioned above when analyzing AADT. Another theory was that the number of lanes could potentially decrease the number of crashes. As the number of available lanes to non-available lanes ratio would be higher, therefore less vehicles would need to merge last second resulting in fewer sideswipes. This was not the case to be the case for this data set. Increasing the number of lanes increased the total number of crashes using the model above while holding all other variables constant. The predicted number of crashes increased by 1.24 times when the number of lanes increased from 1 to 5 lanes on each direction of traffic. This result was similar to the one found in the study previously mentioned by Kononov, Bailey and Allery in 2008.

Another variable with high correlation was the speed limit and the number of crashes in the work zone. The minimum value recorded for the speed limit variable was 35 mph while the maximum number was 65 mph. Overall, this variable didn’t make a huge impact on the predicted number of crashes. The predicted number of crashes increased by 1.05 times when the speed limit was 65 mph compared to when the speed limit was 35 mph. All other variables were held constant during the analysis. Speed limit was one of the last variables included after using Figure 6. Random Importance Plot. Therefore, it would make sense that speed limit didn’t have a large impact on the number of crashes predicted. In comparison to the variables previously mentioned, changing the speed limit had the least impact on number of crashes. Though the impact that speed limit had on the number of crashes in a work zone was minor, the results are like those reported earlier by Chambless in 2002.

The second highest correlation to predicted crashes was the length of the work zone denoted in the data set as distance. This is most likely due to long work zones having a longer exposure to vehicles. Meaning the likelihood of a crash occurring increases because vehicles have more roadway to traverse. The correlation in the Figure 5 is negative indicating that the longer the work zone the less crashes occur in the work zone. In the final model equation both the Distance^2 and the interaction between percent trucks and distance both have negative coefficients, while the distance coefficient itself is positive. The maximum value for work zone distance was 15.82 miles while the minimum distance was 0.82 miles. While keeping all other factors constant increasing the distance variable from its minimum to maximum values increases the number of crashes by 9.6 times. The result is expected as it is reasonable to assume that increasing the area of a work zone would increase opportunity for a crash to occur. However, the scale that the number of crashes increased based off changing the distance variable is much higher than the other variables previously mentioned. In addition, due to the Distance^2 being included in the model, the distance follows a parabolic shape. Due to this, the maximum number of crashes does not occur at the maximum work zone distance. It occurred at work zones with approximately 10 miles in length. At this inflection point the number of crashes is 47.1 times higher than the minimum number of crashes while all other factors are held constant. The scale at which the distance affects the number of crashes appears to be much higher than initially expected. The cause of the extreme increase in crashes may be due to more rear-end collisions occurring caused by extreme congestion in long distance work zones.

Another notable variable included in the final model was whether the roadway was divided or not. In the model and data set, a value of “1” for this variable denotes that the roadway was divided and a value of “0” denotes that it is not divided. From the final model above, work zones with divided roadways had 2.22 times more crashes than those that were not divided when all other variables were held constant. This result was unexpected, it was hypothesized that roadways that were not divided would potentially seem more head on collisions. This would theoretically be enhanced when a work zone reduced the number of lanes and/or lane width on each side of the roadway. A likely solution is that divided roadways have a much larger volume and number of lanes that lead to more crashes.

One concern with work zone safety is visibility. Semi-trucks are much larger than normal passenger vehicles and can provide obstructions for other drivers as well as blind spot challenges from the truck drivers perspective. With the addition of work zone obstructions, it could be possible that the number of trucks that pass-through work zones impact the number of crashes in a negative way. In the data set the minimum value that the percentage of vehicles were trucks was 2% and the maximum was 35%. When the percentage of trucks increased from 2% to 35% using the model above, the number of crashes decreased by 1.30 times while keeping all other variables constant. This result implies that work zones with less trucks have more crashes. This could be due to a combination of factors. One could be that they tend to know of work zone locations better than most of the public and avoid them for faster routes. Another could be that they are more strictly following the speed limits and other safety protocol. While several other factors contribute the overall result is that the percentage of trucks increasing does not negatively affect the number of crashes in the work zone. This result was contradictory to that of the study done by Daniel, Dixon and Jared in 2000. More work should be done using this variable to analyze the impact that trucks have on work zones.

Another variable in the final model to look at in depth is the estimated project duration in days. It was hypothesized that increasing the number of days a work zone is open would also increase the number of crashes, simply because there would be more volume of traffic in the time period. However, the work zone may not have as many operation hours per day or even be open every day like that of work zones that are finished in much shorter days. Not to mention that work zones that are finished more quickly may have many more workers and vehicles as well as other distractions that could affect driver responses negatively. Using the model to estimate crashes while keeping all other factors constant, the resulting number of crashes decreased by 1.76 times when the project duration of the work zone went from the minimum value of 7 days to the maximum number of 540 days. As previously mention this is due to a combination of factors most likely due to the hours of operation, whether at low volume hours such as nighttime or a low number of hours during the day. The road sections also may have been completely closed off from the public at certain times during construction.

The curvature variable was recorded in the data set to denote if the work zone occurred in a roadway section that was primarily curved or that included some sort of crossover. During these types of work zones line of sight can be a major issue for drivers and could potentially lead to an increase in number of crashes. In the data set a value of “1” denotes that there is a noticeable curvature in the roadway and/or there is a cross over present while a “0” denotes that there is not. From the model above, while keeping all other factors constant it was found that having a curved roadway section or crossover in the work zone increased the number of crashes by 1.36 times compared to straight roadways with long lines of sight.

The final variable included in the final model was whether the roadway was in an urban or rural area as denoted by the crash data set. The findings show that this variable has a very small impact to the number of crashes. When the location of a work zone changes from urban to rural there is a 1.002 times increase in the number of crashes in the work zone. The overall change was very slight and for most work zones in Iowa it didn’t change the number of crashes by +/- 1. While extremely small, the result differs from previously mentioned study by Ullman in 2018. More research should be done using the rural and urban variable to determine the effects it has on work zones. In addition, more data could be incorporated from other states urban and rural work zones.

Model Validation

After examining some key variables, analyzing the resulting predicted and actual crash values was done to determine the accuracy of the model. Each value printed above the bars is the number of crashes from the data set that occurred, while the percentage was calculated using:

$$Calculated\ Percentage = \left(\frac{Predicted\ Crashes}{Actual\ Crashes}\right) * 100 \qquad (3)$$

The color scheme in Figure 7 denotes whether values were over predicted, under predicted or acceptable. Values in red were determined to be over predicted and values in blue were determined to be under predicted. Percentage values between 90% and 110% were deemed to be in the good/acceptable range, these values are shown in grey. As seen in Figure 7, 17 out of 32 or 53% of crashes were deemed as acceptable while 29% were under predicted and 18% were over predicted. As seen in the Figure 7 in addition to the numbers previously mentioned, the model struggles to predict work zones with a small number of crashes. All work zones that are denoted as over predicted or under predicted in the model are work zones with less than 10 crashes. Except for work zone 1af, 6aa and 3g, which have predicted values of 18, 15 and 11 respectively.

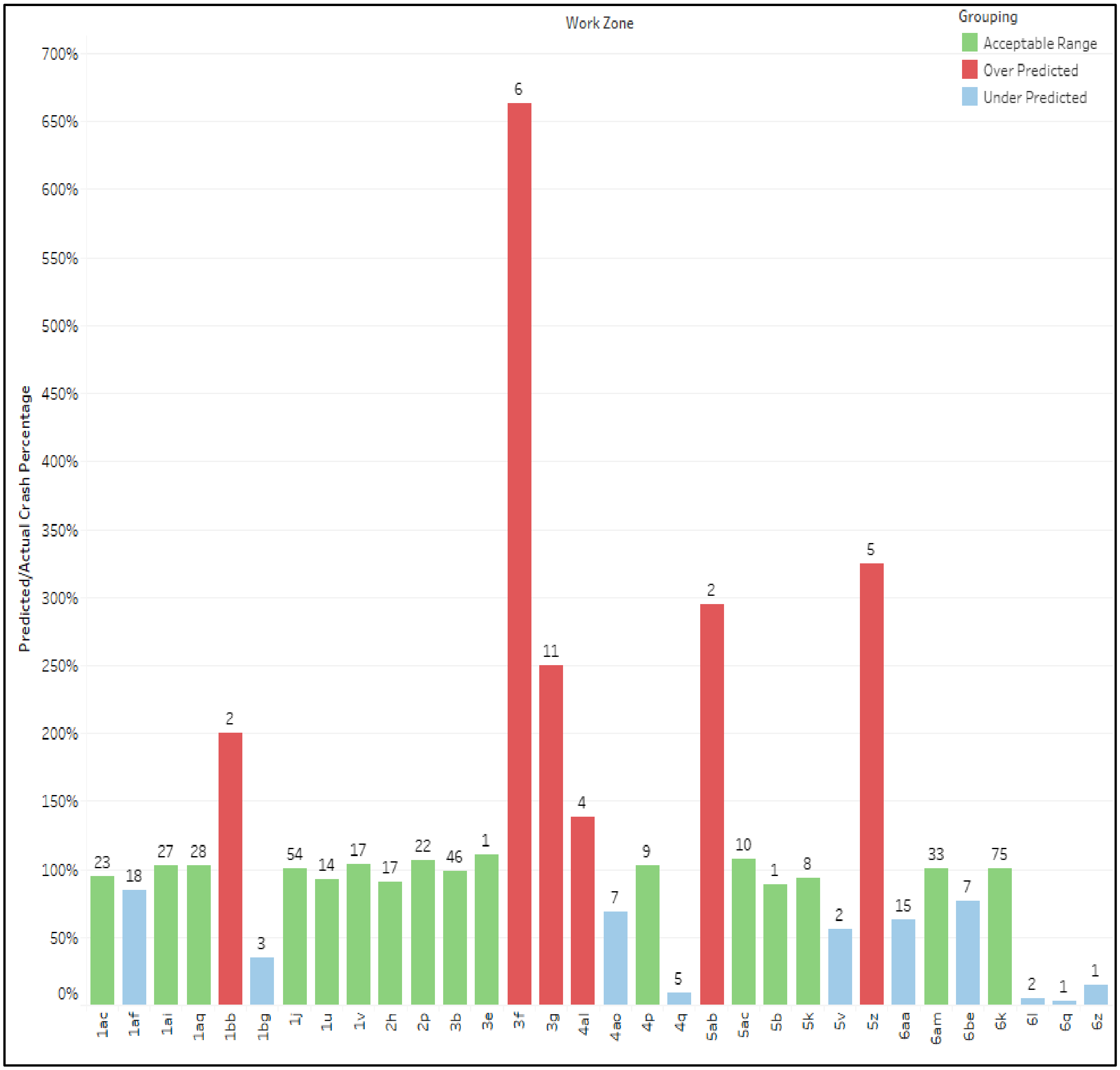


Figure 7. Predicted to Actual Crash Ratio Plot

CHAPTER 6. CONCLUSION

Key Factors Analysis

This study intended to help predict the number of crashes and determine factors that negatively affect crashes in order to mitigate them in the future. This research identified key variables that contribute to work zone collisions. The NB model was applied to determine which variables significantly impacted crashes. The 15 variables in Table 6 were determined to best represent factors that affect work zone crashes in Iowa. Note that all other variables are held constant in the analysis. Below is an overview of key factors discovered in the model in order from most impactful to least impactful:

- Work Zone Length: Increasing the work zone length (mi) increased the number of crashes by 47.1 times from its minimum value to maximum value. Followed a parabolic shape with the worst performing distance being 10 miles.
- DOT District: Values were compared to district 1 as the reference value. Districts 2,4 and 6 have 6.8, 5.8 and 2.6 times more crashes respectively in work zones than in district 1. While districts 3 and 5 have 4.17 and 1.5 times less work zone crashes than district 1 work zones.
- Divided Roadways: Work zone roadways that were divided saw an increase in work zone crashes 2.22 times more than the non-divided counterparts.
- AADT: When increasing the volume of vehicles from the minimum value of 3700 to the maximum value of 98,500, the number of vehicle crashes increased by 1.99 times.
- Estimated Project Duration: As the number of estimated days for project completion increased from the minimum number of days to the maximum number in the data set, the number of work zone related vehicle crashes decreased by 1.76 times.

- Road Curvature: Work zones that include a majority of curved roadway sections or crossover sections increased the number of vehicle crashes by 1.36 times their straight roadway counterparts.
- Percent Trucks: When the percentage of trucks went from its minimum value of 2% to its max of 35% the number of work zone crashes decreased by 1.3 times.
- Number of Lanes: When the number of lanes in one direction increase from 1 lane to 5 lanes the number of work zone crashes increased by 1.24 times.
- Work Zone Activity, Speed Limit, and Rural/Urban: All three mentioned variables had small impacts on the number of crashes. Each of which was less than 1.1 times increase/decrease in the number of crashes.

The other important conclusion that this study incorporates is the accuracy of predicting work zone crashes. As seen in Figure 7 the model predicted a large percentage of crashes accurately but struggled to predict work zones with a small number of crashes. There are two possible insights that come with using this model. First, work zones with low crash frequency are not as significant of a concern, therefore the model works well. Second, if the model cannot predict all crash frequency levels accurately it is not a good model and should be improved. As mentioned below a large portion of data available to work zones is either inaccessible or inaccurate. It is recommended in the near future a similar study should be performed with improved data to increase the accuracy of predicting crashes at all levels.

Data Limitations

One of the clearest limitations of the data was the accuracy of when work zones are active. While the Crash and INRIX data sets are currently accurate work zone data is not as readily available or as precise. As currently there is no database the has all information on work zones. All work zone data was collected manually using work zone pdf files. As not all work zone plans are created by the same affiliation some information was challenging to find for each individual work zone.

One such variable was the dates of operation as the dates are not always recorded on the work zone documentation. For this studies analysis, the crash data was collected for the entire year that the work zone was active. This is clearly not the case and more accurate dates of activity could drastically improve the performance of the model as well as reduce the number of data points. One potential fix for future work would be to not include any crashes that occurred in winter months. This approach would assume that the work zones would not be active in this time period. While it is unlikely that workers would be present during these months, there may be equipment or lanes closed still during this time. Another approach would be to use dates recorded in the work zone pdf noted by shoulder work or drainage work. These dates may not have been the exact starting points but could be used as a rough estimate for the starting dates. Therefore, it could reduce a large portion of data. The best possible solution would be to create a robust automatically reported work zone database. Where contractors or DOTs would be required to input certain information that would be compiled into one data set.

Another area of inaccuracy is in the distance of the active work zone. Currently the value recorded for the work zone distance was taken from the ArcGIS work zone file. While this most likely is a somewhat accurate representation of the entire work zone area, in this study each individual work zone phase was not factored into the equation. Meaning, that in the current set-up crashes were most likely included that were in a different phase of the work zone. Either one that was already completed or yet to be started. In the near future work zones will have GPS locations at the starting and ending points of the work zone to provide a much more accurate distance calculation. This will also help track when a work zone is going through a number of different phases in its life cycle. This GPS locator may also improve the starting and ending dates and time of day of each work zone. Knowing the start and end dates of each individual work zone will allow for the INRIX data to be filtered down much more as well as more accurately report speed data for when the work zone was active. Reducing the size of the data may also allow for better computation speeds for other variable selection methods potentially outputting a better model using different model selection criteria.

Data size is another limitation that this project dealt with. Two years of work zone data provided over 100 work zones to work with, but after filtering out missing data only 32 work zones were analyzed in the study. Increasing the number of years to approximately 5 years would likely provide more work zones to analyze as well as increase the quality of data in the years to come. As mentioned above this would most likely result in a similar data size as the one used in this study, as increasing the accuracy of work zone locations would likely reduce the overall number of data points included. Another solution to this limitation would be to work with other state DOTs to increase the number of work zone data available. A result from this would allow the model to be used more accurately throughout the entire United states instead of in only Iowa.

Future Work/ Recommendations

The result of this thesis has many recommendations that mostly pertain to the limitations of the data that were mentioned previously. The first of which requires some improvements in data collection methods mentioned above, being the accuracy of work zone location, date and time of day when active. In the near future when both of these can be improved using GPS locations at work zone sites the data collected will be more accurate than current data.

The second being streamlining work zone data collection. Currently both INRIX and Crash data sets are accurate and automatically recorded. While for this study work zone data was collected manually. Currently, work is being done to create a single data set in which relevant data would be recorded by each organization working on the project. This will allow for an easier and more accurate analysis than performed in this study. Specifically, this data set would improve the accuracy of work zone dates and distance by noting the starting and ending dates as well as each individual phase length.

The number of work zones analyzed in the study was originally around 200 work zones in Iowa but was reduced to only 32 as most of the work zone pdfs were not available. In order to increase the number of work zones included in the study either more years of data, roughly 5 years, or by reaching out to other state DOT's. Including more states in the data set would allow for a more generalized model for predicting work zone crashes in the United States.

As mentioned above and in the results section, computational power was a limiting factor in the number of methods for the modeling portion of this thesis. As previously mentioned, there is work being done on creating more accurate data for both work zone dates and location. This will reduce the number of data points used in the INRIX data by a huge proportion. This will allow for much easier calculations. Due to this reduction it may be more viable to try other model selection criteria, specifically lasso and elastic net attempted in this study. Another possible solution is to change aggregation of the INRIX data by the day instead of by the hour. If data size is a huge issue again. It may also be a good idea to not include INRIX data if it is not a significant predictor with new better data, as it increases the number of data points in the study by a large amount.

As previously mentioned, it may be a good idea to not include the average speed variable from INRIX data to reduce some computational requirements. After going through the modeling process in this study some variables were identified particularly in the work zone plan data that could be added in the future. The first of which would be the phases of each work zone. At a minimum having a variable for the total number of different phases each work zone has could be a significant factor in predicting crashes, as the number of changing work zone areas could confuse drivers who frequently drive on the road sections.

Another change to the data set would be to separate the curvature variable into two different variables. One of which would be a variable for if the work zone included a crossover to one divided section of the roadway and another to denote if the roadway was curved. The curvature variable could also be improved to denote some specific degree of curvature whether radius or distance or some other method instead of a simple yes or no input. If any other additional variables are proposed to be tested, they should be added if data is available.

Lastly some recommendations will be made in for each of the variables with a large negative impact on the number of work zone crashes. First, the distance of the work zone had a huge impact on the number of vehicle crashes. Either the work zones that span a larger distance should be split into more phases or additional safety measures should be implemented. One suggestion is to reduce the speed limit or to increase the awareness of the work zone. Either by increasing the advanced warning distance, increasing knowledge of detour routes or increasing awareness of the incoming work zone weeks in advance.

Similarly, adding additional safety features to work zones high with AADT, high number of lanes, divided roadways, and other variables that depend solely on the roadways could decrease the number of crashes. Finally, the study found that the worst performing work zones occurred in district 2, 4 and 6. More collaboration between district work zone groups could provide insights to improve the poor performing work zones in each individual district.

The final recommendation applies to the usage of the NB model presented in Table 6. As mentioned before for prediction modeling, all variables are recommended to be between the maximum and minimum values in the study. The accuracy of the equation degrades more rapidly outside of those boundaries. In addition, if any variable is unknown for a work zone using the average value or an estimated value will provide better results than leaving it as a blank field.

REFERENCES

1. Akepati, S. R., & Dissanayake, S. (2011). Characteristics of the Work Zone Crashes. *Transportation and Development Institute Congress 2011*. doi: 10.1061/41167(398)122

2. Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*(4), 2249–2260. doi: 10.1016/j.csda.2007.08.015

3. Capelli, R., Carloni, P., & Parrinello, M. (2019). Exhaustive Search of Ligand Binding Pathways via Volume-Based Metadynamics. *The Journal of Physical Chemistry Letters*, *10*(12), 3495–3499. doi: 10.1021/acs.jpclett.9b01183

4. CDC - Highway Work Zone Safety - NIOSH Workplace Safety and Health Topic. (2019, October 3).

5. Chambless, J., Ghadiali, A. M., Lindly, J. K., & McFadden, J. (2002). Multistate work-zone crash characteristics. *Institute of Transportation Engineers.ITE Journal*, *72*(5), 46.

6. Chen, C., & Xie, Y. (2016). Modeling the effects of AADT on predicting multiple-vehicle crashes at urban and suburban signalized intersections. *Accident Analysis & Prevention*, *91*, 72–83. doi: 10.1016/j.aap.2016.02.016

7. Cho S, Kim H, Oh S, Kim K, Park T: Elastic-net regularization for genomewide association studies of rheumatoid arthritis. BMC proceedings. 2009, 3 (suppl 7): s25-10.1186/1753-6561-3-s7-s25.

8. Daniel, J., Dixon, K., & Jared, D. (2000). Analysis of Fatal Crashes in Georgia Work Zones. *Transportation Research Record: Journal of the Transportation Research Board*, *1715*(1), 18–23. doi: 10.3141/1715-03

9. Garber, N. J., & Zhao, M. (2002). Distribution and Characteristics of Crashes at Different Work Zone Locations in Virginia. *Transportation Research Record: Journal of the Transportation Research Board*, *1794*(1), 19–25. doi: 10.3141/1794-03

10. Harb, R., Radwan, E., Yan, X., Pande, A., & Abdel-Aty, M. (2008). Freeway Work-Zone Crash Analysis and Risk Identification Using Multiple and Conditional Logistic Regression. *Journal of Transportation Engineering*, *134*(5), 203–214. doi: 10.1061/(asce)0733-947x(2008)134:5(203)

11. Inrix. (n.d.). Iowa DOT. Retrieved from https://inrix.com/case-studies/iowa-dot/

12. Khattak, A. J., Khattak, A. J., & Council, F. M. (2002). Effects of work zone presence on injury and non-injury crashes. *Accident Analysis & Prevention*, *34*(1), 19–29. doi: 10.1016/s0001-4575(00)00099-3

13. Kononov, J., Bailey, B., & Allery, B. K. (2008). Relationships between Safety and Both Congestion and Number of Lanes on Urban Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, *2083*(1), 26–39. doi: 10.3141/2083-04

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

15. Mao, L.-Z., Zhu, H.-G., & Duan, L.-R. (2012). The Social Cost of Traffic Congestion and Countermeasures in Beijing. *Sustainable Transportation Systems*. doi: 10.1061/9780784412299.0010

16. Minimal Crash Report. (n.d.). Retrieved February 6, 2020, from http://accidentreports.iowa.gov/index.php?pgname=search_results&lookup=dqsrumjoctc3u2p8mand9sl9i7

17. Mohan, S. B., & Gautam, P. (2000). Cost of Highway Work Zone Injuries. *Construction Congress VI*. doi: 10.1061/40475(278)129

18. Schrank, D., T. Lomax, and B. Eisele. TTI's 2011 Urban Mobility Report, Texas Transportation Institute, Texas A&M University, College Station, TX, September 2011

19. Ullman, G. L., Pratt, M., Fontaine, M. D., Porter, R. J., & Medina, J. (2018). Analysis of Work Zone Crash Characteristics and Countermeasures. doi: 10.17226/25006

20. Zou H, Hastie T: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B. 2005, 67: 301-320. 10.1111/j.1467-9868.2005.00503.x.