2018-06-01

Crash Severity Distributions for Life-Cycle Benefit-Cost Analysis of Safety-Related Improvements on Utah Roadways

Conor Judd Seat
Brigham Young University

Crash Severity Distributions for Life-Cycle Benefit-Cost Analysis

of Safety-Related Improvements on Utah Roadways

Conor Judd Seat

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Mitsuru Saito, Chair
Grant G. Schultz
W. Spencer Guthrie

Department of Civil and Environmental Engineering

Brigham Young University

ABSTRACT

Crash Severity Distributions for Life-Cycle Benefit-Cost Analysis
of Safety-Related Improvements on Utah Roadways

Conor Judd Seat
Department of Civil and Environmental Engineering, BYU
Master of Science

The Utah Department of Transportation developed life-cycle benefit-cost analysis spreadsheets that allow engineers and analysts to evaluate multiple safety countermeasures. The spreadsheets have included the functionality to evaluate a roadway based on the 11 facility types from the Highway Safety Manual (HSM) with the use of crash severity distributions. The HSM suggests that local agencies develop crash severity distributions based on their local crash data. The Department of Civil and Environmental Engineering at Brigham Young University worked with the Statistics Department to develop crash severity distributions for the facility types from the HSM.

The primary objective of this research was to utilize available roadway characteristic and crash data to develop crash severity distributions for the 11 facility types in the HSM. These objectives were accomplished by segmenting the roadway data based on homogeneity and developing statistical models to determine the distributions. Due to insufficient data, the facility types of freeway speed change lanes and freeway ramps were excluded from the scope of this research. In order to accommodate more roadways within the research, the facility type definitions were expanded to include more through lanes.

The statistical models that were developed for this research include multivariate regression, frequentist binomial regression, frequentist multinomial, and Bayesian multinomial regression models. A cross-validation study was conducted to determine the models that best described the data. Bayesian Information Criterion, Deviance Information Criterion, and Root-Mean-Square Error values were compared to conduct the comparison. Based on the cross-validation study, it was determined that the Bayesian multinomial regression model is the most effective model to describe the crash severity distributions for the nine facility types evaluated.

Keywords: crash severity, crash severity distribution, life-cycle benefit-cost analysis, Utah

ACKNOWLEDGEMENTS

This research was made possible with funding from the Utah Department of Transportation. I acknowledge those who had an impact in the research and who have supported me throughout my academic career. First, I thank the members of the Utah Department of Transportation Technical Advisory Committee, especially including Scott Jones. In addition, multiple professors and students have positively contributed to this research and assisted in creating the final product. Their effort has been invaluable. Next, I thank Dr. Saito, Dr. Schultz, and Dr. Guthrie for their feedback, advice, mentorship, and support and for serving on my graduate committee. Each member has taught me technical skills that will aid in my professional career. They have helped me grow and improve in ways that I could not have imagined. In addition, I thank all of the faculty and staff in the Department of Civil and Environmental Engineering who have been part of my academic journey and have taught me principles, concepts, and life skills. Finally, I thank my main support system. First, I am grateful to my parents, who have encouraged me to be the very best person I can be. They have supported me in everything I have done in my life. Next, I thank my siblings, Brooke, Melanie, Abbey, and Eric. They have taught me how to work hard and have fun doing it. Finally, I thank my wife, Marlee, who has aided me in all of my endeavors with a smile.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Roadway safety is one important aspect taken into consideration when roadways are rebuilt, rehabilitated, or maintained. There continues to be a large portion of research in the United States relating to the safety of roadways. One facet of safety-related research is the development of life-cycle benefit-cost analysis, which helps determine which safety countermeasure provides the best benefit for the lowest cost. A previous study funded by the Utah Department of Transportation (UDOT) developed life-cycle benefit-cost analysis spreadsheets (Saito et al. 2016) using the method presented in the Highway Safety Manual (HSM) that is applicable to various highway types included in the manual (AASHTO 2010). The outcome of a life-cycle benefit-cost analysis is significantly affected by crash severity distributions used to predict the number of crashes of each severity type that will be reduced on the roadway after safety-related improvements are implemented.

This research was conducted to develop crash severity distributions using UDOT's crash data for the life-cycle benefit-cost analysis spreadsheets developed in a previous study (Saito et al. 2016). This chapter presents the background information related to this research, explains the purpose and need for this research, and describes the organization of the report.

1.1 Background

Safety has become increasingly important on roadways over the last several decades. UDOT has made roadway safety one of their top priorities, which is expressed in their campaign: "Zero Fatalities: A Goal We Can All Live With™." The goal of zero fatalities is "all about eliminating fatalities on [Utah] roadways" (UDOT 2016). One technique that UDOT uses to reduce fatalities on roadways is continued investment in transportation safety research. Through transportation safety research, safety-related improvements on roadways can be evaluated to understand which improvements will be most effective.

The HSM, originally published in 2010, presents the preferred methods for performing life-cycle benefit-cost analysis of safety-related improvements (AASHTO 2010). UDOT recently adopted the most reliable method for determining the change in crashes known as the Part C Predictive Method (AASHTO 2010). The Part C Predictive Method is an 18-step method for predicting average crash frequencies. The Part C Predictive Method was applied through a series of Excel-based life-cycle benefit-cost spreadsheets developed by the Brigham Young University (BYU) safety research team (Saito et al. 2016). The purpose of the spreadsheets is to provide engineers and analysts with a tool to evaluate multiple countermeasures and their life-cycle benefits so that the engineer or analyst can select improvements for highway segments and intersections that will contribute the most to the prevention of future crashes.

One component of life-cycle benefit-cost analysis is the use of crash severity distributions. Crash severity distributions describe the distribution of crashes by severity type. Crash severity distributions are important to life-cycle benefit-cost analysis because they are used to generate estimates of cost savings by predicting the severity of crashes that will be reduced as a result of implementing a countermeasure. Although there is a single crash severity distribution given in the HSM, it is recommended that separate severity type distributions be developed for all highway types included in the HSM. The HSM recommends that each agency calibrate the predictive models in order to apply the models to their jurisdiction (AASHTO 2010). The spreadsheets recently developed for UDOT include the analysis for 11 facility types, as outlined in the HSM; however, they all use the same default crash severity distribution included in the HSM. UDOT currently has only one crash severity type distribution that has been used for conducting life-cycle benefit-cost analyses for evaluating safety-related improvements. There was a need to develop multiple crash severity distributions by roadway type to more accurately evaluate safety-related improvements.

1.2 Purpose and Need

The purpose of this research was to develop crash severity distributions for the 11 facility types outlined in the HSM. The crash severity distributions were developed using various statistical models. Available crash data together with highway mile point and functional classification data were used for the input file for developing statistical models to generate crash severity type distributions. The scope of this research included a comprehensive literature review, a crash severity distribution survey, the development of several statistical models of crash severity distribution, and conclusions and recommendations. Specifically, the crash severity distributions included the following 11 facility types included in the HSM (AASHTO 2010):

1. Rural two-lane two-way (TLTW) highways
2. Undivided rural multilane highways
3. Divided rural multilane highways

4. Two-lane undivided suburban/urban arterials

5. Three-lane suburban/urban arterials including a two-way left-turn lane (TWLTL)

6. Four-lane undivided suburban/urban arterials

7. Four-lane divided suburban/urban arterials

8. Five-lane suburban/urban arterials including a TWLTL

9. Rural and urban freeway segments

10. Freeway speed change lanes

11. Freeway ramps

The need for this research arose as UDOT does not currently use multiple crash severity distributions for its current life-cycle benefit-cost analysis; a single crash severity distribution is used for the entire UDOT roadway system. The HSM states that the purpose of the calibration procedure "is to adjust the predictive models which were developed with data from one jurisdiction in another jurisdiction" (AASHTO 2010). Calibration will account for differences between jurisdictions in factors such as climate, driver populations, animal populations, crash reporting thresholds, and crash report system procedures.

By including multiple crash severity distributions within the current life-cycle benefit-cost analysis, the analysis will significantly improve. The number of crashes by severity type prevented by countermeasures will be more accurate than using a single set of severity distributions. Safety engineers and analysts will more effectively allocate tax payer money to projects that will reduce vehicle crashes, especially severe vehicle crashes.

As part of previous research efforts by BYU, the Utah Crash Prediction Model (UCPM) and Utah Crash Severity Model (UCSM) were developed (Schultz et al. 2015). These models are only used for roadway segments and cannot be applied to intersections and interchanges at the time of this research. Since this research effort was performed in conjunction with the research effort for these models, intersections and interchanges were not included as part of this research.

1.3 Organization

This thesis consists of six chapters. Chapter 1 presents an overview of the report along with a stated purpose, scope, and need for this research. Chapter 2 contains the literature review, which is a summary of findings related to the research. Chapter 3 outlines the content and results of a survey distributed to state departments of transportation (DOTs). Chapter 4 presents the methodology pertaining to the creation of the crash severity distributions including the data preparation and development of the statistical models. Chapter 5 contains the results from the statistical models for the crash severity distributions. Chapter 6 presents the conclusions and recommendations for future research.

2 LITERATURE REVIEW

A comprehensive literature review has been performed on general aspects of traffic safety and crash severity distributions. This process consisted of gathering information that could contribute to this study. Several topics are addressed in this literature review. First, life-cycle benefit-cost analysis is reviewed along with UDOT's current approach to this analysis. Next, general aspects of traffic safety are addressed, including crash severity, monetary benefit of crashes, the predictive method, and facility types outlined in the HSM. Lastly, a summary of previous research completed by BYU, including the UCPM and the UCSM, is discussed.

2.1 Life-Cycle Benefit-Cost Analysis

The HSM can be considered as the basis for anything related to safety on roadways (AASHTO 2010). Life-cycle benefit-cost analysis of safety-related improvements and the method explained in the HSM can be considered as the preferred method to complete such analysis. This section describes the methods outlined in the HSM and the current method used by UDOT for life-cycle benefit-cost analysis. Finally, the use of crash severity distributions in life-cycle benefit-cost analysis is discussed.

2.1.1 HSM Techniques for Life-Cycle Benefit-Cost Analysis

The safety benefits for a project are determined using the crash information for a site. One of the most important parts of the life-cycle benefit-cost analysis of safety-related improvements is to estimate the change in the number of crashes resulting from a proposed project. The HSM outlines four different methods for estimating the change in expected average crash frequency of a proposed project or project design alternative. The four methods listed in the Part C Predictive Method are presented in order of most to least reliable (AASHTO 2010):

- *Method 1* – Apply the Part C Predictive Method to estimate the expected average crash frequency of both the existing and proposed conditions.
- *Method 2* – Apply the Part C Predictive Method to estimate the expected average crash frequency of the existing condition and apply an appropriate project crash modification factor (CMF) from Part D to estimate the safety performance of the proposed condition.
- *Method 3* – If the Part C Predictive Method is not available, but a safety performance function (SPF) applicable to the existing roadway condition is available, use that SPF to estimate the expected average crash frequency of the existing condition and apply an appropriate project CMF from Part D to estimate the safety performance of the proposed condition. A locally derived project CMF can also be used in Method 3.
- *Method 4* – Use observed crash frequency to estimate the expected average crash frequency of the existing condition and apply an appropriate project CMF from Part D to the estimated expected average crash frequency of the existing condition to obtain the estimated average crash frequency for the proposed condition. This method is applied to facility types not addressed by the Part C Predictive Method.

When a CMF is used in one of the four method outlined above, the standard error of the CMF can be applied to develop a confidence interval around the estimated expected average crash frequency. With this range, analysts can see the type of variation associated with implementing a countermeasure (AASHTO 2010).

2.1.2 Current UDOT Method

The current recommended method for UDOT for life-cycle benefit-cost analysis uses the most reliable method in the HSM. Saito et al. (2016) developed a series of spreadsheets that would implement the Part C Predictive Method for use with life-cycle benefit-cost analysis. The spreadsheets allow the life-cycle benefit-cost estimates to be analyzed for the 11 facility types outlined in the HSM. Although default values are given throughout the spreadsheet, crash costs, expected average crash frequency, and crash severity distribution can be changed in order to fit the specific site that is being analyzed. The use of crash severity distributions in life-cycle benefit-cost analysis is discussed briefly in the next section.

2.1.3 Use of Crash Severity Distributions

One important aspect associated with life-cycle benefit-cost analysis of safety-related countermeasures is determining the total benefits. The main benefit in the case of safety-related improvements is the expected reduction in crashes within the study site. Crash severity distributions are used to predict the severity crash types that will be reduced as a consequence of the safety-related countermeasure. It is important to note that different countermeasures may reduce different crash types.

2.2 Traffic Safety

In the HSM, safety is defined as "the crash frequency or crash severity, or both, and collision type for a specific time period, a given location, and a given set of geometric and operational conditions" (AASHTO 2010). There are two types of safety analyses: subjective and objective. Subjective safety analysis relates to how safe a person feels on a roadway, while objective safety analysis refers to the use of quantitative measures that are independent of the observer. The HSM focuses on objective safety analysis. In the evaluation and estimation methods presented in the HSM, crash frequency is used as a fundamental indicator of safety (AASHTO 2010). This section first defines crash severity. Next, the predictive method from the HSM is presented. Finally, the facility types outlined in the HSM are presented.

2.2.1 Crash Severity

One major component of traffic safety is crash severity, defined as the level of injury or property damage of the crash (AASHTO 2010). Although many injuries may be inflicted during a crash, crash severity is defined as the most severe level of injury that is caused by the crash. In most agencies, crash severity is divided into categories known as the KABCO scale. The five KABCO categories used in the HSM are (AASHTO 2010):

- K: Fatal injury: an injury that results in death;
- A: Incapacitating injury: any injury, other than a fatal injury, that prevents the injured person from walking, driving, or normally continuing the activities the person was capable of performing before the injury occurred;
- B: Non-incapacitating evident injury: any injury, other than a fatal or incapacitating injury, that is evident to observers at the scene of the crash in which the injury occurred;
- C: Possible injury: any injury reported or claimed that is not fatal, incapacitating, or non-incapacitating evident injury and includes claim of injuries not evident;
- O: No injury: Property damage only (PDO).

UDOT uses similar crash severity categories based on the KABCO scale. The numerical values 5 to 1 are used rather than the alphabetical characters KABCO. Severity 1 is PDO, while severity 5 is a fatal injury crash. Traffic safety can be improved in two ways: a decrease in average crash frequency or a decrease in the average crash severity. This section explains the monetary benefits of crash frequency based on FHWA and UDOT standards. Next, factors contributing to crashes are discussed.

2.2.1.1 Monetary Benefit of Crash Frequency

After the change in crash frequency has been estimated for a project, the benefits from reducing the crashes needs to be converted to a monetary value. There are many different opinions as to how much value should be placed on the different severity levels of crashes. The Federal Highway Administration (FHWA) has completed a significant amount of research that establishes a basis for quantifying, in monetary value, the human capital crash costs to society of fatalities and injuries from highway crashes (AASHTO 2010). The FHWA values for each crash severity type are shown in Table 2-1.

Table 2-1: FHWA Benefit Value Per Crash for Each Crash Type (AASHTO 2010)

Severity	Severity Category	Severity No.	Value
PDO	O	1	$ 7,400.00
Possible Injury	C	2	$ 44,900.00
Evident Injury	B	3	$ 79,000.00
Disabling Injury	A	4	$ 216,000.00
Fatal	K	5	$ 4,008,900.00

State and local jurisdictions often have adopted the crash costs by crash severity and collision type. For example, UDOT has their own monetary values that they use in determining the value of each crash severity level (Wall 2016). As would be expected, monetary values generally increase as the severity level increases. However, UDOT equalizes the scale for disabling injury crashes and fatal crashes. This is done in order to lessen the benefit of reducing fatal crashes and increase the benefit of reducing disabling crashes. It is obvious that fatal crashes should most definitely be prevented; however, in many cases, disabling crashes become more expensive in the long run due to medical costs and because the persons involved in these incapacitating injuries are prevented from ever working again though still requiring full-time care in many cases. The UDOT values for each crash severity type are shown in Table 2-2.

Table 2-2: UDOT Benefit Value Per Crash for Each Crash Type (Wall 2016)

Severity	Severity Category	Severity No.	Value (Year 2015 value)
PDO	O	1	$ 3,200.00
Possible Injury	C	2	$ 62,500.00
Evident Injury	B	3	$ 122,400.00
Disabling Injury	A	4	$ 1,961,100.00
Fatal	K	5	$ 1,961,100.00

2.2.1.2 Factors Contributing to a Crash

Although many crashes refer to the “cause” of a crash, it is more accurate to attribute crashes to many contributing causes. Cause may include time of day, driver attentiveness, speed, vehicle condition, road design, and many other factors. Traditionally, these factors can be classified into three different categories: human, vehicle, and roadway or environmental. Human factors refer to characteristics of the drivers including the age, judgement, skill, attention, fatigue, experience, and sobriety. Vehicle factors refer to the design, manufacture, and maintenance of the involved vehicles. Roadway or environmental factors refer to the geometric alignment, traffic control devices, surface friction, weather, and visibility of the surrounding area (AASHTO 2010).

By understanding the various factors that influence the sequence of events, crashes and crash severities can be reduced by implementing specific measures to target specific contributing factors. In 1979, research was completed to describe the distribution of contributing crash factors to vehicle crashes and their relationships to each other (Treat et al. 1979). The findings from this research are shown in Figure 2-1.

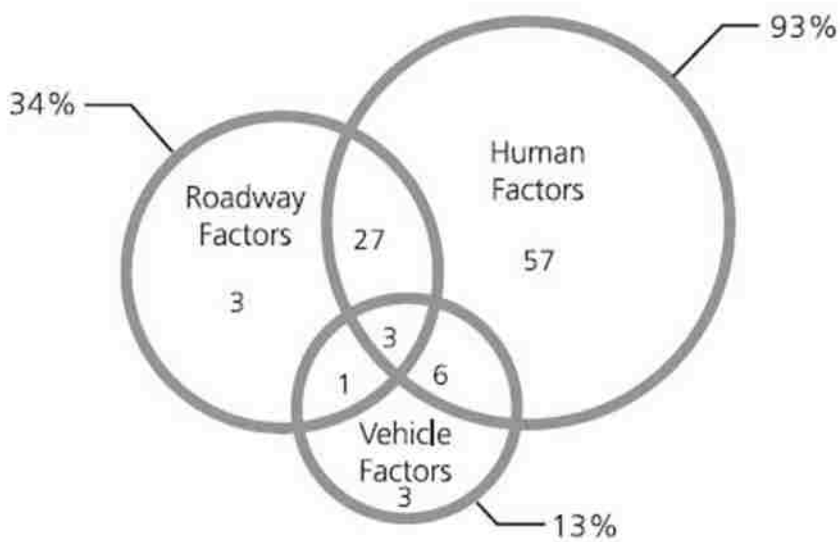


Figure 2-1: Contributing crash factors to vehicle crashes (Treat et al. 1979).

It was observed that 34 percent of crashes were caused, at least in part, by roadway factors, while human factors and roadway factors together caused 27 percent of crashes (Treat et al. 1979). While there are many strategies for reducing crashes and severity, the majority of these strategies are not within the scope of the HSM. As such, the HSM focuses mainly on crashes where it is believed that the roadway or environment is a contributing factor, whether wholly or in part.

2.2.2 Predictive Method

Within the HSM, the predictive method is explained. The predictive method is the methodology in Part C of the HSM that is used to estimate the "expected average crash frequency of a site, facility, or roadway under given geometric design and traffic volumes for a specific period of time" (AASHTO 2010). The predictive method outlined in the HSM uses the Empirical Bayes (EB) method. One clear advantage to using the EB method is, once the model is calibrated for a particular site, the model can be readily applied to the region for which it was calibrated (AASHTO 2010).

There are two basic elements of the predictive method. First, the predictive method estimates the average crash frequency for a specific site type. This is accomplished using a statistical model developed from data for a number of similar sites and adjusted for specific site and local conditions. Second, the expected crashes and observed crashes for the site are combined. A weighting factor is applied to the two estimates to reflect the model's statistical reliability. Currently, the HSM provides a detailed predictive method for three facility types: rural TLTW, rural multilane highways, and urban and suburban arterials (AASHTO 2010).

There are some major advantages to using the predictive method. First, regression-to-the-mean analysis focuses on long-term expected crash frequency rather than short-term observed crash frequency. Another major advantage is that the reliance on availability of limited crash data is reduced by incorporating predictive relationships based on data from similar sites. In addition, the predictive method accounts for the fundamentally nonlinear relationship between crash frequency and traffic volume. Last, the SPFs in the HSM are based on the negative binomial distribution, which are better suited to modeling the high natural variability of crash data than traditional modeling techniques based on the normal distribution (AASHTO 2010).

2.2.3 Facility Types

The HSM outlines 11 facility types for which the predictive method is applicable. The 11 facility types are (AASHTO 2010):

1. Rural TLTW highways
2. Undivided rural multilane highways
3. Divided rural multilane highways
4. Two-lane undivided suburban/urban arterials
5. Three-lane suburban/urban arterials including a TWLTL
6. Four-lane undivided suburban/urban arterials
7. Four-lane divided suburban/urban arterials
8. Five-lane suburban/urban arterials including a TWLTL
9. Rural and urban freeway segments
10. Freeway speed change lanes
11. Freeway ramps

Each facility type has different attributes according to urban code, the number of through lanes, TWLTLs, median type, and functional class.

2.3 Previous BYU Research

Several research efforts have been conducted by BYU with regards to traffic safety and the HSM predictive method. The UCPM and UCSM were developed by Schultz et al. (2015) and are explained in further detail in the following sections.

2.3.1 UCPM

The UCPM was developed to help UDOT identify segments of roadway that have a higher number of crashes than expected (Schultz et al. 2015). This model uses a variety of parameters such as vehicle-miles traveled (VMT), number of lanes, speed limit, and others to create a crash distribution for different roadways. The median of the distribution is used as the expected number of crashes that might occur on a specific segment based on the characteristics of that segment. Using the Bayesian horseshoe selection method, a pre-selection process is performed that takes all possible parameters in the dataset to produce a list of the significant ones that should be used. The parameters can be used to predict a distribution of the number of expected crashes for a given severity group (Schultz et al. 2015).

To start the procedure, a statistical model must be chosen to provide the base dataset in the analysis and identification of the problem segments or "hot spots." Crash data for the years 2008 to 2012 were used in this project's statistical model (Schultz et al. 2015). From this model, the total crash counts for each segment and the count of crashes for each attribute were selected by the Bayesian horseshoe selection method. The UCPM required 100,000 iterations to obtain posterior predictive distributions on the expected number of crashes to occur on each segment. Crash counts were available for all severity levels combined and severity levels K and A (Schultz et al. 2015).

Because this model can be used to determine the number of crashes that are expected to occur on a given roadway segment, it can help determine the number of crashes that will be reduced on each roadway segments when the values of selected variables change. The same can be applied to severe crashes. After the number of crashes reduced is determined, the benefit can be calculated by comparing different possible treatments to improve safety (Schultz et al. 2015).

2.3.2 UCSM

The UCSM is used to determine the probability of a severe crash occurring. Three types of input data are required for this analysis: 1) the probability that a severe crash occurs given that a crash has occurred on a selected segment, 2) the predicted number of severe crashes, and 3) the probability that the respective number of severe crashes occurred. Each segment may then be assigned a ranking based on the difference between the actual and the predicted number of crashes to find the most dangerous road segments (Schultz et al. 2015).

This model can be run with the same dataset as the crash prediction model with one exception. Not only must the UCSM have a count of every crash that happened on that segment in the given time period, but it must also have a count of crashes occurring in the severity group (Schultz et al. 2015).

This model is helpful in determining which roadways are more dangerous according to crash severity. If more severe crashes are occurring than what is predicted, then it is recommended that the road be analyzed further for possible safety-improvement measures (Schultz et al. 2015).

2.4 Chapter Summary

Crash severity distributions are used in life-cycle benefit-cost analysis by determining the total benefits of safety-related improvements. The benefits are typically the expected reduction in crashes as a consequence of the countermeasure. The method outlined in the HSM is the most preferred method of such analysis. To determine the monetary value of a reduced crash, the five crash severity levels have associated monetary values. In Utah, the values for fatal injury crashes and disabling crashes have been equalized because disabling crashes may become more expensive in the long run. The facility types outlined in the HSM are based on varying attributes of urban code, number of through lanes, TWLTLs, median type, and functional class. Previous BYU research has developed the UCPM and the UCSM, which locate hot spots on roadway segments that perform worse than expected in terms of safety.

3 CRASH SEVERITY DISTRIBUTION SURVEY

Due to the lack of literature relating to crash severity distributions, a survey, which was conducted using Qualtrics software (Qualtrics 2017), was distributed to each state DOT in the United States. The purpose of this survey is to determine the uses of crash severity distributions in conducting life-cycle benefit-cost analyses of countermeasures. This chapter presents the content of the survey, the survey distribution, the survey results, and additional information on crash severity distributions obtained through this survey.

3.1 Survey Content

The content of the survey on crash severity distributions focused on the uses, benefits, and derivation of life-cycle benefit-cost analysis and crash severity distributions across the United States. The survey was designed in such a way that different questions could be asked based on the answers the respondent gave to questions earlier in the survey. The maximum number of questions a respondent was required to answer was 13. It was estimated that the survey would take about 5 minutes to complete. The majority of questions were multiple choice questions; however, some questions allowed respondents to input unique text and upload relevant documents. Appendix A includes the full survey and survey flow that was used to collect the data.

3.2 Survey Distribution

Contacts for the DOTs across the United States were obtained from lists provided by the UDOT Safety Division and the Subcommittee of Safety Management found on the American Association of State Highway and Transportation Officials (AASHTO) website (AASHTO 2018). The list provided by UDOT was consulted first, and 34 of the 50 contacts were found from this list. The remaining 16 contacts were obtained from the AASHTO website.

The survey was distributed to all 50 state DOTs in the United States on February 23, 2017, at approximately 12:30 PM MST and was closed at approximately 10:00 AM MDT on March 30, 2017. Two reminder emails were distributed on March 8, 2017, at 11:30 AM MST and March 27, 2017, at 12:30 PM MDT to respondents that had not completed the survey.

3.3 Survey Results

Of the 50 DOTs that were sent the crash severity distribution survey, 27 DOTs responded to the survey. However, three responses were not included in the data analysis because their responses were incomplete, reducing the number to 24 respondents. The states that were represented in the data analysis include, in alphabetical order, Alabama, Delaware, Hawaii, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Massachusetts, Minnesota, Mississippi, Missouri, Nebraska, New Jersey, New Mexico, New York, North Dakota, Oklahoma, Oregon, South Carolina, South Dakota, and Vermont.

One of the purposes of this survey was to understand the uses of life-cycle benefit-cost analysis throughout the United States. According to the results of the survey, 20 of the 24 respondents use life-cycle benefit-cost analysis.

The types of distributions applied to the life-cycle benefit-cost analysis were also surveyed. Results regarding the types of distributions applied to life-cycle benefit-cost analysis are shown in Figure 3-1. Nearly half of all respondents indicated that they use multiple distributions that were derived from their state's crash data. Additionally, 21 percent of respondents indicated that their single crash severity distribution was derived from their state's crash data. Only three of the 24 respondents said that they currently use the crash severity distribution given in the HSM.

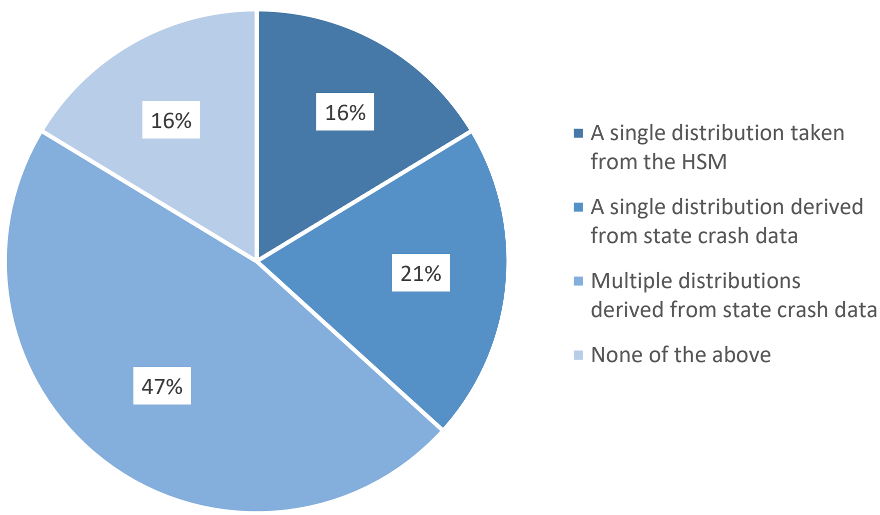


Figure 3-1: Crash severity distributions used from survey respondents.

Another question on the survey explored the DOTs interest in and ability to develop crash severity distributions for the 11 facility types in the HSM. Figure 3-2 shows the summary of the results of the question. Forty-five percent of the respondents indicated that their respective DOTs have considered developing crash severity distributions for the 11 facility types. Twenty-two percent of respondents indicated that their DOTs have not considered deriving crash severity distributions for the 11 facility types outlined in the HSM. If the respondent answered that their

DOT had not considered deriving these crash severity distributions, they were encouraged to specify a reason why they had not considered this. The most common reason for not developing crash severity distributions for the 11 facility types was that they had insufficient data.

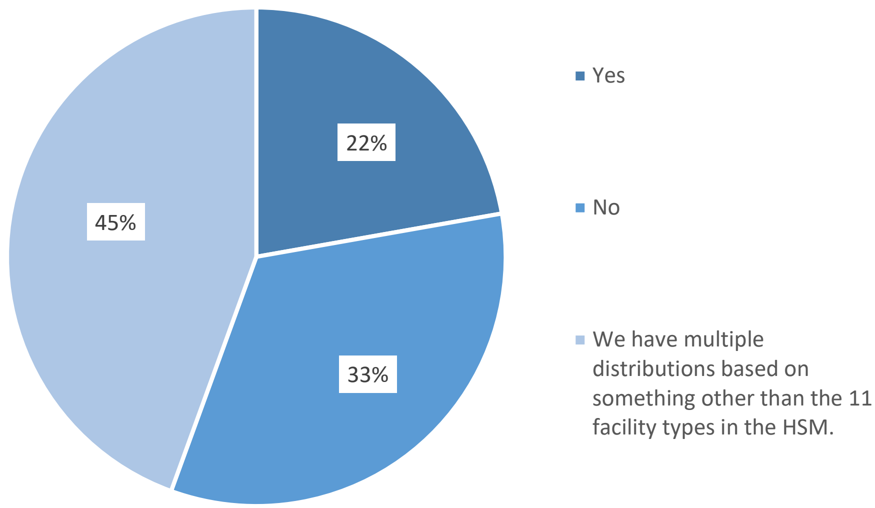


Figure 3-2: Consideration to derive crash severity distribution on 11 facility types.

3.4 Crash Severity Distribution Information

As part of the survey, agencies were given the opportunity to upload literature relating to their states' research on the life-cycle benefit-cost analysis or crash severity distributions. Of the 24 respondents to the survey, four agencies uploaded literature. Even though none of the uploaded literature describe the methodology for creating crash severity distributions, some of the literature show the distributions that the states use in their analyses.

The New York DOT uploaded literature pertaining to their crash severity distributions. A sample of the document is shown in Figure 3-3. The document indicates that the New York DOT defines crash severity on the categories of fatal (K), injury (A, B, and C), and PDO (O) crashes. The segments for the crash severity distributions are classified based on level of access (full, partial, free); urban code (urban, rural); median type (divided, undivided); and number of lanes. Based on these classifications, the New York DOT has 74 different crash severity distributions for roadway segments (NYSDOT 2013).

Classification**			Accident Severity Distribution (percent)			
			Fatal	Injury	Fatal/Injury	PDO*
53	L	FREE ACCESS, RURAL, UNDIVIDED, 4 LANES	0.48	18.59	19.07	80.93
54	A	FREE ACCESS, RURAL, UNDIVIDED, 4 LANES	0.50	19.69	20.19	79.81
55	L	FREE ACCESS, RURAL, UNDIVIDED, ALL LANES	0.48	18.59	19.07	80.93
56	A	FREE ACCESS, RURAL, UNDIVIDED, ALL LANES	0.50	19.69	20.19	79.81
57	L	FREE ACCESS, URBAN, DIVIDED, 2 LANES	0.31	29.55	29.86	70.14
58	A	FREE ACCESS, URBAN, DIVIDED, 2 LANES	0.30	30.35	30.65	69.35
59	L	FREE ACCESS, URBAN, DIVIDED, 4 LANES	0.31	29.55	29.86	70.14
60	A	FREE ACCESS, URBAN, DIVIDED, 4 LANES	0.30	30.35	30.65	69.35
61	L	FREE ACCESS, URBAN, DIVIDED, 6 LANES	0.31	29.55	29.86	70.14
62	A	FREE ACCESS, URBAN, DIVIDED, 6 LANES	0.30	30.35	30.65	69.35
63	L	FREE ACCESS, URBAN, DIVIDED, 7 LANES	0.31	29.55	29.86	70.14
64	A	FREE ACCESS, URBAN, DIVIDED, 7 LANES	0.30	30.35	30.65	69.35
65	L	FREE ACCESS, URBAN, DIVIDED, ALL LANES	0.31	29.55	29.86	70.14
66	A	FREE ACCESS, URBAN, DIVIDED, ALL LANES	0.30	30.35	30.65	69.35
67	L	FREE ACCESS, URBAN, UNDIVIDED, 2 LANES	0.26	24.70	24.96	75.04
68	A	FREE ACCESS, URBAN, UNDIVIDED, 2 LANES	0.25	25.45	25.70	74.29
69	L	FREE ACCESS, URBAN, UNDIVIDED, 3 LANES	0.26	24.70	24.96	75.04
70	A	FREE ACCESS, URBAN, UNDIVIDED, 3 LANES	0.25	25.45	25.70	74.29
71	L	FREE ACCESS, URBAN, UNDIVIDED, 4 LANES	0.26	24.70	24.96	75.04
72	A	FREE ACCESS, URBAN, UNDIVIDED, 4 LANES	0.25	25.45	25.70	74.29
73	L	FREE ACCESS, URBAN, UNDIVIDED, ALL LANES	0.26	24.70	24.96	75.04
74	A	FREE ACCESS, URBAN, UNDIVIDED, ALL LANES	0.25	25.45	25.70	74.29
75	I	3 LEG, RURAL, SIGNAL, ALL LANES	0.52	23.90	24.42	75.58
76	I	3 LEG, RURAL, SIGN, ALL LANES	0.52	23.90	24.42	75.59
77	I	3 LEG, RURAL, NONE, ALL LANES	0.52	23.90	24.42	75.58
78	I	3 LEG, URBAN, SIGNAL, 1-4 LANES	0.23	30.36	30.59	69.42

Figure 3-3: Crash severity distributions for New York segments (NYSDOT 2013).

The Vermont DOT also uploaded literature pertaining to their crash severity distributions. Like many other state agencies, the Vermont crash severity distributions use the KABCO scale. The crash severity distributions are classified based on urban code and functional classification. The Vermont DOT has developed 13 crash severity distributions based on these classifications as shown in Figure 3-4. In addition to crash severity distributions, the literature also includes the prediction model for two-lane rural highways, which is shown in Equation 3-1 (VTrans 2005). This equation predicts the number of crashes per mile per year on two-lane rural highways based on the average daily traffic (ADT), lane width, average paved shoulder width, average unpaved shoulder width, and a roadside rating. By comparing the predicted number of crashes to the actual number of crashes, the Vermont DOT can evaluate the safety of roadway segments.

Not a Junction	Fatal	Injury A	Injury B	Injury C	PDO
RURAL					
01 Interstate, Rural	2.1%	8.0%	22.5%	6.7%	60.7%
02 Principal Arterial	4.8%	7.4%	23.2%	6.7%	57.9%
06 Minor Arterial	3.0%	9.7%	21.0%	8.9%	57.5%
07 Major Collector	2.5%	9.0%	24.5%	6.8%	57.2%
09 Local, Rural	1.9%	7.4%	23.3%	8.8%	58.6%
URBAN					
11 Interstate, Urban	1.0%	6.2%	22.7%	5.2%	64.9%
12 Freeway/Expressway	0.0%	12.5%	37.5%	6.3%	43.8%
14 Principal Arterial	0.4%	3.6%	17.4%	10.4%	68.2%
16 Minor Arterial	0.8%	5.5%	19.1%	12.7%	61.9%
17 Urban Collector	0.0%	4.8%	17.0%	7.8%	70.4%
19 Local, Urban	0.0%	5.3%	15.2%	6.6%	73.0%
GENERAL					
Urban	0.4%	4.6%	17.7%	10.1%	67.2%
Rural	2.6%	8.3%	23.1%	8.0%	57.9%
Urban + Rural	1.7%	7.0%	20.8%	8.1%	62.3%

Figure 3-4: Crash severity distributions for Vermont segments (VTrans 2005).

$$N = (0.0015)(ADT)^{0.9711}(0.8897)^{W}(0.9403)^{PA}(0.9602)^{UP}(1.2)^{H} \tag{3-1}$$

Where,
N = Number of crashes per mile per year
ADT = Average daily traffic
W = Lane width
PA = Average paved shoulder width (feet)
UP = Average unpaved shoulder width (feet)
H = Road side rating (values range from 1 to 7)

3.5 Chapter Summary

This chapter presents the content and results from a survey distributed to every state DOT in the United States regarding the use of life-cycle benefit-cost analysis and crash severity distributions. The survey was expected to take about 5 minutes to complete. The results indicated that 83 percent of the respondents used life-cycle benefit-cost analysis. Nearly half of the respondents that used life-cycle benefit-cost had multiple crash severity distributions based on their respective state's crash data. Forty-five percent of respondents indicated that their respective DOTs have considered developing crash severity distributions for the 11 facility types in the HSM. During the survey, respondents had the option of uploading files relating to their crash severity distributions. Respondents from the New York and Vermont DOTs uploaded the distributions they use in their analysis.

4 METHODOLOGY

This chapter presents the methodology used to meet the objectives of this study. First, the facility types, as outlined in the HSM, are defined. Second, the datasets used for the data preparation are outlined. Next, the segmentation program created for previous BYU research is reviewed, including modifications made to the program to fit the needs of this research. Next, the output of the segmentation program is presented. The straight proportion methodology for calculating crash severity distributions is then discussed briefly. Finally, the statistical models created for developing crash severity distributions are described.

4.1 Facility Type Definition

The HSM describes 11 facility types and the roadway characteristics associated with each type (AASHTO 2010):

1. Rural TLTW highways
2. Undivided rural multilane highways
3. Divided rural multilane highways
4. Two-lane undivided suburban/urban arterials
5. Three-lane suburban/urban arterials including a TWLTL
6. Four-lane undivided suburban/urban arterials

7. Four-lane divided suburban/urban arterials

8. Five-lane suburban/urban arterials including a TWLTL

9. Rural and urban freeway segments

10. Freeway speed change lanes

11. Freeway ramps

Each facility type has different attributes according to urban code, the number of through lanes, TWLTLs, median type, and functional class. The attributes for the first nine facility types, as described in the HSM, are shown in Table 4-1. Based on the quality and type of data available, it was determined to exclude the facility types for freeway change lanes (facility type 10) and freeway ramps (facility type 11) from further analysis.

Table 4-1: Facility Type Attributes (AASHTO 2010)

Facility Type Code	Urban Code	Through Lanes	TWLTL	Median	Functional Class
1	Rural	2	0	Undivided	-
2	Rural	4	0	Undivided	-
3	Rural	4	0	Divided	-
4	Urban	2	0	Undivided	Other Principal Arterial/ Major Arterial
5	Urban	2	1	Undivided	Other Principal Arterial/ Major Arterial
6	Urban	4	0	Undivided	Other Principal Arterial/ Major Arterial
7	Urban	4	0	Divided	Other Principal Arterial/ Major Arterial
8	Urban	4	1	Undivided	Other Principal Arterial/ Major Arterial
9	Either	4, 6, 8, or 10	0	Either	Interstate/ Other Freeway or Expressway

4.2 Datasets

Several different datasets were used in this research that have been received through UDOT's Open Data Portal (UDOT 2017) and other UDOT contacts. The datasets used include Historic Annual Average Daily Traffic (AADT), 2014 Medians, 2014 Lanes, Functional Class, and Urban Code. Crash data, crash location, crash rollup, and crash vehicle data, spanning from 2010-2014, were provided by the UDOT Traffic and Safety Division for the project. Route and mile point data were essential for this study and are included in each dataset. This section expounds on the uniform characteristics in each dataset, critical data columns for datasets retrieved from UDOT's Open Data Portal, and critical data columns for each crash dataset.

4.2.1 Data Uniformity

Each dataset downloaded from the UDOT Open Data Portal has separate attributes corresponding with that dataset; however, uniform data fields exist that allow the datasets to be related linearly or spatially. Four roadway identification fields were used to relate the datasets for analysis. These fields include "ROUTE_ID," "DIRECTION," "BEG_MILEPOINT," and "END_MILEPOINT" for every dataset.

The "ROUTE_ID" field corresponds to the federal and state highway numbering system. The direction of traffic flow is described by the "DIRECTION" field. "BEG_MILEPOINT" and "END_MILEPOINT" identify the beginning and ending mile point, respectively, on the route that the roadway segment characteristics exist.

4.2.2 Critical Data Columns for UDOT Open Data Portal Datasets

Each roadway characteristic dataset has individual attributes that correspond with each dataset. According to the UDOT Data Portal, the AADT dataset has data dating from the most recent year back to 1981 on some segments. Additionally, the traffic counter station number and single and combination truck percentages are included in this dataset. The critical data columns for the AADT dataset include route number, beginning mile point, ending mile point, seven years of AADT data, single-unit truck percentage, and combination-unit truck percentage, as shown in Table 4-2. Similar tables for all of the UDOT Data Portal used in this project are summarized in Appendix B.

Table 4-2: Critical Data Columns for the AADT Dataset

Heading	Description
ROUTE	Route ID: numeric route number of a given road segment
BEGMP	Beginning Mile Point: beginning mile point of the road segment
ENDMP	End Mile Point: ending mile point of the road segment
AADT[YEAR]	AADT[YEAR]: historical dataset of AADT data from each year; at least 7 years of data are needed (i.e., AADT2012 through AADT2018)
SUTRK2015	Single-Unit Truck Percentage: single-unit truck percentage of the road segment
CUTRK2015	Combination-Unit Truck Percentage: combination-unit truck percentage of the road segment

4.2.3 Critical Data Columns for Crash Datasets

The datasets obtained from the UDOT Traffic and Safety Division include crash data, crash location, crash rollup, and crash vehicle data. Each dataset includes a column called CRASH_ID and CRASH_DATETIME, which correspond to each crash that occurred. This labeling system is consistent throughout each crash data file. This allows the information about a specific crash to be found quickly in each dataset.

Aside from a uniform crash ID column, each crash dataset contains different information about the crash. The crash data dataset has information regarding the crash severity, weather conditions, pavement conditions, the type of collision, and other roadway conditions. The crash rollup dataset includes information regarding the number of injuries, whether pedestrians or bicyclists were involved, and related circumstances for the crash that occurred. The crash vehicle dataset has information on the posted speed limit, estimated speeds at time of crash, the number of occupants in each vehicle, and the vehicle make and model. The crash location dataset describes the location of the crash in terms of route number and mile point. Tables depicting the critical data columns for each crash dataset collected for this project are provided in Appendix B.

4.3 Data Preparation

Microsoft Excel was used to prepare the data for more detailed analysis and to create homogeneous segments. The Roadway and Crash Data Preparation Workbook was originally created by Schultz et al. (2016) as a means to segment roadways based on homogeneous characteristics or a specified length from multiple datasets. Modifications were made to this Segmentation Workbook to add more datasets and change the programming code to segment the datasets in a different way than that utilized in the original Workbook. This section briefly addresses the original Excel workbook that was created and the modifications made to the original Workbook for this project.

4.3.1 Original Workbook

The original Roadway and Crash Data Preparation Workbook was created in 2015 and is comprised of two parts (Schultz et al. 2016). The two parts are roadway segmentation and crash data combination. The roadway segmentation part uses five datasets to create roadway segments. The five datasets included in this Workbook are AADT, functional class, speed limit or sign faces, lanes, and urban code. Once all of the roadway datasets have been imported, the user can choose whether to segment the data based on homogeneity or length. The final product of the roadway segmentation process is an Excel spreadsheet with the segmented data.

Combining crash data uses four crash datasets: crash location, crash data, crash rollup, and crash vehicle. Once these datasets are imported into the Workbook, the “Combine Crash

Data” button appears which creates two spreadsheets when executed. One spreadsheet contains all of the crash data and the other contains vehicle data related to the crash.

This Workbook was coded using Visual Basic Application (VBA) software that allows the user to input data and create new spreadsheets by executing commands. Figure 4-1 shows the interface of the Workbook. When the import button corresponding to a particular dataset is executed (i.e., Historic AADT), it allows the user to select a data input file. Once the user selects the input file, the VBA macros copy the data using the critical data columns, such as beginning and ending mile point, route, and data specific to that dataset (e.g., AADT for every year), into the Workbook on a new worksheet. Once the dataset is imported, the “Status” bar next to the import button turns green, signifying the dataset has been properly imported.

When all the datasets are imported into the Workbook, a new button appears on the interface that allows the user to choose whether the data will be segmented by change in the data or by a specified maximum length that the user chooses. Figure 4-2 shows the new button. When the “Combine Roadway Data” button is executed by the user, the VBA code ensures that each dataset has been imported. Next, the code cycles through each dataset and deletes routes that are not present in all five datasets and verifies that each dataset has the same ending mile point for each route. The dataset mile point columns are found in each imported data sheet, and the lowest mile point is the beginning mile point for the segmented data. Again, the VBA code cycles through the imported data sheets, and, every time a change in a dataset is found, a new segment begins. Once the data are segmented, headers are added to the spreadsheet, and the user selects a folder location to save the segmented data.

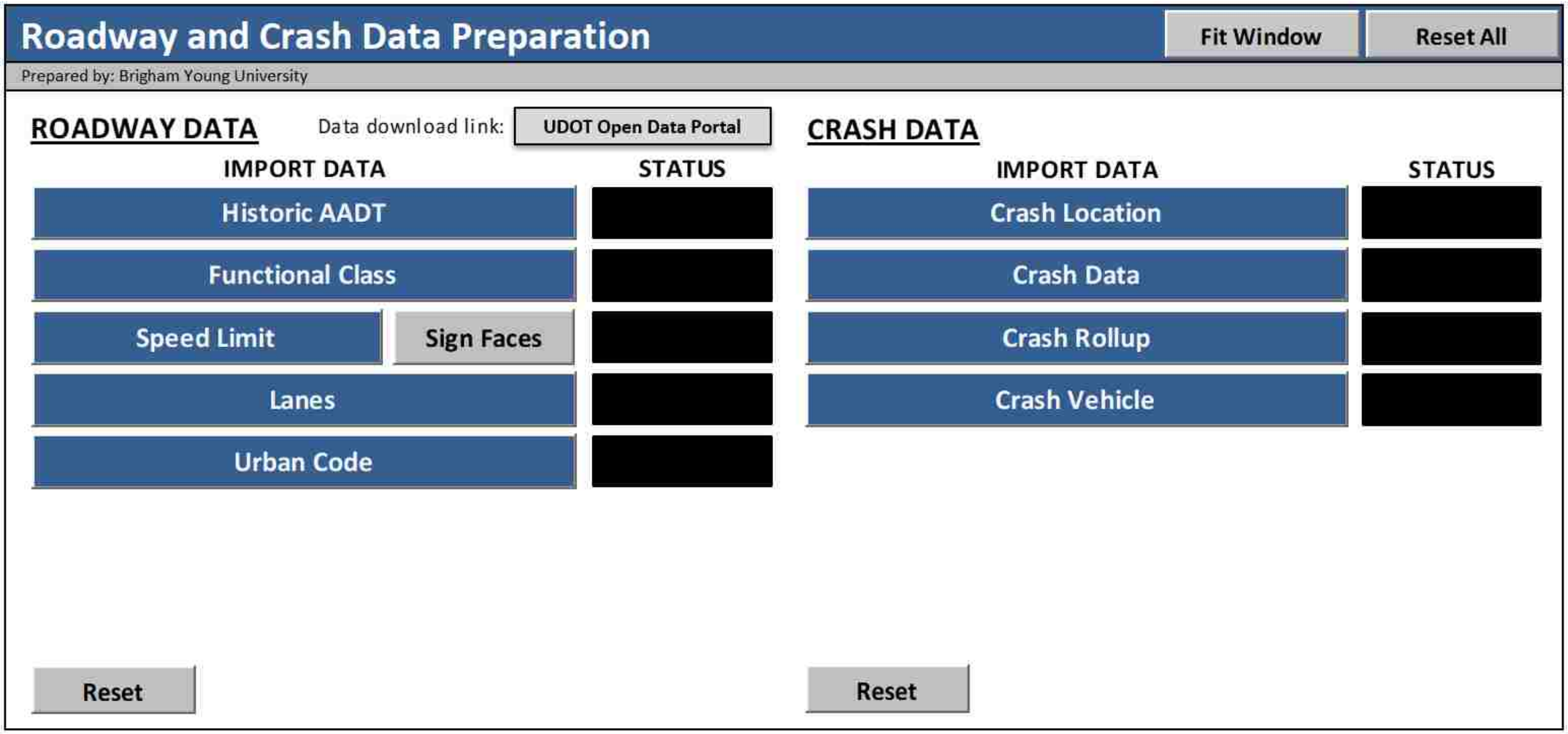


Figure 4-1: Original Roadway and Crash Data Preparation Workbook (Schultz et al. 2016).

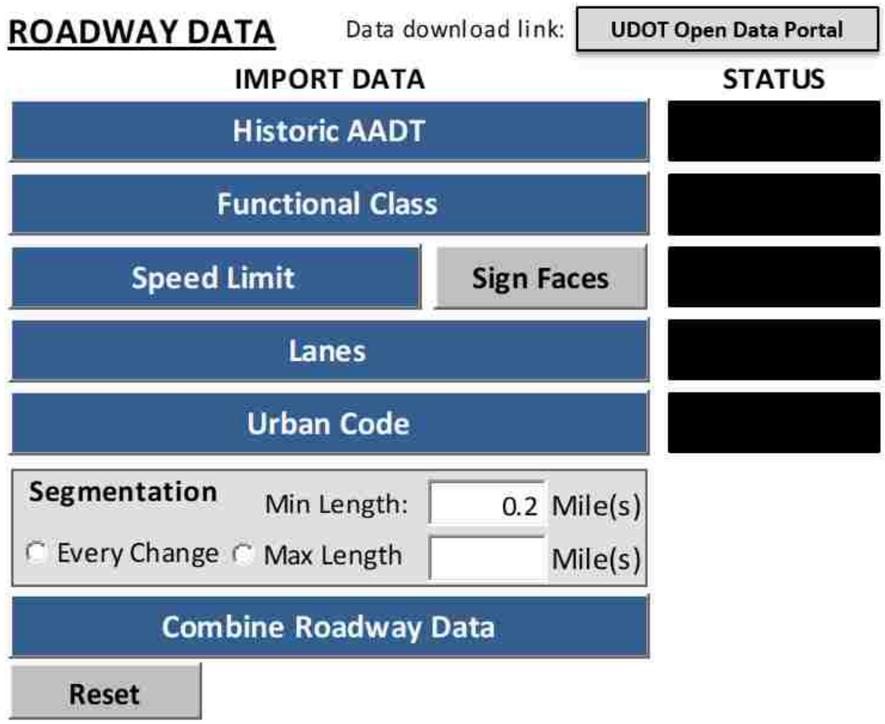


Figure 4-2: Segementation options and combine segmentation button (Schultz et al. 2016).

4.3.2 Modifications to Roadway and Crash Data Preparation Workbook

As a result of the purpose and scope of this project, several changes were made to the Roadway Data Preparation Workbook. These modifications made it possible to add more datasets and combine the roadway data in a different manner than that used in the original workbook. The revised user interface is pictured in Figure 4-3. The portion of the interface associated with the Roadway Data Preparation Workbook is found on the left side of Figure 4-3. Median and crash location were added to the roadway data section of the Workbook. Speed limit was omitted from the roadway data section. In addition, changes were made throughout the VBA code for the lane data, and new codes were added throughout the Workbook to adjust for the specific needs of this research. No changes were made to the Crash Data portion of this Workbook. This section summarizes the modifications made to the roadway data portion of the

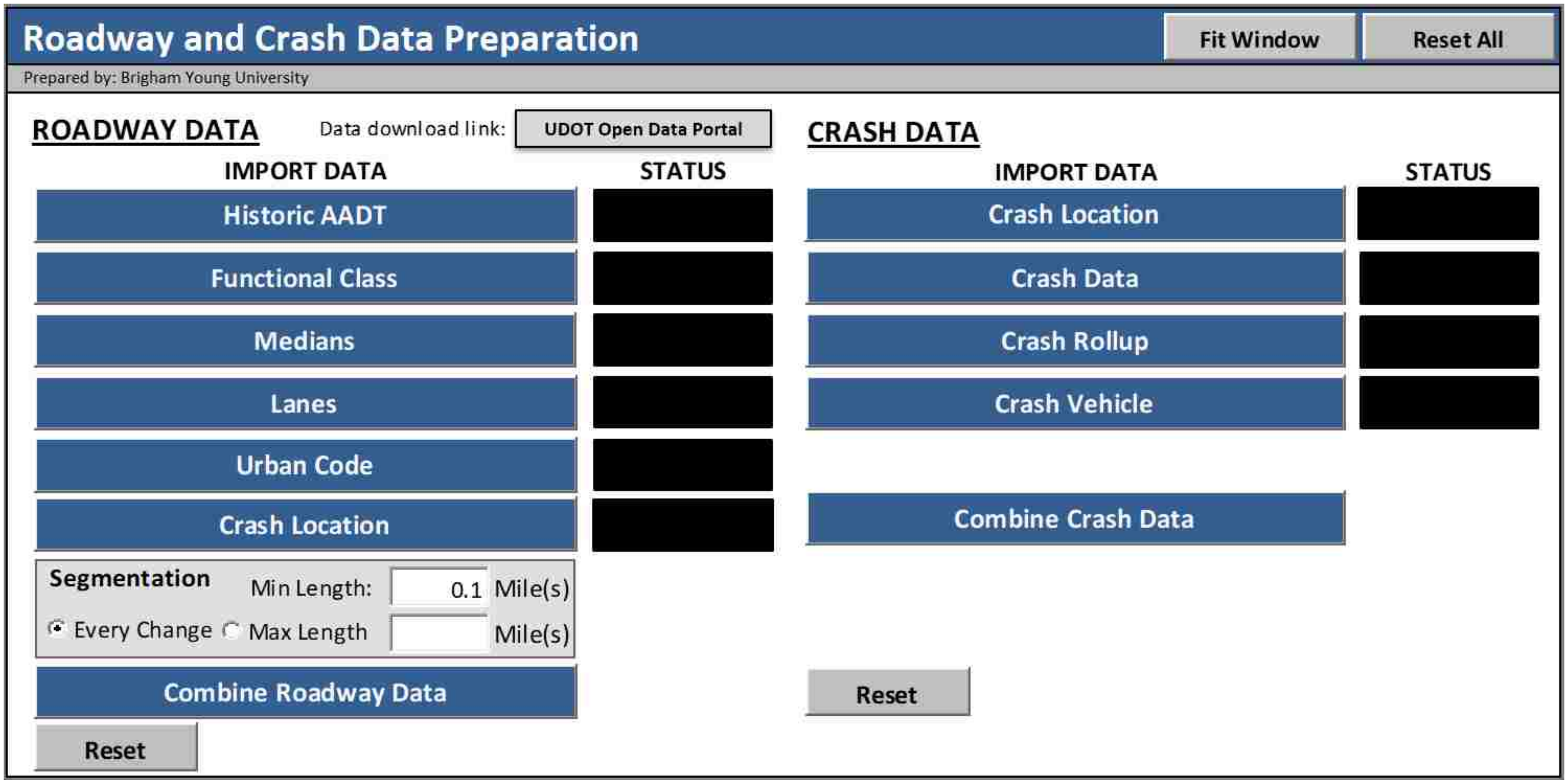


Figure 4-3: Modified Roadway and Crash Data Preperation Workbook (Schultz et al. 2016).

Workbook, including changes made to the lane data, the addition of the median data, crash data additions, and facility type modifications added for this research.

4.3.2.1 Lane Data Modifications

Originally, through-lane data were the only lane type included in the segmentation process. For this study, however, TWLTL data needed to be included. The process of adding the TWLTL lane data into the segmentation process was done by adding TWLTLs to the critical data columns. In addition, the code was altered so that the roadways were segmented according to the addition of the TWLTL data.

4.3.2.2 Median Data Additions

Similar to the lane data, the roadway data were not segmented based on the median type. In the original Workbook, the data were segmented based on speed limit. For this research, speed limit was not necessary, so the median type data replaced the speed limit data.

The original median data are comprised of 10 different median types. These median types include depressed, no median, other divided, painted, railroad, raised island, raised median rapid transit, separate grades, and undivided. Working with so many different types of medians proved to be difficult since the median type would change frequently along the majority of roadway corridors. A proposal was presented to the Technical Advisory Committee (TAC) to consolidate the medians into divided and undivided categories for this project. Upon the approval of the TAC, the consolidated median included the no median, painted median, and undivided median in the undivided category and the remaining median types in the divided category.

4.3.2.3 Crash Data Additions

Another modification that was made to the Segmentation Workbook was the inclusion of the crash data. This dataset was used to calculate the number of crashes that occurred on each roadway segment. Five additional columns were added to the final spreadsheet to include the number of crashes per five-year period analyzed for each severity level.

4.3.2.4 Facility Type Modifications

Once the roadway data were segmented, a facility type was assigned to each roadway segment. If a segment did not strictly fit into any of the nine facility types, an "ERROR" string was entered into the cell for that respective segment. Upon inspection of these results, it was found that 2,443 segments of the 5,732 total segments (42.6 percent) did not meet the criteria for any of the facility types. With approval of the TAC, expanded definitions of the HSM facility types were applied. The expanded facility type attributes are shown in Table 4-3, which can be compared with the HSM definitions of facility type attributes in Table 4-1. With the expanded definitions of facility types, only 12.7 percent of the total segments did not meet the criteria of any of the facility types, compared to the 42.6 percent previously.

Next, it was observed that adjacent segments on the same route had the same facility type. The only varying attribute for the segment was AADT. Segments meeting these criteria were condensed for two reasons: 1) to eliminate short segments that may skew crash distributions and 2) to reduce uncertainty in lane data. When adjacent segments on the same route were condensed, the number of total segments decreased from 5,732 to 1,947 segments. To account for the combining of adjacent segments, a weighted average for AADT based on segment length was included.

Table 4-3: Expanded Facility Type Attributes

Facility Type Code	Urban Code	Through Lanes	TWLTL	Median	Functional Class
1	Rural	2 or 3	0	Undivided	-
2	Rural	4 or more	0	Undivided	-
3	Rural	4 or more	0	Divided	-
4	Urban	2 or 3	0	Undivided	Other Principal Arterial/ Major Arterial
5	Urban	2 or 3	1	Undivided	Other Principal Arterial/ Major Arterial
6	Urban	4 or more	0	Undivided	Other Principal Arterial/ Major Arterial
7	Urban	4 or more	0	Divided	Other Principal Arterial/ Major Arterial
8	Urban	4 or more	1	Undivided	Other Principal Arterial/ Major Arterial
9	Either	4-10	0	Either	Interstate/ Other Freeway or Expressway

4.4 Output

The output for both the original and the modified Roadway Segmentation Workbook is a single Excel spreadsheet that has several different data columns compiled from all the input datasets. Data included in the original output are the beginning and ending mile points of the segment, route, UDOT Region, seven years of AADT data, functional class, urban code, number of through lanes, speed limit, single truck percentages, and combination truck percentages. The modified output includes all of the data included in the original output as well as TWLTL lanes, facility type code, facility type description, vehicle miles traveled, and crash counts for each severity level. Table 4-4 shows all of the column headers that are in the amended output and an example value for each header. The modified segmentation has 1,947 segments, while the original had 6,091 segments.

Table 4-4: Sample Modified Workbook Output

Column Header	Example
Label	0006P
Beg_Milepoint	0
End_Milepoint	46.017
Length	46.017
Route	0006
Route_ID	0006
Direction	P
FC_Code	3
FC_Type	Other Prinicpal Arterial
County	MILLARD
Region	4
UC_Code	99999
UC_Type	Rurall
Median	UNDIVIDED
Through_Lane	2
TWLTL_Lane	0
FT_Code	1
FT_Type	Rural TL TW Highway
AADT_2014	350
AADT_2013	330
AADT_2012	325
AADT_2011	330
AADT_2010	340
AADT_2009	355
AADT_2008	345
Single_Per	0.24961
Combo_Perc	0.232449
VMT_2014	16105.95
Crash: Severity 1	16
Crash: Severity 2	2
Crash: Severity 3	6
Crash: Severity 4	2
Crash: Severity 5	2

4.5 Straight Proportion Method

Once the roadways were segmented based on facility type, a straight proportion was taken for each facility type to create crash severity distributions. The straight proportion method was employed to understand the crash severity distributions for past crash data. The proportions were calculated by dividing the total number of crashes of each severity of a facility type by the total number of crashes on the respective facility type.

4.6 Statistical Model Development

Several statistical models were developed to predict the crash severity distribution for a roadway facility type. The four models that were developed were the multivariate regression, frequentist binomial regression, frequentist multinomial regression, and Bayesian multinomial regression models. These models were chosen because the results are expressed in proportions, which is required for the crash severity distributions since there are more than two crash severities. These four statistical models were developed after exploring the data and examining model diagnostics after fitting each model. This section describes the general development of each model. More information on the detailed development of the statistical models is given by Clegg (2018).

4.6.1 Statistical Foundation

To predict the distribution of crash severities for a facility type, a vector of the probabilities of crash severity based on facility type is required. For each segment, the vector of crashes is as shown in Equation 4-1.

$$\boldsymbol{y}_j = [y_{1j}, y_{2j}, y_{3j}, y_{4j}, y_{5j}] \tag{4-1}$$

Where, $\boldsymbol{y_j}$ = Vector of total number of crashes of all five severities on segment j

y_{ij} = Number of crashes of severity i on segment j

i = Severity of crash (i.e., 1, ..., 5)

j = Roadway segment

Because a crash severity distribution is a vector, the data were assumed to be distributed according to a multinomial distribution, as illustrated in Equation 4-2. This distribution is typically used to describe situations with a discrete number of possible outcomes.

$$\boldsymbol{y}_j \sim Multinomial(n_j, \boldsymbol{\pi_j}) \tag{4-2}$$

Where, $\boldsymbol{y_j}$ = Vector of total number of crashes of all five severities on segment j

n_j = Total number of crashes on segment j

$\boldsymbol{\pi_j}$ = Vector of probabilities for a crash on segment j

i = Severity of crash (i.e. 1, ..., 5)

j = Roadway segment

The vector $\boldsymbol{\pi_j}$ is comprised of the probabilities of crash severity i on segment j, as shown in Equation 4-3.

$$\boldsymbol{\pi_j} = [\pi_{1j}, \pi_{2j}, \pi_{3j}, \pi_{4j}, \pi_{5j}] \tag{4-3}$$

Where, $\boldsymbol{\pi_j}$ = Vector of probabilities for a crash on segment j

π_{ij} = Probability of crash severity i on segment j, $\sum_{i=1}^{5} \pi_{ij} = 1$ because the sum of the probabilities must equal 1.

i = Severity of crash (i.e. 1, ..., 5)

j = Roadway segment

This section describes the general development of the four statistical models that were created as part of this research. The four models include multivariate regression, frequentist binomial regression, frequentist multinomial regression, and Bayesian multinomial regression.

4.6.1.1 Multivariate Regression Model

For the multivariate regression analysis, two initial assumptions are made. The first assumption is that the proportion of each type of crash is distributed normally according to a multivariate normal distribution, as illustrated in Equation 4-4.

$$p_{ij} = \frac{y_{ij}}{n_j} \tag{4-4}$$

Where, p_{ij} = Proportion of crashes of severity *i* on segment *j*

y_{ij} = Number of crashes of severity *i* on segment *j*

n_j = Total number of crashes on segment *j*

i = Severity of crash (i.e., 1, ..., 5)

j = Roadway segment

With regards to these proportions, it was also assumed that these proportions follow a multivariate normal distribution.

In order to perform multivariate regression on the statistic p_{ij}, where p_{ij} is between 0 and 1, the data are required to be transformed so that the data span all the real numbers using function *f(x)*. For a given set of data, linear regression is used to find the line that best describes the major trends in the data. Regression is illustrative of the relations between a set of covariates, *X*, and their response, *Y*. It is commonly associated with a best-fit line. Possible transformations that are commonly used for this type of analysis are the logit, probit, and arcsine transformation, shown in Equations 4-5, 4-6, and 4-7, respectively. These three transformations were used to manipulate the data so that a linear regression model could fit the data better for this project.

$$f(x) = \log(\frac{x}{x-1}) \tag{4-5}$$

$$f(x) = \int_{-\infty}^{x} \frac{1}{2\pi} e^{-\frac{x^2}{2}} dx \tag{4-6}$$

$$f(x) = \sin^{-1}(\sqrt{x}) \tag{4-7}$$

Where, x = Proportion of crashes of severity *i* on segment *j*

It is important to note that multivariate regression does not account for the variability in total segment crashes n_j. For crash severity distributions, it is expected that the distributions will sum to 1. In multivariate regression, the predicted probabilities for the crash severities will not necessarily always sum to 1. Multivariate regression also allows for nonlinear elements in the matrix of covariates. Specifically, the analysis applied varying numbers of natural splines to certain variables.

4.6.1.2 Frequentist Binomial Regression Model

In the multivariate regression analysis, it was assumed that the proportion of crashes of a certain severity on a segment are representative of the actual probability of a crash occurring. With multivariate regression, the model is predicting the proportion of crashes of each severity, not directly estimating the probability of crashes of each severity occurring.

Similar to transformation functions used in the multivariate regression model, frequentist binomial regression models use link functions to estimate the parameters of a distribution. Mathematically, the frequentist binomial regression model can be written as illustrated in Equation 4-8.

$$y_{ij} \sim Binomial(n_j, \pi_{ij}) \tag{4-8}$$

$$f(\pi_{ij}) = \boldsymbol{X_j \beta_i} = \beta_o + X_{ij}\beta_i + \cdots + X_{kj}\beta_{ki} \tag{4-9}$$

Where, y_{ij} = Number of crashes of severity i on segment j

n_j = Total number of crashes on segment j

π_{ij} = Probability of a crash of severity i on segment j

$\boldsymbol{X_j}$ = Vector regression covariates for segment j

$\boldsymbol{\beta_i}$ = Vector of regression coefficients

i = Severity of crash

j = Roadway segment

k = Number of covariates selected

The function f in Equation 4-9 refers to one of the three link functions considered for this analysis. While many link functions exist, the logit, probit, and complementary log-log link functions are the most commonly used for this type of analysis. The logit and probit functions are identical to Equations 4-5 and 4-6. The complementary log-log function is defined in Equation 4-10.

$$f(x) = \log(-\log(1-x)) \tag{4-10}$$

Similar to the multivariate regression, the probabilities for each crash of severity i are normalized so that the probabilities sum to 1. Also, nonlinear elements are allowed in the matrix of the covariates. One limitation of this model is that it does not account for any dependence between the probabilities π_{1j}, π_{2j}, π_{3j}, π_{4j}, and π_{5j}.

4.6.1.3 Frequentist Multinomial Model

Due to the dependence between the probabilities for each crash severity, a frequentist multinomial model was considered for this analysis. The frequentist multinomial model was defined previously in Equation 4-2.

Once again, link functions are used to link the probabilities of each crash severity to the real number line, ℝ. Similar to the binomial regression model, it is not assumed that the probabilities follow a normal distribution. For the frequentist multinomial model, the only link function that was analyzed was the logit function in Equation 4-5.

Multinomial regression performs regression on the odds of one class as compared to a reference class. For the purposes of this analysis, crash severity 1 was set as the reference class. The model of the log-odds of a class of interest compared to the reference class of crash severity 1 can be written as shown in Equations 4-11, 4-12, and 4-13.

$$\pi_{\{i\}j} = \frac{e^{n_{ij}}}{1+\sum_{i=2}^{5} e^{n_{ij}}} \tag{4-11}$$

$$\log(\frac{\pi_{\{i\}j}}{\pi_{\{1\}j}}) = n_{ij} = X_j\beta_i \tag{4-12}$$

$$\pi_{\{i\}j} = \frac{1}{1+\sum_{i=2}^{5} e^{n_{ij}}} \tag{4-13}$$

Where, $\boldsymbol{\pi_j}$ = Vector of probabilities of a crash severity i on segment j

i = Severity of crash

j = Roadway segment

n_j = Total number of crashes on segment j

Once again, many different versions of this model were considered in this analysis, including natural spline functions to account for nonlinearity in the numeric variables.

4.6.1.4 Bayesian Multinomial Regression Model

Finally, a Bayesian multinomial regression model was fit so that a predictive probability distribution on each probability within the crash severity distribution could be used. In order to obtain the probability distribution, the logit function, shown previously in Equation 4-5, was used to link the elements of π_j to the real number line. The Bayesian multinomial regression model that employs the logit link function is written as shown in Equations 4-14, 4-15, and 4-16.

$$y_i = Multinomial\ (n_j, \boldsymbol{\pi_j}), \text{such that for } i \neq 1 \tag{4-14}$$

$$\log\left(\frac{\pi_{\{i\}j}}{\pi_{\{i\}j}}\right) = \eta_{ij} = \boldsymbol{X_i \beta_i} = \beta_{i0} + X_{1j}\beta_{i1} + \cdots + X_{kj}\beta_{ik} \tag{4-15}$$

$$\log\left(\frac{\pi_{\{i\}j}}{\pi_{\{1\}j}}\right) = \eta_{1j} = 0 \tag{4-16}$$

Where, $\boldsymbol{\beta_1} = \boldsymbol{0}$

$\boldsymbol{\pi_{ij}}$ = Vector of probabilities of a crash severity *i* on segment *j*

n_j = Total number of crashes on segment *j*

η_{ij} = Vector of regression covariates for segment *j* and regression coefficients

$\boldsymbol{X_j}$ = Vector regression covariates for segment *j*

$\boldsymbol{\beta_i}$ = Vector of regression coefficients

i = Severity of crash

j = Roadway segment

k = Number of regression covariates included in the model

This Bayesian multinomial regression model can be exploited further. It is expected that for certain variables, the effect will vary depending on crash severity. For example, an increase in the AADT of a road will affect the probability of severity 1 crashes differently than it will affect the probability of severity 2 crashes. For this reason, additional coefficients were added to the model. The Bayesian multinomial regression model calculates a coefficient differently for each of *i* classes when calculating the best-fit line.

In order to account for various intercepts and coefficients, several models were considered. By substituting these various models in for Equation 4-15, the crash severity distributions for each of the facility types can be predicted. The models that were analyzed for this research were the random intercepts model with respect to severity and facility type, random coefficients model with respect to severity and facility type, and random coefficient mixture model with respect to severity and facility type. In addition, a nonsense model was used to evaluate the effectiveness of each model. A nonsense model was created by assuming that all crash severities were equally likely. By creating a nonsense model, each of the other models could be compared to the nonsense model to understand how well the data were predicted by each model. In other words, the nonsense model creates a baseline to which the other models may be compared.

4.6.2 Methods of Evaluation

Three different methods were used in order to evaluate the model fit and predictive ability. The three methods are Bayesian Information Criterion (BIC), Deviance Information Criterion (DIC), and Root Mean Squared Error (RMSE).

BIC evaluates model fit by examining the likelihood of the model given the observed data. Once the likelihood is determined, deviance is extracted. High deviance indicates a low likelihood and poor fit. An additional penalty is assessed to models with additional parameters to account for the loss of the information in adding another covariate. This penalty is known colloquially as the curse of dimensionality. Models with more parameters naturally fit the data better; therefore, a greater penalty is assessed to offset the loss of information that may be present. A high BIC value relative to other BICs of similar models on the same data indicates a relatively poor fit.

DIC is very similar to BIC. However, it is impossible to compare the model fit between models using BIC for evaluation and Bayesian models using DIC for evaluation. The values calculated for BIC and DIC cannot be translated from one to the other. If comparisons are to be made between Bayesian and other types of models, RMSE is recommended.

RMSE evaluates the model's predictive ability by examining how far the calculated estimates are from the actual values observed. For example, if for segment j the model predicts $n_j * p_{ij}$ crashes of severity i, where p_{ij} is the proportion of crashes of severity i on segment j and y_{ij} is the actual number of severity i crashes on segment j, the RMSE is calculated as shown in Equation 4-17.

$$RMSE_i = \sqrt{\frac{1}{N_s}\sum_{j=1}^{N_s}(y_{ij} - n_j * R_{mi})^2} \tag{4-17}$$

Where,

N_s = Total number of segments

y_{ij} = Number of crashes of severity i on segment j

n_j = Total number of crashes on segment j

R_{mi} = Index probability for crash severity i for facility type m, within overall crash severity distribution matrix R

i = Severity of crash

j = Roadway segment

m = Facility type

4.7 Chapter Summary

This chapter presented the methodology required to develop crash severity distributions for nine facility types in the HSM. The datasets used in this analysis were historic AADT, functional class, median, lane, urban code, and crash location. All of these datasets, with the exception of crash location, were downloaded from UDOT's Open Data Portal (UDOT 2017). A roadway data preparation Excel-based spreadsheet developed in previous research was used to segment the roadway based on homogeneity. Several changes were made to the Workbook in order to meet the scope and purpose of the research including adding TWLTL data, median data, crash counts for each severity level, and facility type modifications. The output was a single Excel spreadsheet with all of the data pertaining to each roadway segment. The total number of segments was 1,947 after the expanded definitions of the HSM were applied. The methods for developing the statistical models were also discussed. The four models that were developed for this research were the multivariate regression, frequentist binomial regression, frequentist multinomial regression, and Bayesian multinomial regression models. For evaluation of goodness of model fit to actual crash occurrences, BIC, DIC, and RMSE were used.

5 RESULTS

This chapter presents the results for the crash severity distributions from the straight proportion method and statistical models that were developed for the nine facility types listed in the HSM. First, the results for the straight proportion method are presented. Next, the evaluation results including the crash severity distributions for each of the four statistical models are discussed. The four models include multivariate regression, frequentist binomial regression, frequentist multinomial regression, and Bayesian multinomial regression. Finally, the preferred model for determining the crash severity distributions for the nine facility types is selected. More information regarding the results of the statistical models is provided by Clegg (2018).

5.1 Straight Proportion Method

The straight proportion method was used to understand the crash severity distribution from past crash data. The results for the straight proportion method for the nine facility types are shown in Table 5-1. Although the straight proportion method is quick and simple, it is not the best to determine a crash severity distribution. One shortcoming of taking a straight proportion is that it does not take into account the random nature of crashes like a statistical model would. Another shortcoming of this method is that it only accounts for the facility type to determine the crash severity distribution, whereas statistical models can include a number of relevant variables.

Table 5-1: Crash Severity Distribution for Straight Proportion Method

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7393	0.1046	0.1062	0.0373	0.0126
2	0.7849	0.0952	0.0841	0.0253	0.0105
3	0.7798	0.1116	0.0832	0.0166	0.0088
4	0.6900	0.1764	0.1110	0.0195	0.0031
5	0.6645	0.2024	0.1072	0.0197	0.0062
6	0.6407	0.2132	0.1212	0.0214	0.0035
7	0.6660	0.2064	0.1086	0.0171	0.0019
8	0.6482	0.1977	0.1233	0.0265	0.0043
9	0.7554	0.1410	0.0797	0.0183	0.0056

5.2 Multivariate Regression Model

In multivariate regression for transformed proportions, several assumptions must be met. The assumptions that are important for this analysis are that the proportions change linearly relative to each covariate, crash counts from different segments are independent from one another, residuals are normally distributed, and the residuals have an equal variance across the scope of each covariate, or homoskedacity. Since the analysis also considered multiple non-linear regression, the linearity assumption was relaxed.

While much of the data have linear trends, there are some variables that do not have linear trends. Figure 5-1 shows a linear plot for one variable, combination trailer truck percentage that shows nonlinearity The red line indicates the median of the data, which does not follow a linear trend.

As explained previously in section 4.3, efforts were made to ensure homogeneity within a roadway segment. While there may be some dependence between segments, independence has been accounted for in the data cleaning process; therefore, the crash counts are independent from one another.

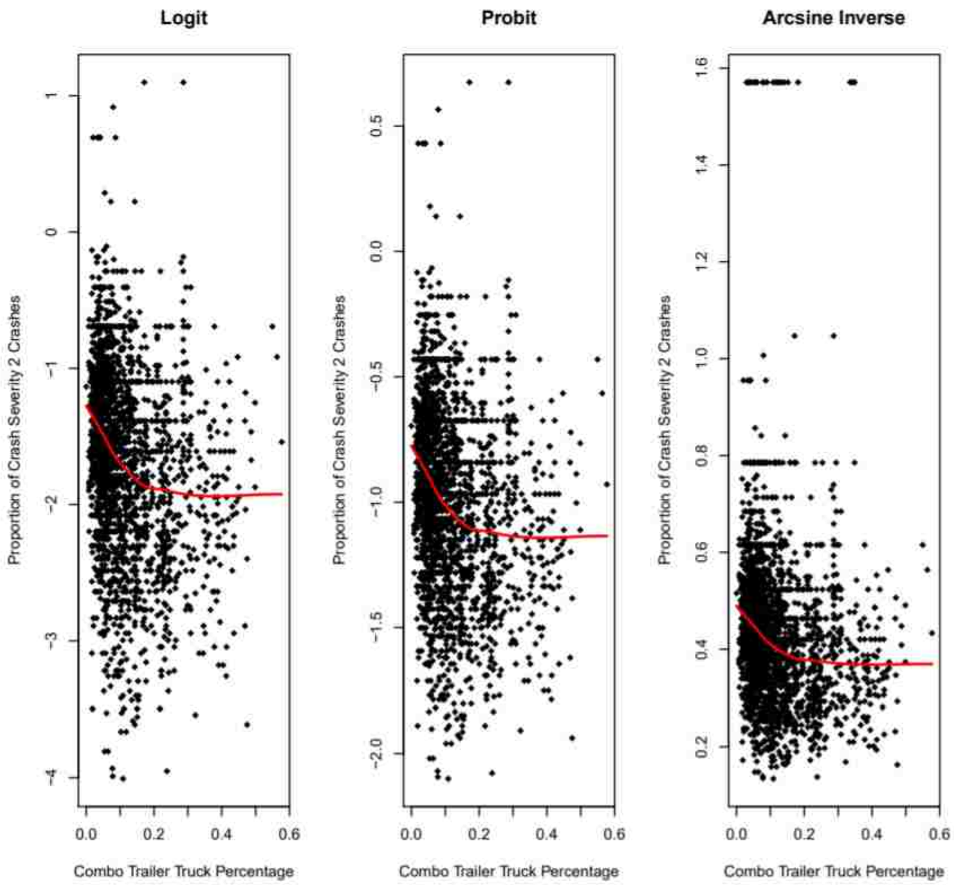


Figure 5-1: Linear plots for percent single trucks (Clegg 2018).

The assumptions that the residuals are normally distributed and have equal variance are unique to the multivariate regression framework. Figure 5-2 show the residual plots, which appear normally distributed, for a multivariate regression model. Additionally, heteroskedacity was examined. It was observed that the equal variance was not guaranteed, which implies that the assumptions that the proportions are normally distributed is unjustified.

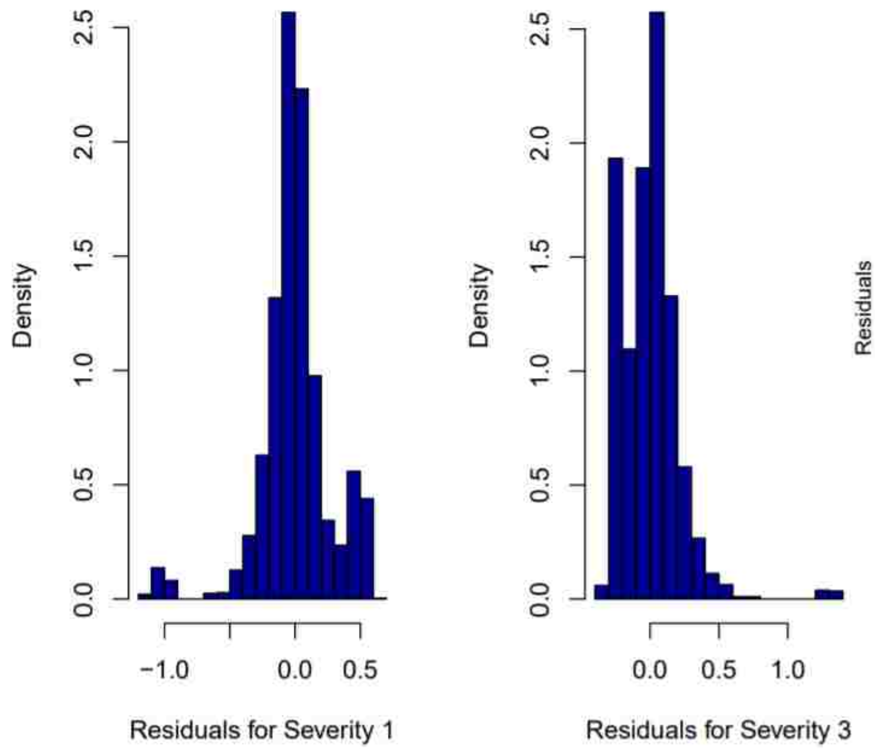


Figure 5-2: Residual plots for multivariate regression (Clegg 2018).

For each of the four models that were explained in section 4.6, many models were considered to achieve the most effective model. An iterative backward variable selection was employed for variable selection in the multivariate regression analysis. The backwards variable selection method began by using all possible covariates. If deleting a covariate lowered the BIC, indicating a better fit, the covariate was removed. This process continued until only three covariates remained: segment length, facility type, and 2014 AADT.

The option was investigated of including county and UDOT region in the model so that crash severity distributions could be specific to smaller geographic areas. The backward variable selection method showed that including either of these variables did not improve model fit. As such, the variables relating to county and UDOT region were not included as variables within any of the four statistical models.

For the multivariate regression, frequentist multinomial regression, and frequentist binomial regression models, nonlinearity was examined using natural splines. Several different numbers of natural splines were included. The best model was chosen using BIC values, as BIC values account for a good model fit while enforcing parsimony, meaning the simplest model with the most predictive power.

For the multivariate regression framework, the best model for all severities is expressed in Equation 5-1, with *f* defined previously in Equation 4-7. The $\mathbb{I}$ symbol refers to an indicator function, which has a value of 1 or 0 depending on the facility type. For example, if Facility Type Code = *m*, the indicator function is assigned a value of 1 because the facility types matches the desired code in the model. The resulting crash severity distribution for the best-fitting model according to BIC for each of the nine facility types is shown in Table 5-2.

$$f(p_{ij}) = \beta_{0i} + \beta_{1m}\left(\mathbb{I}\left(Facility\ Type\ Code_j = m\right)\right) + \sum_{g=1}^{6} \omega_{2g}\left(LENGTH_j\right)\beta_{2gi} + \left(AADT\ 2014_j\right)\beta_{3i} + (Through\ Lanes_j)\beta_{4i} \tag{5-1}$$

Where,
- p_{ij} = Proportion of crash of severity *i* on segment *j*
- β_i = Regression coefficient for severity *i*
- n_j = Total number of crashes on segment *j*
- ω_g = Natural spline function of number *g*
- $LENGTH_j$ = Variable for length of segment *j*
- $AADT\ 2014_j$ = Variable for 2014 AADT value for segment *j*
- $Through\ Lanes_j$ = Number of through lanes for segment *j*
- g = Number of natural spline functions included in the model (6 natural spline functions are included in this model)

i = Severity of crash

j = Roadway segment

m = Facility type

Table 5-2: Crash Severity Distribution for Multivariate Regression Model (Clegg 2018)

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7854	0.0913	0.0902	0.0267	0.0064
2	0.8657	0.0624	0.0563	0.0123	0.0033
3	0.8023	0.1089	0.0746	0.0098	0.0044
4	0.7266	0.1594	0.0977	0.0136	0.0027
5	0.6993	0.1900	0.0960	0.0102	0.0045
6	0.6610	0.2041	0.1169	0.0152	0.0028
7	0.7538	0.1597	0.0753	0.0092	0.0020
8	0.6937	0.1891	0.0970	0.0175	0.0027
9	0.7734	0.1111	0.0880	0.0197	0.0078

5.3 Frequentist Binomial Regression Model

An identical procedure to the multivariate regression models was completed for the Frequentist binomial regression framework. Once again, natural splines were used but were applied to all variables. The best model was chosen based on the median BIC values for the five models of each crash severity. Unlike the multivariate regression models, however, extreme amounts of variability in the diagnostic measures did not exist, indicating that the binomial regression model could account for the other sources of variability.

For the variable selection, the backward variable selection method was used again. The variables that were chosen using this technique were facility type code, length, and AADT 2014. The chosen model for binomial regression is shown in Equation 5-2. The resulting crash severity distribution for the nine facility types is shown in Table 5-3.

$$logit\left(\pi_{imj}\right) = \beta_{0i} + \beta_{1m}\left(\mathbb{I}\left(Facility\ Type\ Code_j = m\right)\right) + \left(LENGTH_j\right)\beta_{2i} + \sum_{g=1}^{7} \omega_{3g}(AADT\ 2014_j)\ \beta_{3gi} \quad (5\text{-}2)$$

Where, π_{imj} = Probability of crash severity i on segment j of facility type m

$LENGTH_j$ = Variable for length of segment j

$AADT\ 2014_j$ = Variable for 2014 AADT value for segment j

n_j = Total number of crashes on segment j

ω_g = Natural spline function of number g

g = Number of natural spline functions included

i = Severity of crash

j = Roadway segment

m = Facility type

Table 5-3: Crash Severity Distribution for Frequentist Binomial Regression Model (Clegg 2018)

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7444	0.1061	0.1026	0.0356	0.0113
2	0.7720	0.0972	0.0920	0.0266	0.0122
3	0.7633	0.1231	0.0890	0.0168	0.0078
4	0.6780	0.1831	0.1140	0.0213	0.0036
5	0.6551	0.2095	0.1073	0.0210	0.0071
6	0.6286	0.2163	0.1291	0.0218	0.0042
7	0.6745	0.2070	0.0997	0.0170	0.0018
8	0.6373	0.2023	0.1277	0.0274	0.0053
9	0.7610	0.1207	0.0869	0.0239	0.0075

5.4 Frequentist Multinomial Regression Model

The multinomial regression models have a different set of assumptions. For multinomial regression, it is required that the transformed proportions change monotonically relative to each covariate. Monotonic relationships occur when one variable increases when another variable increases or when one variable decreases when another variable decreases. It was found that some variables did not have monotonic relationships, such as AADT 2014.

The model that was chosen for the multinomial regression model is illustrated in Equation 5-3.

$$\log\left(\frac{\pi_{\{i\}mj}}{\pi_{\{1\}mj}}\right) = \beta_{0i} + \beta_{1m}\left(\mathbb{I}\left(Facility\ Type\ Code_j = m\right)\right) + \sum_{g=1}^{38} \omega_{2g}(LENGTH_j)\beta_{2gi} + \sum_{g=1}^{38} \omega_{3g}(AADT\ 2014_j)\beta_{3gi} \tag{5-3}$$

Where, $\pi_{\{i\}mj}$ = Probability of crash severity i on segment j within facility type m

ω_g = Natural spline function of number g

g = Number of natural spline functions included in the model (38 natural spline functions are included in this model)

i = Severity of crash

j = Roadway segment

m = Facility type

The covariates included in this model were the same as the frequentist binomial regression models: segment length, 2014 AADT, and facility type. While this multinomial model performed well, it did not vastly outperform the other models considered with regard to RMSE. Also, it dealt with possible monotonicity in AADT and accounted for dependence within elements of the $\boldsymbol{\pi}_{\mathbf{mj}}$ vector. The crash severity distribution relating to the frequentist multinomial regression model is shown in Table 5-4.

Table 5-4: Crash Severity Distribution for Frequentist Multinomial Regression Model (Clegg 2018)

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7679	0.1006	0.0916	0.0314	0.0085
2	0.7855	0.0884	0.0916	0.0248	0.0097
3	0.7637	0.1238	0.0886	0.0183	0.0056
4	0.6985	0.1902	0.0964	0.0136	0.0013
5	0.5918	0.2403	0.1305	0.0303	0.0071
6	0.6535	0.2113	0.1145	0.0178	0.0029
7	0.6680	0.2100	0.1011	0.0195	0.0014
8	0.6404	0.2066	0.1257	0.0234	0.0039
9	0.7770	0.1188	0.0768	0.0221	0.0053

5.5 Bayesian Multinomial Regression Model

The final model under the Bayesian multinomial framework was the most complex considered. It introduces coefficients for every combination of facility type and crash severity. Covariates were selected by the Bayesian Lasso algorithm rather than the backward variable selection algorithm used for the other models. The covariates of segment length, number of through lanes, number of deceleration lanes, percentage of single trucks, percentage of combination trucks, and 2014 AADT had a probability greater than 0.8 of being non-zero.

The DIC was used to determine which of the Bayesian models fit the actual data the best. The DIC values for each of the Bayesian models is shown in Table 5-5. Based on the DIC values, the random coefficients model with different coefficients for each crash severity and facility type combination had the lowest DIC value. Therefore, this model was chosen as the best Bayesian model from this analysis. The model can be written as shown in Equation 5-4. The resulting crash severity distribution for the nine facility types for the best Bayesian multinomial regression model is illustrated in Table 5-6.

Table 5-5: DIC Values for Bayesian Models (Clegg 2018)

Model	DIC Value
Random Intercepts (Severity)	28271
Random Intercepts (Severity & Facility Type)	27230
Random Coefficients (Severity)	27186
Random Coefficients (Severity & Facility Type)	**26577**
Nonsense	51590
Mixed	27786

$$\log\left(\frac{\pi_{\{i\}j}}{\pi_{\{1\}j}}\right) = \beta_{0i} + X_{ij}\beta_{1i} + \cdots + X_{kj}\beta_{ki} \qquad \text{(5-4)}$$

Where, π_{ij} = Probability of crash severity *i* on segment *j*

i = Severity of crash

j = Roadway segment

K = 7 with covariates for each of the seven chosen variables

Table 5-6: Crash Severity Distribution for Bayesian Multinomial Regression Model (Clegg 2018)

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7308	0.1138	0.1079	0.0364	0.0111
2	0.7958	0.1062	0.0728	0.0192	0.0060
3	0.8308	0.1036	0.0591	0.0019	0.0046
4	0.6839	0.1784	0.1110	0.0239	0.0028
5	0.6896	0.1852	0.1008	0.0200	0.0044
6	0.6435	0.2073	0.1243	0.0211	0.0038
7	0.6522	0.2218	0.1069	0.0185	0.0006
8	0.6433	0.2091	0.1184	0.0248	0.0044
9	0.7569	0.1098	0.0970	0.0240	0.0123

5.6 Final Model Selection

A final comparison between each of the statistical models was made using BIC values. RMSE would be the best diagnostic, but because the predictions are in terms of decimals while observed crashes are in terms of discrete numbers, RMSE is an unreliable diagnostic. Nevertheless, the RMSE values were found to be informative and were included in Table 5-7. Table 5-8 displays the BIC values for the best models from each framework.

Table 5-7: RMSE Values for Best Models (Clegg 2018)

Model	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
Freq MLR	5.6509	3.9571	3.2292	1.4151	0.5788
Freq Logit	5.4687	3.7006	3.2177	1.4228	0.5707
Freq Multinomial	5.3610	3.7291	3.0640	1.4036	0.5488
Random Intercepts (Severity)	5.3721	3.8134	3.0586	1.3852	0.5688
Random Intercepts (Severity & Facility Type)	5.3310	3.5141	3.0582	1.3385	0.5432
Random Coefficients (Severity)	5.3673	3.6241	3.0288	1.3383	0.5465
Random Coefficients (Severity & Facility Type)	5.3324	3.5098	3.0508	1.3345	0.5406
Nonsense	3.3722	7.9741	3.4130	4.1216	5.0276
Mixed	5.3315	3.6384	3.0967	1.4090	0.5764

Table 5-8: BIC Values for the Best Model from Each Framework (Clegg 2018)

Model	BIC
Multivariate Regression	-15342
Frequentist Binomial Regression	10493
Frequentist Multinomial Regression	24927577
Bayesian Multinomial Regression	30666

Based on the RMSE values, the frequentist multinomial regression and binomial regression models appear to perform well. However, the BIC values show that the multinomial regression does not fit the data well. Additionally, the likelihood for the multivariate regression and the binomial regression is misapplied because the assumption that the data are distributed according to the multivariate distribution or frequentist binomial distributions cannot be met. Additionally, vehicle crashes are believed to be distributed according to a multinomial distribution.

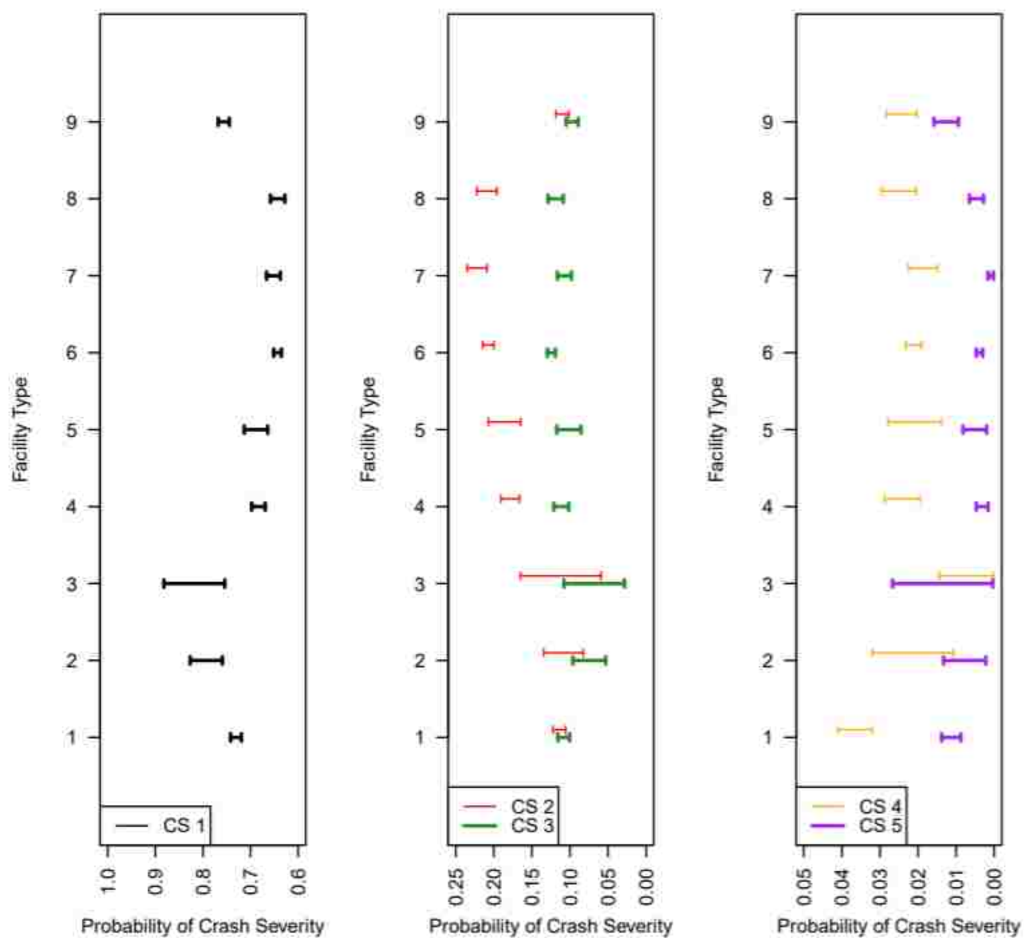
Based on the BIC and RMSE values, it was concluded that the Bayesian multinomial regression model predicted the crash severity distributions more accurately compared to the other models analyzed. The Bayesian models are more flexible in evaluating effects of certain similar segments of roads. The Bayesian model is also favored due to its interpretability. The crash severity distribution for the Bayesian multinomial regression model was shown previously in Table 5-6. In addition, there is a 95 percent probability that the crash severity distribution will fall within the respective values in Table 5-9 and Table 5-10. Figure 5-3 shows the 95 percent credible intervals for the Bayesian multinomial regression model. CS 1-5 refer to crash severity, with 1 representing PDO and 5 representing fatal crashes.

Table 5-9: 95 Percent Credible Upper Bound for Bayesian Multinomial Regression Model (Clegg 2018)

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7419	0.1223	0.1159	0.0411	0.0138
2	0.8269	0.1343	0.0962	0.0320	0.0133
3	0.8819	0.1650	0.1080	0.0144	0.0267
4	0.6975	0.1905	0.1211	0.0288	0.0047
5	0.7132	0.2072	0.1174	0.0279	0.0082
6	0.6514	0.2146	0.1298	0.0232	0.0047
7	0.6664	0.2350	0.1165	0.0225	0.0016
8	0.6581	0.2226	0.1289	0.0296	0.0065
9	0.7679	0.1185	0.1051	0.0283	0.0158

Table 5-10: 95 Percent Credible Lower Bound for Bayesian Multinomial Regression Model (Clegg 2018)

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7192	0.1057	0.1005	0.0320	0.0088
2	0.7594	0.0825	0.0532	0.0107	0.0022
3	0.7542	0.0593	0.0286	0.0001	0.0004
4	0.6693	0.1665	0.1017	0,0195	0.0016
5	0.6642	0.1649	0.0858	0.0139	0.0020
6	0.6352	0.1999	0.1192	0.0191	0.0030
7	0.6375	0.2090	0.0980	0.0149	0.0002
8	0.6278	0.1960	0.1087	0.0205	0.0028
9	0.7446	0.1018	0.0892	0.0203	0.0094



(a) (b) (c)

Figure 5-3: 95 percent credible intervals for Bayesian multinomial regression model: (a) crash severity 1, (b) crash severity 2 and 3, and (c) crash severity 4 and 5 (Clegg 2018).

5.7 Crash Severity Distribution Comparison

Once the Bayesian multinomial crash severity distribution was chosen, a comparison between the chosen model, the straight proportion method, and the HSM crash severity distributions was performed. The comparison of these crash severity distributions is shown in Figure 5-4, where FT refers to the facility type that the distribution represents. The crash severity distribution found within the HSM is developed using roadway data of facility type 1, rural TLTW roadways. Thus, the crash severity distribution is compared to the other crash severity distributions for facility type 1. Though the crash severity distributions for the straight proportion method and Bayesian multinomial model are nearly identical, there are some differences between these distributions and the distribution from the HSM. The difference between the proportion of crashes for crash severity 1 in the HSM and the Bayesian multinomial model is 5.2 percent. For crash severity 2 and crash severity 4, the differences are 3.1 and 1.8 percent, respectively.

The distributions for the straight proportion method and the Bayesian multinomial model for facility types 2 through 9 can also be compared to understand how the results for each method differ from each other. The largest discrepancies between the Bayesian multinomial statistical model and the straight proportion method occur for facility types 3 and 5. For example, the difference between the crash severity 1 proportion between the Bayesian multinomial statistical model and the straight proportion method for facility type 3 is about 5.1 percent. Likewise, for crash severity 3 the difference is approximately 2.4 percent for facility type 3.

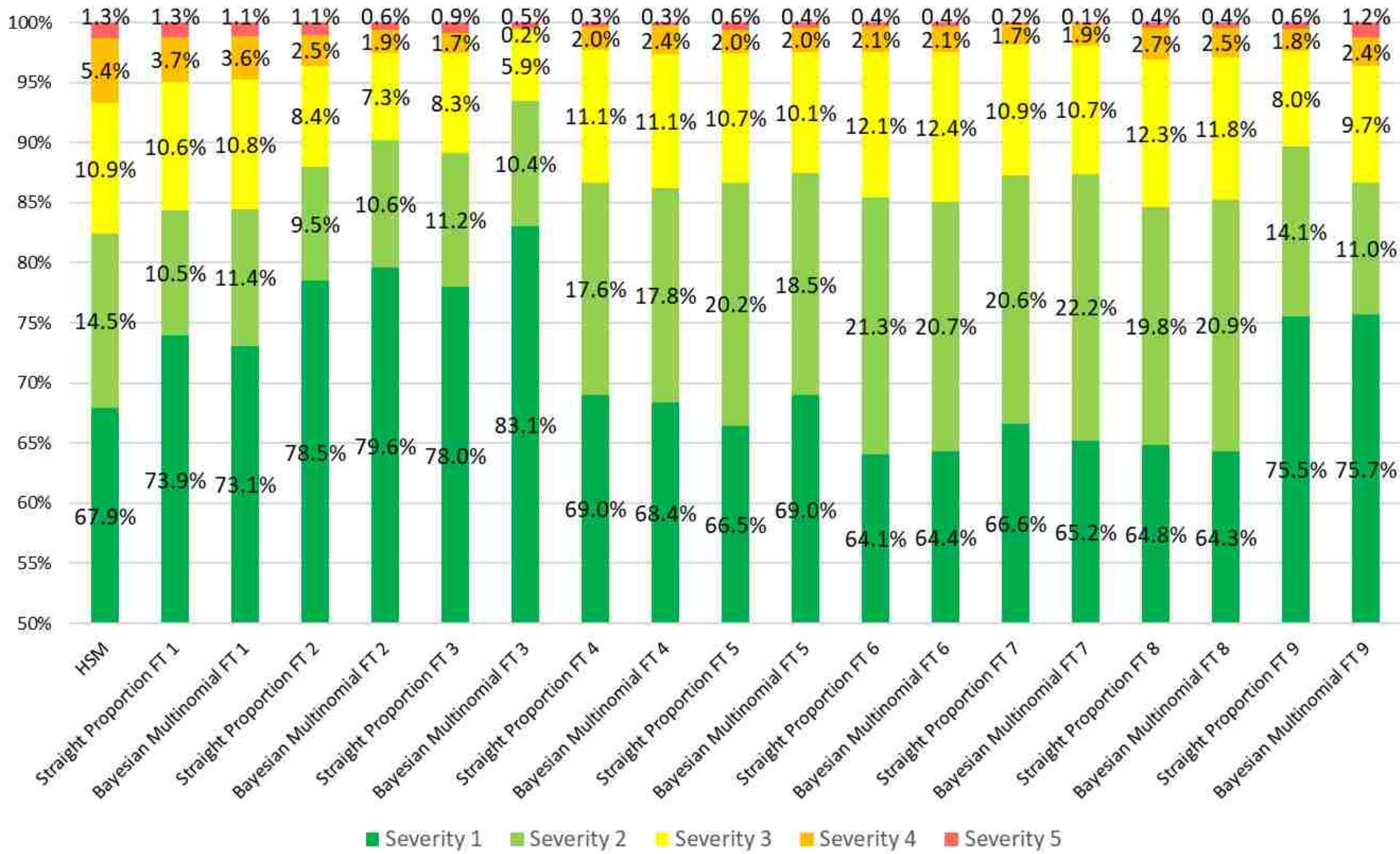


Figure 5-4: Crash severity distribution comparison between HSM (facility type 1), straight proportion method, and Bayesian multinomial model.

Due to the crash costs associated with the crash severity levels in life-cycle benefit-cost analysis, the discrepancies in more severe crashes will have a larger impact in life-cycle benefit-cost analysis. As stated previously, the crash severity distribution for the straight proportion method was not chosen because it is not adequate to describe the contributing factors of crashes. The straight proportion method uses facility type as the only variable to develop the crash severity distributions, whereas the Bayesian multinomial model has a number of variables that are taken into account.

5.8 Chapter Summary

This chapter presented the results for the crash severity distributions for the four models developed for this research. To choose the best model for each framework, it was necessary for the models to be cross-validated. The models were cross-validated using appropriate BIC, DIC, and RMSE values. The model specifications were presented with the significant covariates. In addition, the crash severity distributions for the nine facility types were presented. Next, a single model, the Bayesian multinomial regression model, was presented as the most suitable crash severity distribution. The crash severity distributions for the Bayesian multinomial statistical model, the straight proportion method, and the HSM distribution were compared.

6 CONCLUSIONS AND RECOMMENDATIONS

The purpose of this research was to develop crash severity distributions for the 11 facility types outlined in the HSM. Due to insufficient data, crash severity distributions were developed only for the first nine facility types. The preceding chapters have discussed the procedures used to complete the analysis, including a crash severity distribution survey that was distributed to each of the 50 DOTs in the United States, the roadway segmentation process, and the development of several statistical models.

Due to insufficient research for the development of crash severity distributions, the research team conducted a brief survey to understand the uses of crash severity distributions in relation to life-cycle benefit-cost analysis within DOTs. Next, an existing automated Excel workbook was modified to analyze roadway data segment based on a change in roadway characteristics. The roadway segments were classified based on the HSM facility types. Changes were made to the definition of the facility types in order to include more roadway segments in the analysis. The expanded definition of the facility types included additional through lanes. Statistical models were developed using R-programming. The four models that were developed for this research were the multivariate regression, frequentist binomial regression, frequentist multinomial regression, and Bayesian multinomial regression models.

This chapter summarizes the methodology and final results of this research. First, the roadway segmentation process is summarized. Next, the statistical models developed for this project are discussed. Finally, recommendations and future research opportunities are presented.

6.1 Roadway Segment Summary

After the roadway data were acquired, an automated workbook created in previous BYU research was modified to segment the data using several roadway characteristics. The median type roadway characteristic was included. Although 10 types of medians were included in the dataset, each type was classified as divided or undivided, as the HSM only specifies these two median types. The definitions of the facility types were expanded to include more through lanes. By changing the definition of the facility types, more segments were included in the analysis. Additionally, TWLTLs were also added to the data segmentation process. Once changes were made to the segmentation process, it was observed that several adjacent segments had the same facility type. The only difference between adjacent segments was AADT values. Adjacent segments on the same route were then condensed to reduce the occurrence of short segments. In such cases, a weighted average, based on segment length, of the AADT of combined segments was entered.

6.2 Statistical Model Summary

Once the roadway segmentation process was complete, the data were entered into four statistical models to develop crash severity distributions. The four models that were developed for this research were the multivariate regression, frequentist binomial regression, frequentist multinomial regression, and Bayesian multinomial regression models. To evaluate the models, a cross-validation study was conducted to select the best model for each framework. BIC, DIC, and RMSE values were compared to conduct the analysis. Based on the cross-validation study, it was determined that the Bayesian multinomial regression model was the most effective model to describe the crash severity distributions for the nine facility types evaluated. Table 6-1 shows the crash severity distributions for each facility type for the Bayesian multinomial regression model.

Table 6-1: Crash Severity Distribution for Bayesian Multinomial Regression Model (Clegg 2018)

Facility Type	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
1	0.7308	0.1138	0.1079	0.0364	0.0111
2	0.7958	0.1062	0.0728	0.0192	0.0060
3	0.8308	0.1036	0.0591	0.0019	0.0046
4	0.6839	0.1784	0.1110	0.0239	0.0028
5	0.6896	0.1852	0.1008	0.0200	0.0044
6	0.6435	0.2073	0.1243	0.0211	0.0038
7	0.6522	0.2218	0.1069	0.0185	0.0006
8	0.6433	0.2091	0.1184	0.0248	0.0044
9	0.7569	0.1098	0.0970	0.0240	0.0123

6.3 Recommendations and Future Research

Several suggestions for future research are presented based on the findings of this research. Although the data used in this research worked well throughout the process, one issue with the data was that there were circumstances that involved tedious manipulation to the data in order to achieve the results wanted. UDOT's Light Detection and Ranging (LiDAR) data are of extreme precision but can cause some problems when the data are not required to be at a high level of precision. It is, therefore, recommended that a collection of datasets of varying precision be developed to meet various analysis purposes. For example, for many analyses, a more general, less precise dataset can be used. For this project, the extremely precise LiDAR data were difficult to use at times. By having various datasets, the user will be able to choose a dataset that will meet the needs of his or her research.

Another recommendation for future research is to focus on collecting data for the additional facility types. Originally, the scope of this project included developing crash severity distributions for all 11 facility types; however, due to insufficient data, only nine facility types were developed. The two facility types that were not included were freeway speed change lanes and freeway ramps. In order to develop crash severity distributions for these omitted facility types, additional roadway data and crash data will be required. Perhaps the largest change of data suggested will be in recording the specific lanes in which crashes occur.

The last recommendation for future research is to combine the process of life-cycle benefit-cost analysis with the roadway safety research that continues to be developed. One of the outputs for the safety research is a list of countermeasures to implement in order to increase the safety for that roadway segment. These countermeasures can then be evaluated in the life-cycle benefit-cost analysis to help engineers and decision-makers with choosing the best option for the roadway in terms of safety improvements. Automating this procedure can give the engineer an idea of how much the countermeasure will cost and which countermeasures could be excluded due to cost constraints.

REFERENCES

American Association of State and Highway Transportation Officials (AASHTO). (2010). Highway Safety Manual. Washington, DC.

American Association of State and Highway Transportation Officials (AASHTO). (2018). Subcommittee on Safety Management. < https://safetymanagement.transportation.org/membership/> (April 4, 2017).

Clegg, B. W. (2018). "Predictive Crash Severity Distribution for Utah State Roadways Based on Facility Type." Department of Statistics, Brigham Young University, Provo, UT.

New York State Department of Transportation (NYSDOT). (2013). Average Accident Costs/Severity Distribution State Highways 2011. <https://www.dot.ny.gov/divisions/operating/osss/highway-repository/Revised2010_2011AvrAccCostSev.pdf> (May 7, 2018).

Qualtrics. (2017). Welcome to the Experience Management Platform. <https://www.qualtrics.com/> (November 22, 2017).

Saito, M., Frustaci, J. B., and Schultz, G. G. (2016). "Life-Cycle Benefit-Cost Analysis of Safety Related Improvement on Roadways." Report UT-16.15, Utah Department of Transportation Traffic and Safety Research Division, Salt Lake City, UT.

Saito, M., Schultz, G. G., and Brimley , B. K. (2011). "Transportation Safety Data and Analysis, Volume 2: Calibration of the Highway Safety Manual and Development of New Safety Performance Functions," Report UT-10.12b, Utah Department of Transportation Traffic and Safety Research Division, Salt Lake City, UT.

Schultz, G. G., Bassett, D., Roundy, R., Saito, M., and Reese, C.S. (2015). "Use of Roadway Attributes in Hot Spot Identification and Analysis." Report UT-15.10, Utah Department of Transportation Traffic and Safety, Research Division, Salt Lake City, UT.

Schultz, G. G., Mineer, S.T., Saito, M., Gibbons, J. D., Siegal, S. A., and MacArthur, P. D. (2016). "Roadway Safety Analysis Methodology for Utah." Report UT-16.13, Utah Department of Transportation Traffic and Safety, Research Division, Salt Lake City, UT.

Treat, J. R., Tumbas, N. S., McDonald, S.T., Dhinar, D., Hume, R. D., Mayer, R. E., Stansifer, R. L., and Castellan, N. J. (1979). Tri-Level Study of the Causes of Traffic Crashes: Final

Report- Executive Summary Report No. DOT-HS-034-3-535-79-TAC(S). Institute for Research in Public Safety, Bloomington, IN.

Utah Department of Transportation (UDOT). (2017). Open Data Portal. <http://udot.uplan. opendata.arcgis.com/> (February 2, 2017).

Utah Department of Transportation (UDOT). (2016). Zero Fatalities: A Goal We Can Live With. "Utah Strategic Highway Safety Plan." <http://ut.zerofatalities.com/downloads?SHSP-ZeroFatalities.pdf> (May 7, 2018).

Vermont Agency of Transportation (VTrans). (2005). Highway Safety Improvement Program.

Wall, D. (2016). Utah Department of Transportation. Personal Communication.

APPENDIX A SURVEY

Appendix A includes the details for the survey that was distributed as part of this research. This section includes the survey questions and survey flow and further details regarding the survey results.

A.1 Survey Questions and Survey Flow

The following is the crash severity distribution survey that was sent to each of the 50 State DOTs in the United States. Figure A.1 shows the survey flow.

This is a survey conducted by the Brigham Young University research team to determine the uses of crash severity distributions by State DOT across the United States in conducting life-cycle benefit-cost analyses of safety improvement countermeasures. This survey will take approximately 5 minutes to complete. Completing this survey is voluntary. Please answer each question honestly. The survey is tallied by Qualtrics software. The names of respondents will be kept confidential and will not be reported in any reports, including the final report produced in this study. Please allow those who are most familiar with these subjects to be the representatives for your DOT.

Life-cycle benefit-cost analyses require crash severity distributions in order to predict the types of crashes that will occur on a roadway segment. A crash severity distribution describes the distribution of crash severities for a roadway type, segment or network. There is a single default crash severity distribution described on pages 10-14 through 10-17 of Volume 2 of the Highway Safety Manual for rural two-lane, two-way roads, which is shown below. The Highway Safety Manual encourages state and local agencies to adopt their own crash severity distributions based on their respective crash database. The purpose of this survey is to understand the crash severity distributions that are currently being used or implemented throughout the United States.

Crash Severity Distribution	
Fatal	1.3%
Incapacitating Injury	5.4%
Non-incapacitating Injury	10.9%
Possible Injury	14.5%
Property Damage Only	67.9%
Total	100.0%

Introduction Questions

1. Which State Department of Transportation do you represent?

2. What position do you hold at your DOT?

3. Do you use life-cycle benefit-cost analysis to analyze the cost-effectiveness of safety-related countermeasures?
    a) Yes
    b) No
    c) I don't know

Block A

1. When did you begin using the life-cycle benefit-cost analysis you currently use?
    a) I don't know
    b) More than 10 years ago
    c) Between 5 and 10 years ago
    d) Between 1 and 5 years ago
    e) Less than 1 year ago

2. What crash severity distribution(s) do you use in your life-cycle benefit-cost analysis?
    a) A single distribution taken from the Highway Safety Manual (There is currently only one for rural two-lane two-way highways)
    b) A single distribution derived from our state's crash data
    c) Multiple distributions derived from our state's crash data
    d) None of the above. Other reason specified below. ____________________

Block B

1. Has the DOT you represent considered using different crash severity distribution based on your state's crash data?
    a) Yes, we already have crash severity distributions based on our state's crash data.
    b) Yes, we are currently researching crash severity distributions for our state using our state's crash data.
    c) Yes, but we have not yet started to research it.
    d) No, we feel the crash severity distribution we currently use is sufficient.
    e) No, we don't use crash severity distributions.
    f) No. Other reason specified below. ____________________

2. Has the DOT you represent considered using different crash severity distributions specific to the 11 facility types described in the Highway Safety Manual?
    a) Yes, we already have crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    b) Yes, we are currently researching crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    c) Yes, but we have not yet started to research crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    d) No, we feel the crash severity distribution we currently use is sufficient.
    e) No, we don't use crash severity distributions.
    f) No. Other reason specified below. ____________________

3. Does your DOT have literature relating to the derivation of crash severity distributions for various facility types? Please upload any files in a single compressed file.

4. May we contact you if we have questions about your answers?
    a) Yes
    b) No

Block C

1. Has the DOT you represent considered using different crash severity distributions specific to the 11 facility types described in the Highway Safety Manual?
    a) Yes, we already have crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    b) Yes, we are currently researching crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    c) Yes, but we have not yet started to research crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    d) No, we feel the crash severity distribution we currently use is sufficient.
    e) No, we don't use crash severity distributions.
    f) No. Other reason specified below. ____________________

2. Does your DOT have literature relating to the derivation of crash severity distributions for various facility types? Please upload any files in a single compressed file.

3. May we contact you if we have questions about your answers?
    a) Yes
    b) No

Block D

1. Has the DOT you represent considered using different crash severity distribution based on your state's crash data?
    a) Yes, we already have crash severity distributions based on our state's crash data.
    b) Yes, we are currently researching crash severity distributions for our state using our state's crash data.
    c) Yes, but we have not yet started to research it.
    d) No, we feel the crash severity distribution we currently use is sufficient.
    e) No, we don't use crash severity distributions.
    f) No. Other reason specified below. ____________________

2. Has the DOT you represent considered using different crash severity distributions specific to the 11 facility types described in the Highway Safety Manual?
    a) Yes, we already have crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    b) Yes, we are currently researching crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    c) Yes, but we have not yet started to research crash severity distributions based on the 11 facility types described in the Highway Safety Manual.
    d) No, we feel the crash severity distribution we currently use is sufficient.
    e) No, we don't use crash severity distributions.
    f) No. Other reason specified below. ____________________

3. Does your DOT have literature relating to the derivation of crash severity distributions for various facility types? Please upload any files in a single compressed file.

4. Please describe the crash severity distributions used in your life-cycle benefit-cost analysis. Do you have crash severity distributions for certain facility types? How did you derive your distributions? If literature is available, please attach for reference on the next question.

5. Please attach any literature that is available regarding the previous question. Please upload any files in a single compressed file.

6. When did you begin using the crash severity distributions you currently use?
    a) I don't know.
    b) More than 10 years ago
    c) Between 5 and 10 years ago
    d) Between 1 and 5 years ago
    e) Less than 1 year ago

7. What benefits have you seen from using your crash severity distributions in your crash-related analyses?

8. May we contact you if we have questions about your answers?
    a) Yes
    b) No

Conclusion

If you have questions about this survey, you may contact Dr. M. Saito. Please use the following address when you would like to mail printed materials to us.

If you have questions regarding your rights as a participant in research projects, you may contact Dr. Shane S. Schulthies, Chair of the Institutional Review Board for Human Subjects, 120B RB, Brigham Young University, Provo, UT 84602; phone, (801) 422-5490

Please advance this survey to submit your results. We appreciate your time to participate in this survey.

Block E

1. Has the DOT you represent considered using a life-cycle benefit-cost analysis to analyze the cost-effectiveness of countermeasures to improve safety?
   a) Yes, we are currently researching life-cycle benefit-cost analysis.
   b) Yes, but we have not yet started to research it.
   c) Yes, we have used it in the past but have stopped using it.
   d) No, we are not interested in using it.
   e) No. Other reason specified. ____________________

2. What alternatives methods do you use in order to analyze the cost-effectiveness of countermeasures to improve safety? Please be specific.

3. Does your DOT have literature relating to the derivation of crash severity distributions for various facility types? Please upload any files in a single compressed file.

4. May we contact you if we have questions about your answers?
   a) Yes
   b) No

Other Contact

1. Is there someone else at your DOT that we might contact in order to determine the uses of crash severity distributions at your DOT in conducting life-cycle benefit-cost analyses of safety improvement countermeasures?
   a) Yes
   b) No

Other Contact Information

1. What is their contact information?
   a) Name: ____________________
   b) Phone Number: ____________________
   c) Email Address: ____________________

Contact Question

1. May we contact you if we have questions about your answers?
   a) Yes
   b) No

Contact Information

1. What is your contact information?
   a) Name ____________________
   b) Phone Number ___________________
   c) Email Address ____________________

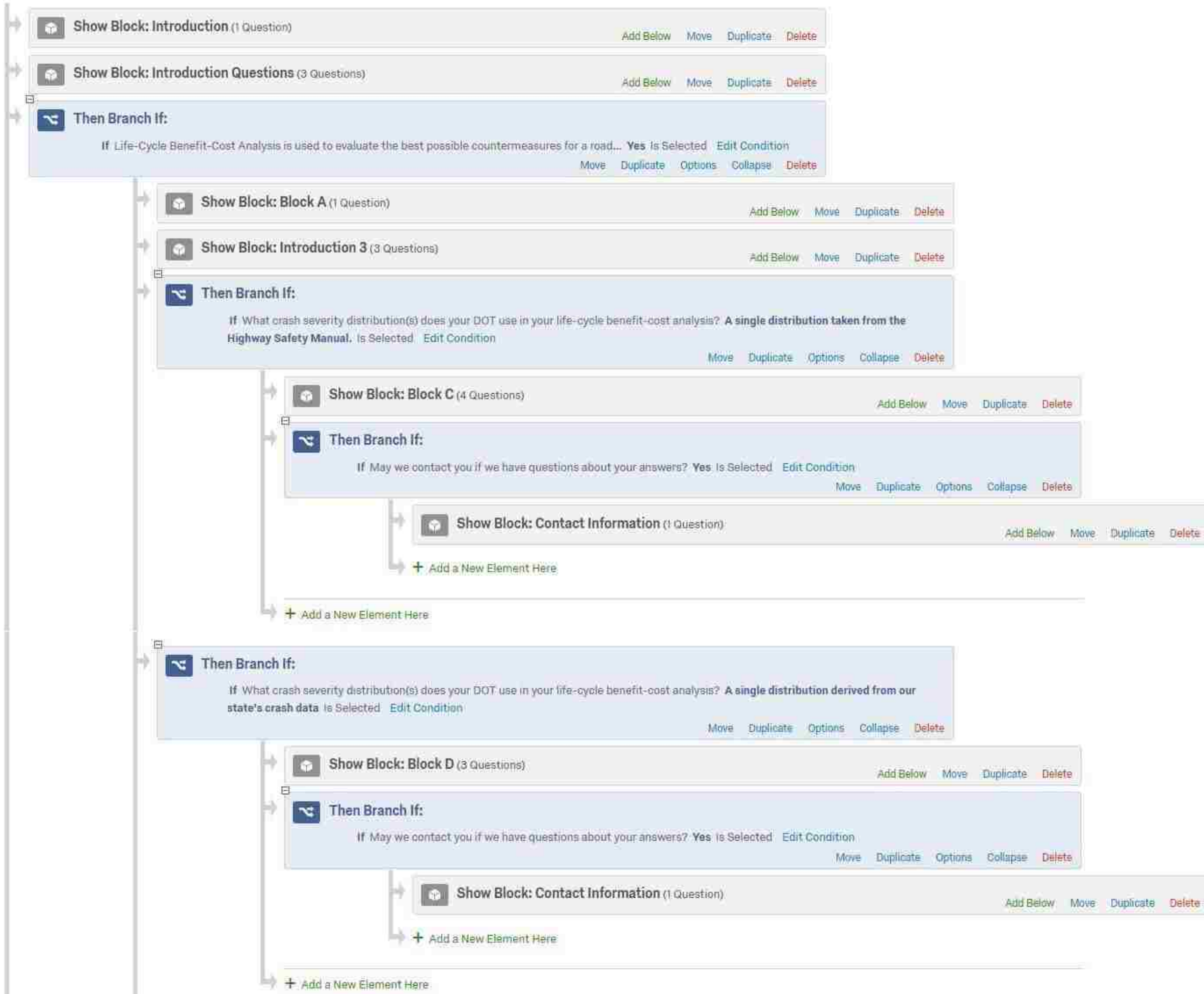


Figure A.1 Crash Severity Distribution Survey Flow

Figure A.1 Continued

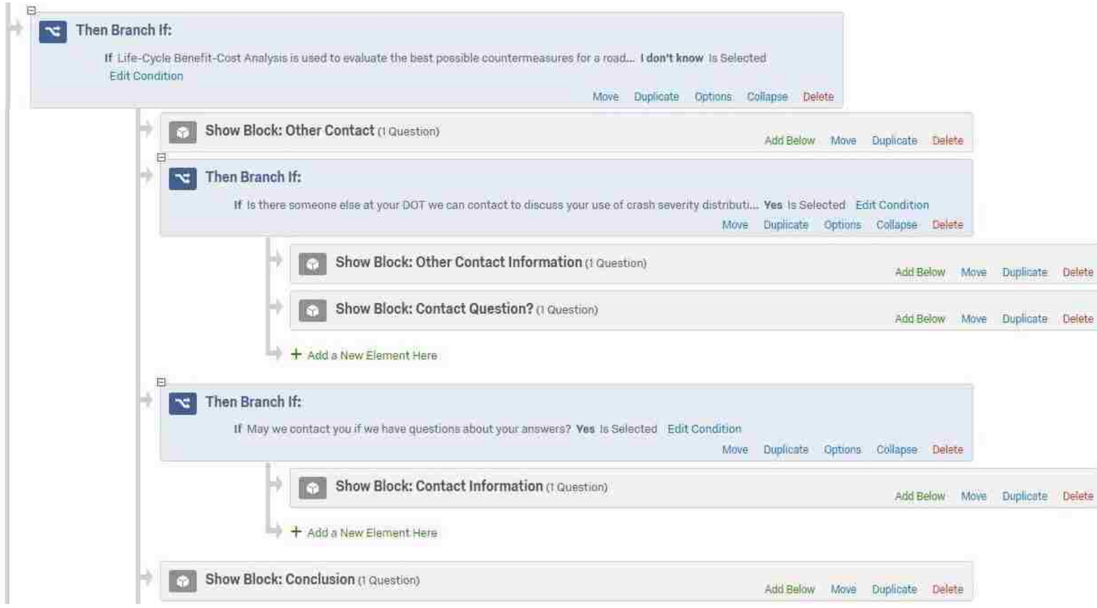


Figure A.1 Continued

APPENDIX B CRITICAL DATA COLUMNS

Appendix B is a collection of tables that provide a list of the critical data columns needed for each dataset. These columns are used in the automated Excel workbook to segment data or combine crash files.

B.1 Roadway Characteristic Datasets

The critical columns for each of the roadway characteristic datasets received from UDOT Traffic and Safety Division are outlined in Table B.1 through Table B.4. These data columns are crucial in the use of the Roadway Characteristic Data portion of the automated Excel workbook.

Table B.1: Critical Data Columns for Functional Class

Heading	Description
ROUTE	Route ID: numeric route number of a given road segment
BEGMP	Beginning Mile Point: beginning milepoint of the road segment
ENDMP	End Mile Point: ending milepoint of the road segment
FC_CODE	Functional Class: number representing the functional class type of the segment

Table B.2: Critical Data Columns for Median Data

Heading	Description
ROUTE_NAME	Route ID: Route ID number with direction letter (i.e., 0089N)
ROUTE_DIR	Direction: Route direction (P, N)
START_ACCUM	Beginning Mile Point: The mile point where the sign appears
END_ACCUM	End Mile Point: The end mile point of the road segment
MEDIAN_TYP	Medan Type: the type of median for the road segment

Table B.3: Critical Data Columns for Lane Data

Heading	Description
ROUTE_NAME	Route ID: numeric route number for a given road segment
START_ACCUM	Beginning Mile Point: beginning mile point of the road segment
END_ACCUM	End Mile Point: end mile point of the road segment
THRU_LANE	Through Lanes: number of through lanes
DECELL_LAN	Deceleration Lanes: number of deceleration lanes
TWO_WAY_LE	Two-Way Left-Turn Lanes (TWLTL): number of TWLTLs
ACCELL_LANE	Acceleration Lanes: number of acceleration lanes
PASSING_LANE	Passing Lanes: number of passing lanes

Table B.4: Critical Data Columns for Urban Code

Heading	Description
ROUTE_NAME	Route ID: numeric route number for a given road segment
START_ACCUM	Beginning Mile Point: beginning mile point of the road segment
END_ACCUM	End Mile Point: end mile point of the road segment
URBAN_CODE	Urban Code: number that represents a description of the surrounding area
URBAN_DESC	Urban Description: description of the surrounding area (i.e., Small-Urban, St. George, rural, etc.)

B.2 Crash Datasets

The critical columns for each of the datasets received from UDOT Traffic and Safety Division are outlined in Table B.5 through Table B.8. These data columns are crucial in the use of the Crash Data portion of the automated Excel workbook.

Table B.5: Critical Columns for Crash Data

Heading	Description
CRASH_ID	Crash ID: unique crash ID number for each crash
CRASH_DATETIME	Crash Date/Time: date and time of crash
CRASH_SEVERITY_ID	Crash Severity ID: numerical severity level of crash (i.e., 1-5)
LIGHT_CONDITION_ID	Light Condition: ID for light condition at time of crash (i.e., 1-6, 88-99)
WEATHER_CONDITION_ID	Weather Condition: ID for weather condition at time of crash (i.e., 1-9, 88-99)
MANNER_COLLISION_ID	Manner Collision: ID for manner of collision in crash (i.e., 1-8, 88-99)
PAVEMENT_ID	Pavement: ID for pavement type (i.e., 1-4, 88-99)
ROADWAY_SURF_CONDITION_ID	Roadway Surface Condition: ID for roadway surface conditions (i.e., 1-9, 88-99)
ROADWAY_JUNCT_FEATURE_ID	Roadway Junction Feature: ID for roadway junction feature (i.e.,1-10, 20-26, 88-99)
WORK_ZONE_RELATED_YNU	Work Zone Related: Y/N to determine whether crash occurred in work zone
WORK_ZONE_WORKER_PRESENT_YNU	Work Zone Worker Present: Y/N to determine whether worker present in work zone
HORIZONTAL_ALIGNMENT_ID	Horizontal Alignment: ID for horizontal curvature of roadway (i.e., 1-2, 88-99)
VERTICAL_ALIGNMENT_ID	Vertical Alignment: ID for vertical curvature of roadway (i.e., 1-4. 88-99)
ROADWAY_CONTRIB_CIRCUM_ID	Roadway Contributing Circumstance: ID for vehicle contributing circumstance related to the crash (i.e., 0-18, 88-99)
FIRST_HARMFUL_EVENT_ID	First Harmful Event: ID for first harmful event resulting from the crash (i.e., 0-62, 88-99)

Table B.6: Critical Data Columns for Crash Location

Heading	Description
CRASH_ID	Crash ID: unique crash ID number for each crash
ROUTE	Route ID: numeric route number for a given road segment
ROUTE_DIRECTION	Direction: route direction (i.e., P, N, or X)
RAMP_ID	Ramp ID: ID indicating a ramp and the type (i.e., 1-4, CD)
MILEPOINT	Mile Point: mile point location of the crash

Table B.7: Critical Columns for Crash Rollup Data

Heading	Description
CRASH_ID	Crash ID: unique crash ID number for each crash
NUMBER_VEHICLES_INVOLVED	Number Vehicles Involved: number of vehicles involved in the given accident
NUMBER_FATALITIES	Number of Fatalities: number of person-fatalities resulting from a given crash
NUMBER_FOUR_INJURIES	Number of Incapacitating Injuries: number of person-incapacitating injuries resulting from a given crash
NUMBER_THREE_INJURIES	Number of Injuries: number of person-injuries resulting from a given crash
NUMBER_TWO_INJURIES	Number of Possible Injuries: number of person-possible injuries resulting from a given crash
NUMBER_ONE_INJURIES	Number of Property Damage Only Events: number of events for property damage only resulting from a given crash
PEDESTRIAN_INVOLVED	Pedestrian Involved: Y/N to determine whether a pedestrian was involved in the crash
BICYCLIST_INVOLVED	Bicyclist Involved: Y/N to determine whether a bicyclists was involved in the crash
MOTORCYCLE_INVOLVED	Motorcycle Involved: Y/N to determine whether a motorcycle was involved in the crash
IMPROPER_RESTRAINT	Improper Restraint: Y/N to determine whether improper restraint was a factor in the crash
UNRESTRAINED	Unrestrained: Y/N to determine whether a driver/passenger was unrestrained in the crash

Table B.7 Continued

Heading	Description
DUI	DUI: Y/N to determine whether driving under the influence was a factor in the crash
AGGRESSIVE_DRIVING	Aggressive Driving: Y/N to determine whether aggressive driving was a factor in the crash
DISTRACTED_DRIVING	Distracted Driving: Y/N to determine whether distracted driving was a factor in the crash
DROWSY_DRIVING	Drowsy Driving: Y/N to determine whether drowsy driving was a factor in the crash
SPEED_RELATED	Speed Related: Y/N to determine whether speed was a factor in the crash
INTERSECTION_RELATED	Intersection Related: Y/N to determine whether the crash occurred at an intersection
ADVERSE_WEATHER	Adverse Weather: Y/N to determine whether adverse weather was a factor in the crash
ADVERSE_ROADWAY_SURF_CONDITION	Adverse Roadway Surface Conditions: Y/N to determine whether adverse roadway surface conditions were a factor in the crash
ROADWAY_GEOMETRY_RELATED	Roadway Geometry Related: Y/N to determine whether roadway geometry was a factor in the crash
WILD_ANIMAL_RELATED	Wild Animal Related: Y/N to determine whether a wild animal was involved in the crash
DOMESTIC_ANIMAL_RELATED	Domestic Animal Related: Y/N to determine whether a domestic animal was involved in the crash
ROADWAY_DEPARTURE	Roadway Departure: Y/N to determine whether a vehicle departed the roadway as a result of the crash
OVERTURN_ROLLOVER	Overturn/Rollover: Y/N to determine whether a vehicle overturned and/or rolled over as a result of a crash
COMMERCIAL_MOTOR_VEH_INVOLVED	Commercial Motor Vehicle Involved: Y/N to determine whether a commercial motor vehicle was involved in the crash
INTERSTATE_HIGHWAY	Interstate Highway: Y/N to determine whether the crash occurred on an interstate roadway
TEENAGE_DRIVER_INVOLVED	Teenage Drive Involved: Y/N to determine whether a teenage driver was involved in the crash

Table B.7 Continued

Heading	Description
OLDER_DRIVER_INVOLVED	Older Driver Involved: Y/N to determine whether an older driver was involved in the crash
URBAN_COUNTY	Urban County: Y/N to determine whether the crash occurred in an urban area
ROUTE_TYPE	Route Type (L/S/U):
NIGHT_DARK_CONDITION	Night/Dark Condition: Y/N to determine whether night or dark conditions was a factor in the crash
SINGLE_VEHICLE	Single Vehicle: Y/N to determine whether a single vehicle was involved in a crash (i.e. not a collision involving multiple vehicles)
TRAIN_INVOLVED	Train Involved: Y/N to determine whether a train was involved in the crash
RAILROAD_CROSSING	Railroad Crossing: Y/N to determine whether the crash occurred at a railroad crossing
TRANSIT_VEHICLE_INVOLVED	Transit Vehicle Involved: Y/N to determine whether a transit vehicle was involved in the crash
COLLISION_WITH_FIXED_OBJECT	Collision with Fixed Object: Y/N to determine whether the crash involved a fixed object (i.e. not another vehicle, nor a person)

Table B.8: Critical Columns for Crash Vehicle

Heading	Description
CRASH_ID	Crash ID: Specific crash ID number for each crash
VEHICLE_NUM	Vehicle Number: Number assigned to each vehicle involved in a given crash
CRASH_DATETIME	Crash Date/Time: Date and time of crash
TRAVEL_DIRECTION_ID	Travel Direction: Direction value of route at the location of the crash (i.e., 1-5)
EVENT_SEQUENCE_1_ID	Event Sequence #1: ID for first crash sequence for non-collision and collision events (i.e., 0-99)
EVENT_SEQUENCE_2_ID	Event Sequence #2: ID for second crash sequence for non-collision and collision events (i.e., 0-99)
EVENT_SEQUENCE_3_ID	Event Sequence #3: ID for third crash sequence for non-collision and collision events (i.e., 0-99)
EVENT_SEQUENCE_4_ID	Event Sequence #4: ID for fourth crash sequence for non-collision and collision events (i.e., 0-99)
MOST_HARMFUL_EVENT_ID	Most Harmful Event: ID for most harmful event resulting from the crash (i.e., 0-99)
VEHICLE_MANEUVER_ID	Vehicle Maneuver: ID for the controlled maneuver prior to the crash (i.e., 1-14, 88-99)
VEHICLE_DETAIL_ID	Vehicle Detail ID: 8-digit ID number that is specific to a vehicle involved in a crash amongst all other vehicle involved in crashes