Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2014-12-01

# Crash Prediction Modeling for Curved Segments of Rural Two-Lane Two-Way Highways in Utah

Casey Scott Knecht
*Brigham Young University - Provo*

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Civil and Environmental Engineering Commons

BYU ScholarsArchive Citation

Knecht, Casey Scott, "Crash Prediction Modeling for Curved Segments of Rural Two-Lane Two-Way Highways in Utah" (2014). *All Theses and Dissertations*. 4352.
https://scholarsarchive.byu.edu/etd/4352

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

# Crash Prediction Modeling for Curved Segments of Rural

# Two-Lane Two-Way Highways in Utah

Casey Scott Knecht

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
C. Shane Reese

Department of Civil and Environmental Engineering

Brigham Young University

December 2014

Copyright © 2014 Casey Scott Knecht

All Rights Reserved

# ABSTRACT

## Crash Prediction Modeling for Curved Segments of Rural Two-Lane Two-Way Highways in Utah

Casey Scott Knecht
Department of Civil and Environmental Engineering, BYU
Master of Science

This thesis contains the results of the development of crash prediction models for curved segments of rural two-lane two-way highways in the state of Utah. The modeling effort included the calibration of the predictive model found in the Highway Safety Manual (HSM) as well as the development of Utah-specific models developed using negative binomial regression. The data for these models came from randomly sampled curved segments in Utah, with crash data coming from years 2008-2012. The total number of randomly sampled curved segments was 1,495.

The HSM predictive model for rural two-lane two-way highways consists of a safety performance function (SPF), crash modification factors (CMFs), and a jurisdiction-specific calibration factor. For this research, two sample periods were used: a three-year period from 2010 to 2012 and a five-year period from 2008 to 2012. The calibration factor for the HSM predictive model was determined to be 1.50 for the three-year period and 1.60 for the five-year period. These factors are to be used in conjunction with the HSM SPF and all applicable CMFs.

A negative binomial model was used to develop Utah-specific crash prediction models based on both the three-year and five-year sample periods. A backward stepwise regression technique was used to isolate the variables that would significantly affect highway safety. The independent variables used for negative binomial regression included the same set of variables used in the HSM predictive model along with other variables such as speed limit and truck traffic that were considered to have a significant effect on potential crash occurrence. The significant variables at the 95 percent confidence level were found to be average annual daily traffic, segment length, total truck percentage, and curve radius. The main benefit of the Utah-specific crash prediction models is that they provide a reasonable level of accuracy for crash prediction yet only require four variables, thus requiring much less effort in data collection compared to using the HSM predictive model.

Keywords: Highway Safety Manual, safety performance functions, crash modification factors, negative binomial, empirical Bayes, safety, horizontal curvature

# ACKNOWLEDGEMENTS

This research would not have been possible without the efforts of many people. I am grateful to the Utah Department of Transportation (UDOT) which provided the funding for this research. Also, I could not have completed this research were it not for the assistance and expertise of the technical advisory committee for this study, especially W. Scott Jones, Tim Taylor, and Travis Jensen of UDOT Traffic and Safety, and Charles Allen of InterPlan Co.

I am also grateful to the faculty and students on the research team at Brigham Young University. Specifically, I would like to thank Dr. Mitsuru Saito and Dr. Grant G. Schultz for their guidance and assistance with this research, in addition to their mentoring throughout my education. Dr. C. Shane Reese provided much assistance with the statistical analysis. Also, I would like to thank Aaron Cook, who provided invaluable assistance with data collection and preparation.

I would also like to acknowledge the continued support and encouragement from my family. Above all, I would like to thank my wife, Kylea, for her patience, support, and encouragement. She has provided more assistance than I could ever describe.

TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# 1 INTRODUCTION

Highway safety is a top priority for everyone. It is important for the governmental agencies that plan, construct, and maintain the highways. It is also important to everyone who uses them. Since the economy of the nation is significantly dependent on an efficient transportation system, it could be argued that highway safety affects everyone, as a crash on a highway does not only affect the people involved in the crash, but also others who are affected by the ensuing delays. Understanding and being able to identify the reasons behind crashes and resolving potential causes are paramount. To do so, researchers have developed crash prediction models that are based on historical crash data to estimate the number of future crashes under prevailing conditions that can be used to evaluate the contributions of physical attributes to crash occurrence.

One of the procedures for crash prediction modeling is using a safety performance function (SPF). SPFs are regression models that estimate average crash frequency for a specific site type as a function of annual average daily traffic (AADT) and segment length (AASHTO 2010, Lord and Persaud 2004). SPFs can be used for predicting the level of safety of a roadway by estimating the number of crashes that might occur given prevailing roadway conditions. The Highway Safety Manual (HSM), which is published by the American Association of State Highway and Transportation Officials (AASHTO), contains an 18-step method for predicting average crash frequencies on rural two-way two-lane highways (AASHTO 2010). The full process is referred to as the Predictive Method. Within the Predictive Method are predictive models that use SPFs along with other factors to predict the number of crashes on a given roadway segment. The SPFs in the HSM were created based on data from Minnesota, Washington, Michigan, Texas, and California. The result is not necessarily a nationwide average crash prediction model; rather, it is an average crash prediction model based on the five states from which the data were collected. Thus, the predictive model requires a calibration factor that adjusts the SPF for local conditions.

Previous research (Saito et al. 2011) developed calibration factors specific to the state of Utah, for the Utah Department of Transportation (UDOT). The calibration factors that were developed were specific to tangent segments of two-lane two-way rural highways in Utah because at the time of their research, no data were available for horizontal curvature. Since that research, UDOT has performed an inventory of all highway curvature within the state of Utah as part of its Light Detection and Ranging (LiDAR) asset management program. With this additional data, UDOT is desirous to calibrate the HSM predictive model specifically for curved segments of two-lane two-way rural highways in Utah. Similarly, UDOT has requested the creation of Utah-specific crash prediction models for two-lane two-way highways exclusive of the HSM predictive model.

This chapter presents the purpose and need for this research as well as the organization of this report.

## 1.1 Purpose and Need

The purpose of this research is to develop crash prediction models for curved segments of rural two-lane two-way highways in Utah using historical crash data and facility data recently collected as part of UDOT's LiDAR asset management program. This will be accomplished by calibrating the HSM crash prediction model for rural two-lane two-way highways as well as by creating Utah-specific models. The crash data come from years 2008-2012, and were assigned to two data groups: a three-year dataset from years 2010-2012, and also the full five-year dataset. These models allow UDOT to better understand the way highway curvature affects crash occurrences. The models will identify which factors play the largest role in crash prediction. With this information, UDOT can focus its efforts on the improvements that will make the most difference in safety.

The need for this research comes from the risk that is present every time someone drives an automobile. Operating an automobile is inherently dangerous and is something to which most people do not give a second thought. For the government agencies that are charged with designing, building, and maintaining highways, safety is a top priority. The most important reason to put safety first is the value of human life. Fatalities from crashes on U.S. highways are far too common. In 2012, there were 33,561 crash-related fatalities in the U.S., 200 of which were in Utah (NHTSA 2013). That is one death nearly every 15 minutes in the U.S. because of a crash. In Utah, crashes on rural roads are 3.3 times more likely to result in a death than crashes on urban roads (UDOT 2013). If an improvement can be made that saves even one life, it is worth it. The main obstacle to improvements that many government agencies face is the lack of funding. Many projects and improvements are shelved due to insufficient funds. With crash prediction modeling, agencies can focus on the most cost-effective measures to improve highway safety.

## 1.2 Report Organization

This chapter presented an overview of the report along with a stated purpose and need for this research. Chapter 2 presents a literature review of topics related to this research. Chapter 3 discusses the data preparation necessary for accurate and complete modeling, and Chapter 4 addresses the methodology for crash prediction modeling.  Chapter 5 presents and evaluates the results of the modeling effort, followed by Chapter 6 which contains conclusions and recommended use of models and further research needs.

# 2 LITERATURE REVIEW

This literature review contains topics related to highway geometry and safety as well as the acquisition and analysis of data, including LiDAR, highway curvature, SPFs, CMFs, calibration factors, and statistical methods.

## 2.1 Light Detection and Ranging (LiDAR)

LiDAR data are well-suited for transportation applications. LiDAR is especially useful when combined with geographic information system (GIS) technology to determine accurate 3D surface representations and characteristics (Pradhan and Rasdorf 2009).

Using LiDAR technology to inventory highway facilities is a practice that many government agencies and private companies are incorporating as one of their tools for asset management (Ellsworth 2013). Manual surveying and observation used to be the only methods available until aerial photography progressed to a point such that horizontal curves and lane widths could be measured with relative ease directly from the imagery. Digital elevation models (DEM) created from aerial photographs and satellite imagery have become widely available and they are generally accurate to +/- 7 meters (Rasdorf et al. 2004).

LiDAR is capable of providing information at high spatial resolutions and accuracies. Pradhan and Rasdorf (2009) discussed the accuracy of LiDAR data, and in 1999, LiDAR data were found to be accurate to +/- 15 centimeters. Figure 2-1 shows a sample LiDAR capture which exemplifies the accuracy level of LiDAR compared to the image captured by Roadview Explorer.

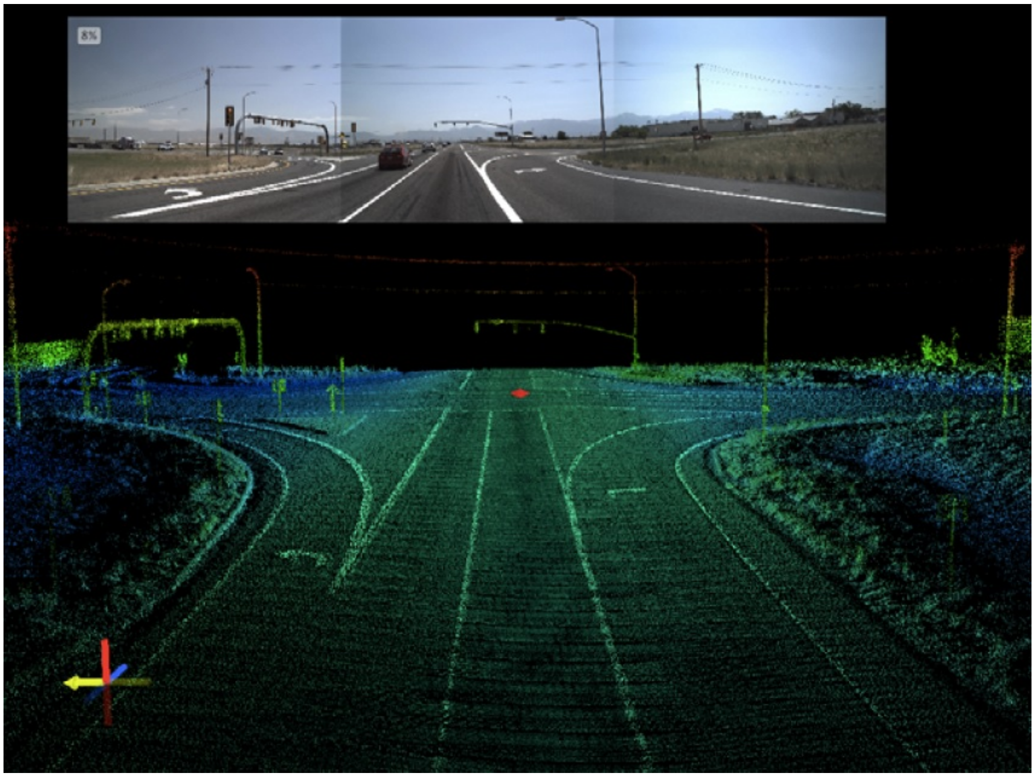


Figure 2-1: LiDAR Capture (Ellsworth 2013)

Many transportation agencies are utilizing mobile vehicles to collect a wide variety of asset data (Findley et al. 2013). In 2011, UDOT commenced a project that would eventually collect highway infrastructure data for every state road in Utah using LiDAR. The data have an average accuracy of +/-3 centimeters (Ellsworth 2013).

As the technology improves and the machinery becomes more sophisticated, accuracy naturally improves with it. LiDAR employs a significantly higher concentration of data points than surveying or DEM (Findley et al. 2012). Thus less interpolation is required and the points create a redundancy to reduce error.

LiDAR data are an excellent supplement to existing highway data inventories as they provide a validation of existing data (Findley et al. 2012). And as previously explained the accuracy of asset data can greatly improve as its technology advances.

## 2.2 Highway Curvature

Highway curvature will play an important role in this research. A previous study by Saito et al. (2011) focused on straight segments because curvature data were not available. However, because of the availability of curvature data from UDOT's LiDAR project, this research was able to study the effect of horizontal and vertical alignment on SPFs.

Curves can be found on almost every highway in the U.S. They require careful design and implementation to maintain a high level of safety. Yet even with the extra precautions taken, approximately 25 percent of all fatal crashes in the United States in 2002 occurred on horizontal curves (Khan et al. 2012). This does not include crashes that occurred on vertical curve segments. Previous research has identified curvature as one of the most significant predictors of crashes (Easa and You 2009, Lord et al. 2010).

Curves can be very different in appearance and design. There are horizontal curves, vertical curves, and curves that are both horizontal and vertical. For horizontal curves, the most apparent distinguishing factor is how sharp the curve is. Sharpness is really a measure of radius or curvature. As expected, there are more crashes on sharper curves, yet that is not the only factor. Narrower curve width, lack of spiral transitions, and increased superelevation deficiency all contribute to higher crash rates on curves (Zegeer et al. 1992). Approximately 70 percent of curve-related fatal crashes were single-vehicle crashes in which the vehicle left the roadway and struck a fixed object or overturned (Srinivasan et al. 2009). Curves are inherently more dangerous than straight sections because drivers are required to maneuver rather than simply maintain their course.

In some studies, road segments must have a minimum degree of curvature to be classified as a curve. Khan et al. (2012) determined that 3.45° was a good break point for what segments behaved like straight sections versus curved sections. Segments with less than 3.45° curvature behaved similarly to straight sections.

Determining segmentation of highway curves can prove difficult as well. Srinivasan et al. (2009) used global positioning system (GPS) coordinates to track horizontal alignments. The data were then used to determine where tangents, arcs, and spirals began and ended.

All else being equal, higher traffic volumes and longer curves were also associated with significantly higher number of curve-related crashes. Ranges of crash reductions for horizontal curves improvements were determined for flattening curves, widening lanes, widening paved shoulders, adding unpaved shoulders, adding a spiral transition, and improving superelevation (Zegeer et al. 1992).

There are several factors that can affect safety on curves including signage, pavement markings, and roadside hazards (Labi 2006, Zegeer et al. 1992, Khan et al. 2012). In a study on rural two-lane roads in Indiana, Labi (2006) found that many of the roads observed had deficiencies in signage and markings. Also cited in the study was a sobering statistic: the death rate for motorists on rural roads was more than 2.5 times the rate for driving on all other roads. Any remediation or added measure of safety that reduces that number of crashes would be a move in the right direction.

When looking at all factors that can affect safety, some have been shown to cause little if any change.  For highway segments with a degree of curvature greater than 3.45, the use of advisory signs is not a significant factor (Khan et al. 2012).  So on sharper curves, other influencing factors must take over.  Hauer (1999) discarded lane width as a factor, specifically 11-foot versus 12-foot lanes.  It would seem that a larger lane would be safer, but in the research presented by Hauer, there were more crashes in the 12-foot lanes than the 11-foot lanes.

All of this information shows that modeling crashes involves various factors.  The most important of these factors is the human factor which is very difficult to quantify and describe.  The best thing that can be done to remedy this is to collect more data and create jurisdiction-specific calibrations of crash prediction models.

## 2.3 Safety Performance Functions (SPFs)

SPFs are regression models that estimate average crash frequency for a specific site type as a function of AADT and segment length (AASHTO 2010, Lord and Persaud 2004).  SPFs developed in a specific jurisdiction or on a general level can be recalibrated for a different jurisdiction.  The HSM contains SPFs and there are documented calibrations that have already been performed (Fitzpatrick et al. 2008).

The HSM contains an SPF for rural two-lane two-way road segments as shown in Equation 2-1.

$$N_{spf} = AADT \times L \times 365 \times 10^{-6} \times e^{-0.312} \qquad (2\text{-}1)$$

where, $N_{spf}$ = number of predicted annual crashes,

$AADT$ = average annual daily traffic, and

$L$ = segment length (mi).

This model assumes that the annual number of predicted crashes is directly proportional to the amount of vehicles that travel through the highway segment. By converting AADT to an annual value (multiplying by 365) and changing the order of magnitude (multiplying by $10^{-6}$), the annual number of predicted crashes becomes million vehicles miles traveled (VMT), which is used as a surrogate of exposure.

The HSM was released in 2010 after much research and preparation. Consequently, very few studies have been published on the HSM crash prediction models for rural two-lane two-way highways since its publication. In 2011, a study was performed on calibrating the HSM to predict total crashes on highways in Oregon (Xie et al. 2011). In this particular study, the guidelines for calibration set forth in the HSM were followed. The study mentioned specifically the difficulty in preparing the data set and the local adjustments made such as adjusting sample sizes for underrepresented facility types. The target number of 100 crashes per year could not be achieved at low-volume intersections. The study shows that sample size estimation procedures were applied to determine how many crashes could reasonably be expected, and then the target number of crashes was modified. Also, some minor road AADT values were difficult to come by. In certain regions, the locally maintained roads carried more traffic than the state highways. Because of this, a model was created to estimate AADT on rural highways. The model used variables including population, income, and distance from freeway, along with geometric design information. These variables allowed the AADT to be estimated in a manner consistent with the rest of the state highway system.

Any number of variables can be used in a model. The key is choosing the variables that are most appropriate and affect the SPF the most. Data collection costs increase as the number of variables increases. The HSM model (AASHTO 2010) uses several variables for rural two-lane two-way highways including lane and shoulder widths, curvature, driveway density, and roadside hazards. There are certain base conditions that the HSM lays out such as 12-foot lanes, six-foot paved shoulders, five driveways per mile, a roadside hazard rating of three, and an absence of curvature, rumble strips, passing lanes, two-way left-turn lanes (TWLTL), lighting, and automated speed enforcement. Deviations from this can still be modeled with crash modification factors (CMFs) which are multiplied to the number of predicted annual crashes found by the base crash prediction model. The HSM specifies that SPFs should incorporate traffic volume and crash frequency, while geometric design and traffic control features should be incorporated through CMFs.

SPFs tend to be simplistic because they often contain predictive rather than actual causal factors (Lord and Persaud 2004). As described above, causal factors including human errors are very difficult to model. Hence there is a need to adjust SPFs by way of CMFs, which will be discussed next.

## 2.4 Crash Modification Factors (CMFs)

CMFs represent the relative change in crash frequency due to a change in one specific condition, estimating the effect of a particular geometric design or traffic control or the effectiveness of a particular treatment or condition (AASHTO 2010). CMFs were originally referred to as Accident Modification Factors (AMFs), but were updated in the final version of the HSM to be CMFs. As such, CMF will be used exclusively in this thesis. CMFs are often preferred by transportation safety analysts because they allow base-line models to be recalibrated for different jurisdictions.

For rural two-lane two-way highway segments, the HSM model has CMFs for 12 design and control features: lane width, shoulder width and type, horizontal curve length and radius, horizontal curve superelevation, grade, driveway density, centerline rumble strips, passing lanes, two-way left-turn lanes, roadside design, lighting, and automatic speed enforcement (AASHTO 2010).

CMFs are developed with two variables: location and time. By keeping time constant, a cross-sectional analysis can be performed. Keeping location constant will render a before-after analysis (Gross et al. 2010). For example, a study on similar roads in California and Texas at the same time can be classified as a cross-sectional analysis. If the road in California is observed this year and compared to same road last year, it can be classified as a before-after analysis. Research from Gross et al. (2010) states that before-after analyses are preferred to cross-sectional analyses because an actual change can be observed.

In many instances, multiple CMFs can be used (Gross et al. 2010). Care must be taken, however, if two or more CMFs are used simultaneously because their effect may be compounded if there is any correlation among them (AASHTO 2010, Fitzpatrick et al. 2008, Gross et al. 2010, Lord et al. 2010). As mentioned above, variables need to be independent of each other if they are to be used together.

When changes are made to highway geometry and/or segmentation, the calibration of CMFs will need to be performed. Hauer (1997) claimed that driver behavior can be affected any time there is a change. For example, a road that is repaved may provide an increased sense of safety even if the actual highway geometry is identical. This increased sense of safety is in addition to the actual increase in safety that comes from replacing pavement that is in poor condition (Labi 2006). Because of the increased sense of safety, drivers may increase speed, thus altering the condition that had existed in previous data. Before-after analyses become cloudy when multiple variables change simultaneously, especially when driver behavior is involved. While a longer study period may help to account for natural variability and regression to the mean, a longer study period increases the likelihood that site conditions have changed. The HSM (AASHTO 2010) recommends estimating expected crash frequency for each year in a study period as a way to address this limitation.

## 2.5 Calibration Factors

As part of the predictive model developed in the HSM, a calibration factor is multiplied with the crash frequency predicted by the SPF to account for differences between the jurisdiction and time period for which the predictive models were developed and the jurisdiction and time period to which they are applied (AASHTO 2010). These calibration factors can adjust for climate, animal population, driver population, crash reporting threshold, and crash reporting practices. The HSM recommends new calibration factors every two to three years. The calibration procedure includes identifying facility types, selecting sites, obtaining data, applying the predictive model to predict total crash frequency at each site, and computing calibration factors. The computation is simply a ratio of the sum of the observed crashes at all sites to the sum of the predicted crashes at all sites. The calibration factor will vary for each facility type.

Calibration for rural two-lane two-way highways requires several data elements such as segment length, AADT, horizontal curve length and radius, lane width, shoulder type and width, and the presence of two-way left-turn lanes (AASHTO 2010). Other data such as spiral transition presence, superelevation, percent grade, lighting presence, driveway density, passing lane presence, short four-lane presence, centerline rumble strip presence, and roadside hazard rating are desirable, but not required. For these optional data, assumptions can be made if the actual data are not available (AASHTO 2010). The assumptions are laid out in the HSM, with most defaulting to the agency design policy.

## 2.6 Statistical Methods

Choosing a statistical method for analysis will depend on the jurisdiction and what variables and factors are important to include. Poisson regression is one of the most suitable techniques for crash prediction modeling because highway crashes are discrete rare events and crash counts are non-negative integer variables (Labi 2006). Labi goes on to explain that the Poisson approach has a crucial weakness, that is, the assumption that the mean and the variance of crash distribution are equal (Labi 2006). This is rarely the case with crash analysis.

The negative binomial (NB) model allows for additional variance representing the effect of omitted variables. Fitzpatrick et al. (2010) used NB regression models to determine the effects of independent variables on crashes on rural four-lane highways in Texas. Srinivasan et al. (2009) used an empirical Bayes (EB) before-after analysis to account for potential selection bias and regression-to-the-mean. The HSM includes the EB model, and has thus established it as the standard method for road safety analysis (AASHTO 2010, Labi 2006). Another benefit of the EB model for safety analysis is that it automatically corrects for the regression-to-the-mean effect (Labi 2006).

## 2.7 Literature Review Summary

Using LiDAR data to account for highway geometry and conditions is a new approach that will be of benefit to any transportation agency. Understanding how curved and straight sections affect crash data will allow for better planning and implementation of new roads and changes to existing roads.  Since curves tend to be more dangerous than straight segments, proper analysis needs to be performed so as to show differences between curved segments and straight segments.

The predictive model laid out in the HSM includes an SPF to predict crash frequency with base conditions, one or more CMFs to account for site-specific conditions, and a calibration factor to adjust the prediction to local conditions at the site.

Curved segments provide an entirely new variable when creating SPFs and CMFs. Proper segmentation of curves and tangents will ensure that calibrations are accurate and reliable.  Understanding the characteristics of variables and the relationship they have with each other will prevent redundancy and overestimation of a CMF.  The proper statistical method and approach may be difficult to choose, but with the right variables and analysis period, SPFs and CMFs obtained through the calibration process can be useful for crash prediction and analysis.

# 3 DATA PREPARATION

The first step of data collection was to randomly select segments that were representatives of rural two-way two-lane highways in Utah. Once the segments for further analysis were selected, the next step was to gather sufficient data on the components of the selected segments that would affect the predictive power of the crash prediction models (i.e., variables in the predictive models included in the HSM). While the segments do not need to meet the base conditions, one must know the values of the components so as to determine the value of an appropriate CMF.

This chapter presents the scope of data collection the Horizontal Alignment Finder (HAF), the resources used for data collection, the limitations of the data and the data collection resources, proper sampling, facility data, and crash data. A summary will conclude the chapter.

## 3.1 Scope of Data Collection

The task for this study included an analysis of tangent segments and curved segments. The tangent segments were selected from the previous research of Saito et al. (2011) in which rural two-way two-lane highways in Utah were randomly selected. The selection was limited to homogeneous tangent sections due to data limitations, especially curve-related data, as well as the scope of the research. Therefore, the main data collection effort of this study was focused on finding curve related data such as curve radius, point of curvature (PC) and point of tangent (PT).

## 3.2 Horizontal Alignment Finder

The HSM crash prediction models consist of a set of variables describing the conditions of the segments selected, of which curve radii are one of the most important information that needs to be available to analyze rural two-lane two-way highways. This piece of information, however, has been difficult to collect and hence has not been available previously. For this reason, the previous research done by Saito et al. (2011) focused on tangent segments of the rural two-way two-lane highways. Now that data for curved segments became available through UDOT's LiDAR program it became possible to analyze curve segments in addition to tangent segments. Curve segments have more variation in their attributes and are more prone to variation in their classifications and accuracy (Findley 2011). Despite the highway asset data available from the LiDAR project, the highway geometric dataset provided by the LiDAR program was inadequate for this research, mostly because it did not clearly and accurately identify the PC, PT, or other attributes of the curve. The major issue was that the dataset obtained from the LiDAR project segmented more than half of all curves in the UDOT-owned highways into more than one segment. In other words, one curve was classified as having multiple PCs and PTs, creating the appearance of multiple curves of varying length and type. This segmentation would not accurately reflect the reality of the curves, and would therefore produce inconsistent results in the analysis of crash data. A method was needed to combine the curve segments within the same curve identified by the LiDAR program as curves into one single section. This method proved to be a crucial ancillary effort on this project.

The algorithm developed in this study is called the HAF and it provided a method by which the curve segments were combined to a reasonably high success rate (85 percent or better). The algorithm is described in more detail in a separate paper (Cook et al. 2015). The dataset provided by the LiDAR program included tabulated data for each segment, including the milepost. The algorithm uses this tabulated data to compare attributes of each consecutive segment and then combines the segments that have sufficiently similar geometry. It performs this combination task by identifying each segment as either a tangent, part of a unique curve, or a unique curve all on its own. It then combines all partial segments to make unique curves from each of these sets of partial segments. After this manipulation a combined curve segment extends from the beginning of the first combined segment to the end of the last combined segment. It also has a filter to catch and remove erroneous curves which often appear in areas near intersections. These three steps—Identification, Combination, and Removal—are the basic idea behind the procedure. Figure 3-1 presents the segmentation process used by this algorithm.

The Identification and Combination steps are conceptually separated, but the implementation of the two are closely connected by the mechanism by which the HAF algorithm associates them. The identification step works by classifying each segment as either a tangent or a curve. Each curved segment is assigned a curve number, which is unique for each curve, but not necessarily for each segment. In other words, a curve with three constituent segments would have each of those segments assigned the same curve number. This is done by comparing each segment to the segment immediately prior, that is, curved segments sufficiently similar are given the same curve number. In the combination step, all segments with the same curve number are grouped into one larger segment which represents the full curve, and have their attributes combined in various ways as discussed below. These two steps are presented visually in Figure 3-2.

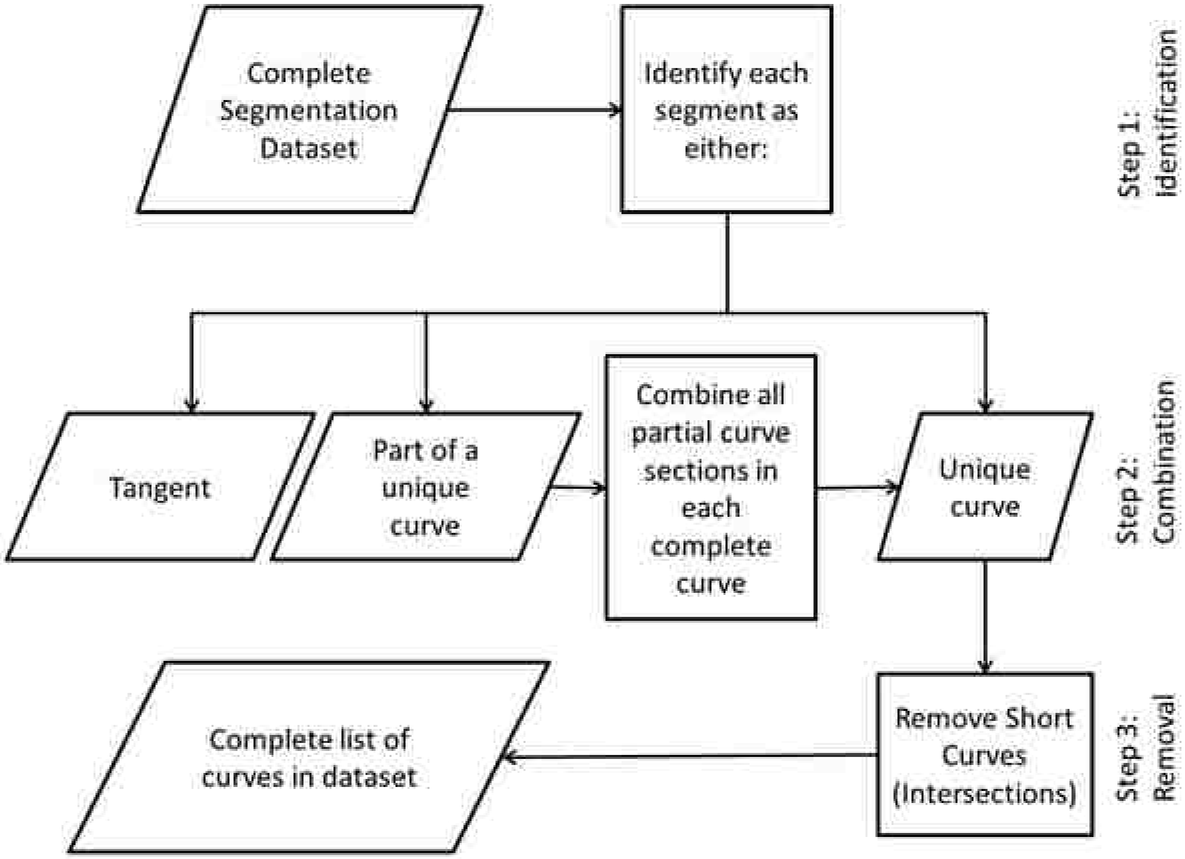

Complete Segmentation Dataset
Identify each segment as either:
Step 1: Identification
Tangent
Part of a unique curve
Combine all partial curve sections in each complete curve
Unique curve
Step 2: Combination
Complete list of curves in dataset
Remove Short Curves (Intersections)
Step 3: Removal

Figure 3-1: The HAF Algorithm Segmentation Steps

The attributes are combined according to the number of segments in each curve number. If there is only one constituent segment, the attributes of the segment become the attributes of the curve. If there is more than one segment, the attributes are combined from all the constituent segments. A detailed breakdown of how these are separated is shown in Figure 3-3.

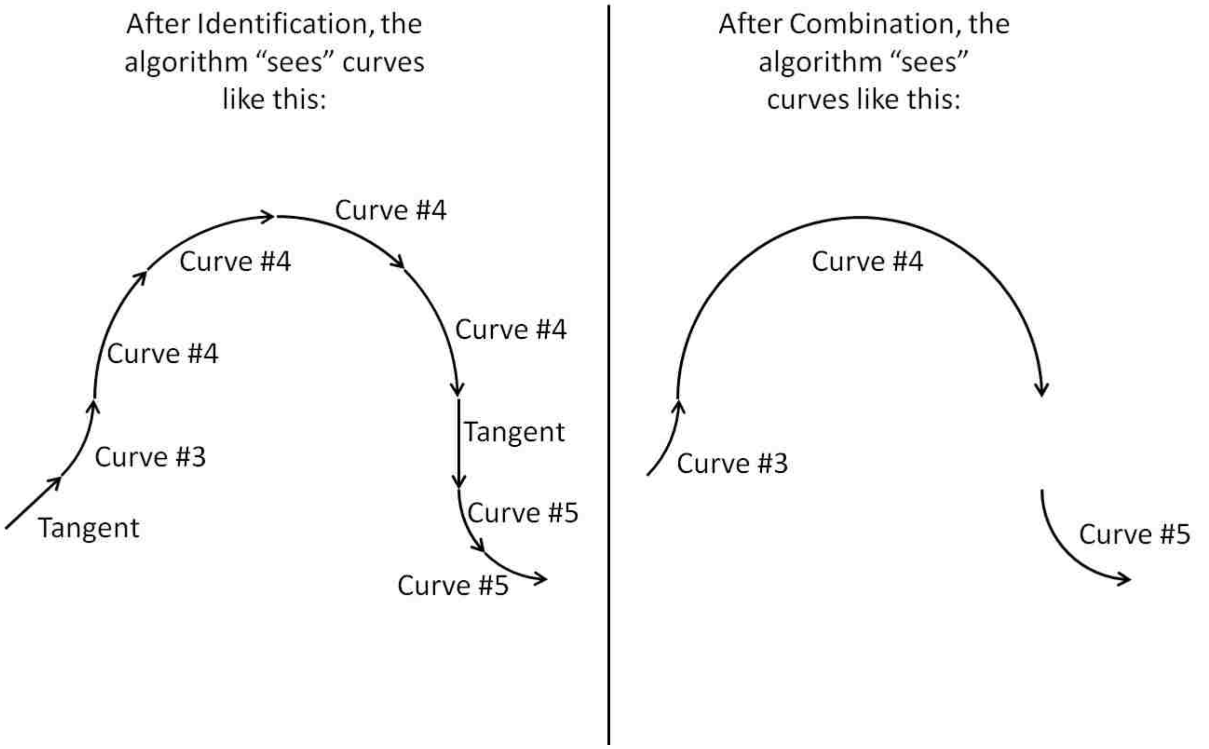

After Identification, the algorithm "sees" curves like this:
Curve #4
Curve #4
Curve #4
Curve #4
Curve #3
Tangent
Tangent
Curve #5
Curve #5
After Combination, the algorithm "sees" curves like this:
Curve #4
Curve #3
Curve #5

Figure 3-2: Representation of Identification and Combination of the HAF Algorithm

Four parameters were used to determine if each segment should be classified as a partial curve, a unique curve, or a tangent. The parameters include Segment Length, Segment Radius, In-Curve Radius/Length Ratio, and Curve Length. The first three are used in identification and combination, while the fourth is used in the Removal stage. The algorithm compares each segment's attributes to those of the previous segment and these parameters. These comparisons become the inputs for a weighting scale from -7 to +7, including 0. The weighting scale is how the algorithm classifies each section as a curve or tangent, and the range of the scale was arbitrarily selected for convenience. A positive weight will classify the segment as a curve, and a negative will classify it as a tangent. A zero weight is reserved for cases in which the road is in a curve, yet the next segment is obviously not part of the same curve for reasons other than the radius of curvature (e.g., the route changes, the direction of travel reverses, or the direction of curvature reverses). No segment is permanently assigned a zero weight; the HAF algorithm will only force the segment to be analyzed without the previous segment's attributes and force a new curve number if the segment is a curve.

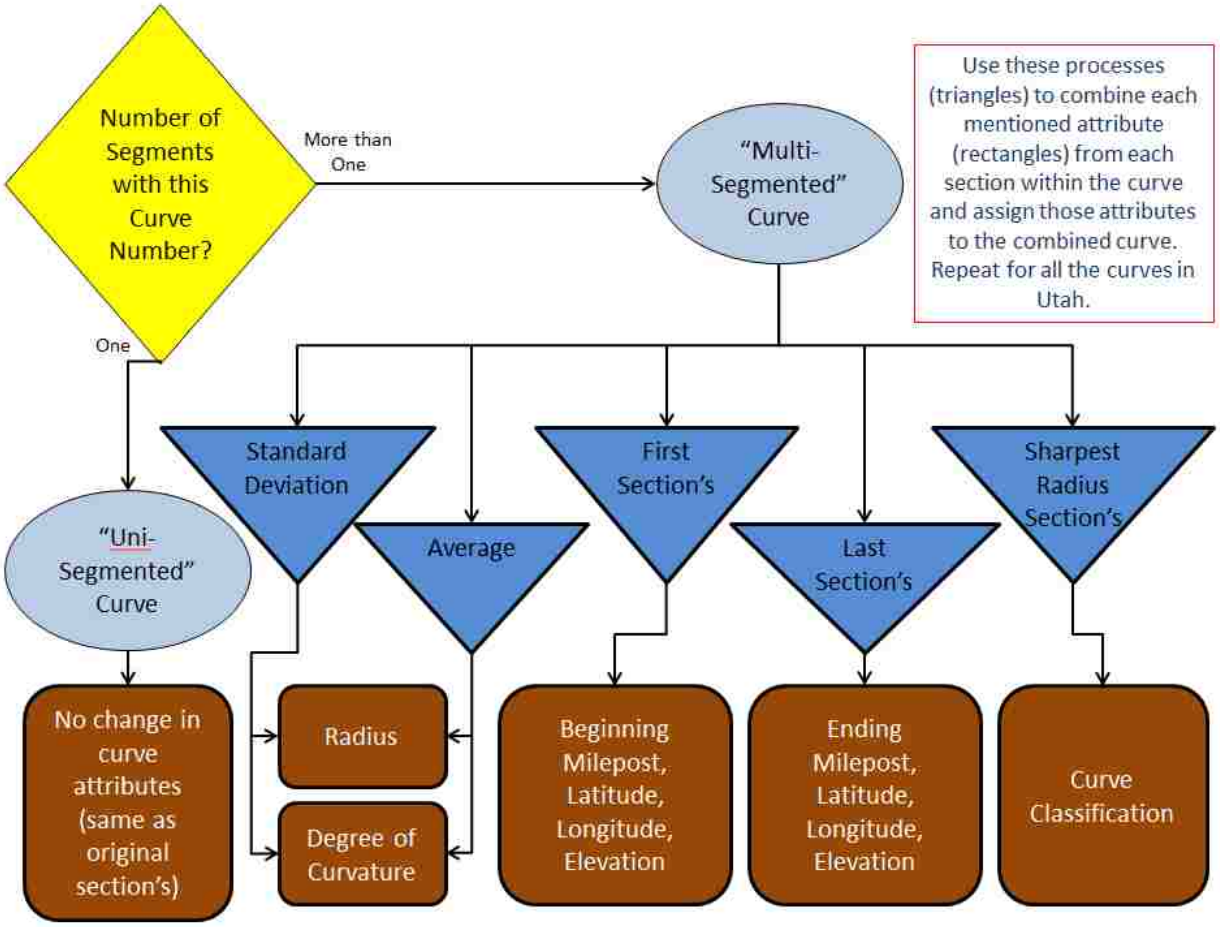

Number of Segments with this Curve Number?
More than One
One
"Multi-Segmented" Curve
Use these processes (triangles) to combine each mentioned attribute (rectangles) from each section within the curve and assign those attributes to the combined curve. Repeat for all the curves in Utah.
"Uni-Segmented" Curve
Standard Deviation
Average
First Section's
Last Section's
Sharpest Radius Section's
No change in curve attributes (same as original section's)
Radius
Degree of Curvature
Beginning Milepost, Latitude, Longitude, Elevation
Ending Milepost, Latitude, Longitude, Elevation
Curve Classification

Figure 3-3: Combination Schema for Each Segment

An analysis of the curves produced by the HAF resulted in a success rate of the locations of the curves being successfully identified of 84.4 – 92.9 percent and a success rate of correctly placing the PC and PT in the proper location of 78.7 – 89.9 percent. Although these are not perfect, they are as good as or better than any other model. It is believed that the randomization of the curve selection and the buffering of the curve to include superelevation runoff and tangent runout mitigated the margin of error in most of the problematic curves.

## 3.3 Resources

The collection of data came from various sources throughout the process. The availability and accessibility of data online allowed for a widespread survey of segments across the state. In many cases, the different resources had redundant features which allowed for verification and validation of different data. The resources used in the data collection process included Google Earth, Roadview Explorer, the UDOT Data Portal, and the UDOT Crash Database.

### 3.3.1 Google Earth

Google Earth (Google 2014) was the source of all aerial imagery used in the data collection. Aerial imagery was used to obtain lane width, lane configuration (including passing ability), and driveway count data. The lane widths were obtained by using the measure tool built into the software. Google Earth was also used to verify shoulder width and the presence of rumble strips and lighting where possible. In addition to the collection of specific attributes, Google Earth was used to gain a general understanding of the segment and the surrounding features. It was during this preliminary observation that many segments were removed from the dataset of randomly selected curved segments. The criteria for segment removal will be discussed later in this chapter.

### 3.3.2 Roadview Explorer

For the attributes that required street-level imagery, Roadview Explorer (UDOT 2012) was the principal resource. This included rumble strip presence and lighting. It was also used to confirm other attributes such as passing ability, driveway count, and lane and shoulder widths. Roadview Explorer is a Java platform application consisting of searchable and navigable imagery taken from a vehicle, mounted with cameras, while driving on every state and federal road in Utah. The program has the ability to jump to a specific milepost, down to the thousandth of a mile, on any road included in the system. It also has the ability to virtually travel the road by advancing the images in a slideshow format, giving the illusion of driving the road. This feature was especially helpful in navigating the road from start to finish for each of the segments selected for analysis. The navigation can move forward and backward for both directions of travel.

### 3.3.3 UDOT Data Portal

The UDOT Data Portal (UDOT 2014a) contains a large volume of data available for download in a variety of formats. The formats include shapefiles for use in GIS software, KML files for use in Google Earth, spreadsheets, and text files. The data for this project that were gathered from the UDOT Data Portal included the actual roadway file, speed limits, shoulder widths, AADT, and truck percentages. These datasets were brought into a GIS map as shapefiles, and from that point, the attributes could be tied to the segments.

### 3.3.4 UDOT Crash Database

The UDOT Crash Database (UDOT 2013) contains all recorded data pertaining to every crash on Utah state roads. Most of the data come from police reports and crash investigators.

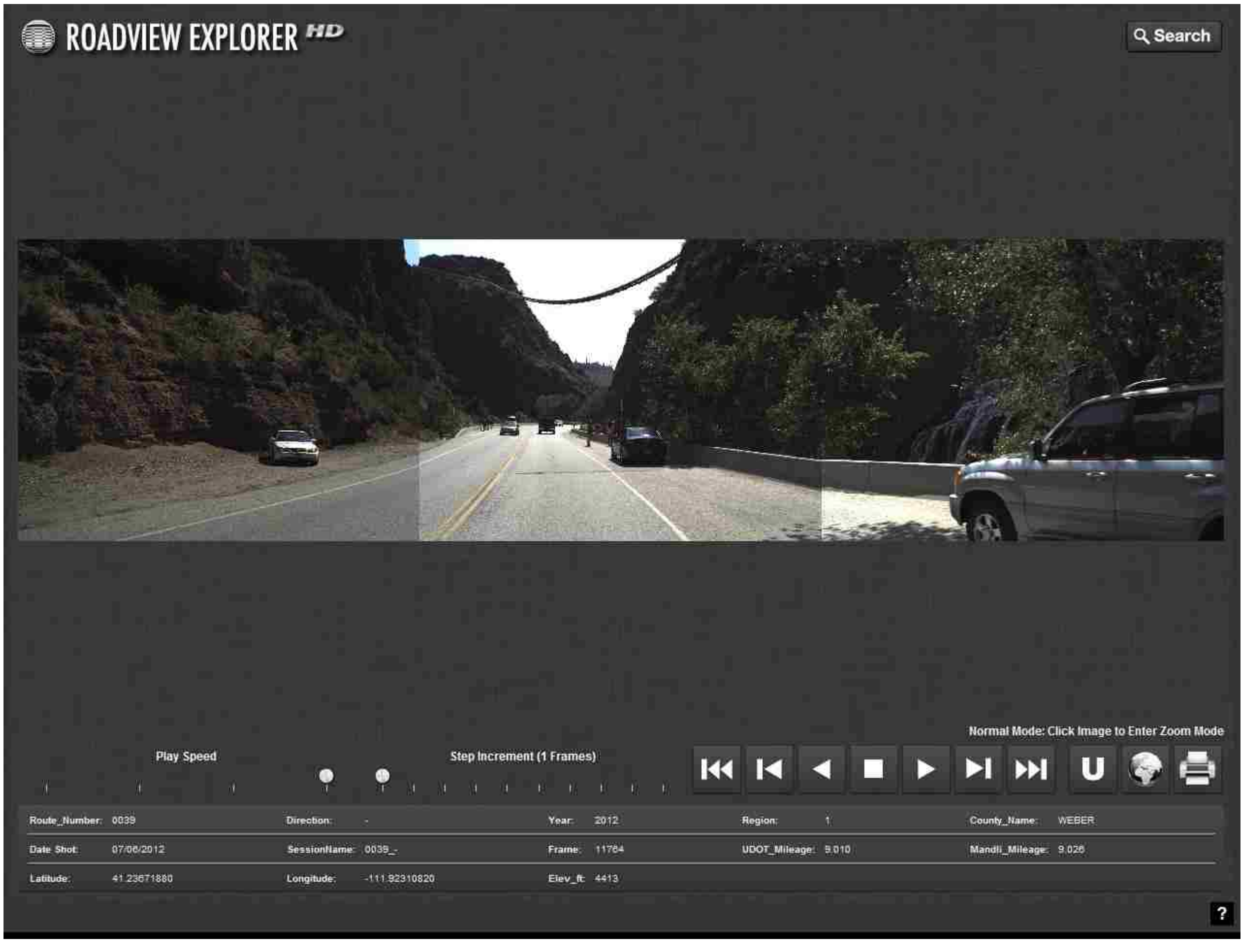
The data recorded for each crash include date and time, route and milepost, weather conditions, cause, vehicles involved, passengers involved, severity, light conditions, work zone conditions, and road surface conditions. The Crash Database is not accessible online. Special permission must be granted to access the specific files.

## 3.4 Data Limitations

After consideration of the crashes attributed to curve segments, it was observed that many crashes occurred just before the start of the curve or just past the end of the curve. This could indicate that the entrance or exit of a curve is dangerous in its own right. Frequently, however, the crash reporting by law enforcement contains various levels of precision and accuracy and may not correctly identify the location of the crash. Some site investigators use a portable measuring wheel to measure from the nearest milepost. Others use a reporting device that is equipped with GPS receiver to pinpoint the site of crash. But unless the reporting takes place at the actual site of the crash, the GPS coordinates will not be accurate. If the reporting takes place in a vehicle parked near the site, then the crash may be recorded at the parking location near the site. Because of these inconsistencies, it was decided to add superelevation runoff and tangent runout lengths to both ends of the curve regardless of the actual presence of these elements. The calculations for these lengths are found in the Facility Data subsection of this chapter.

## 3.5 Proper Sampling

Some segments were removed from the dataset at various points in the data compilation process. Duplicate or overlapping segments were removed immediately based on route number and mileposts. Segments that were in urban areas or residential areas were removed. Segments with speed limits lower than 30 mph were removed. The reasoning for this is because speed limits lower than 30 mph are generally associated with high pedestrian traffic, residential areas, and/or vehicles stopping for roadside attractions (such as the waterfall on the right side of Figure 3-4). While these conditions are not necessarily grounds for removal, they do not fit the purpose of identifying truly rural segments.


ROADVIEW EXPLORER HD
Search
Normal Mode: Click Image to Enter Zoom Mode
Play Speed
Step Increment (1 Frames)
U
?

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Route_Number: | 0039 | Direction: | - | Year: | 2012 | Region: | 1 | County_Name: | WEBER |
| Date Shot: | 07/06/2012 | SessionName: | 0039_- | Frame: | 11764 | UDOT_Mileage: | 9.010 | Mandli_Mileage: | 9.026 |
| Latitude: | 41.23671880 | Longitude: | -111.92310820 | Elev_ft: | 4413 | | | | |

Figure 3-4: Example of Segment with High Pedestrian Traffic (UDOT 2012)

If the lanes were not striped or if they were less than 9 feet wide, the segment was removed. Segments that contained a stop sign, signal, or other traffic control device for the main directions of traffic were removed since their inclusion would be better suited for an intersection analysis. In some areas, the segment included a 90° or near-90° turn from one cardinal direction to another with a very small curve radius. This was encountered where a route traveled on an east/west roadway and then the route designation changed to a north/south roadway. Thus, the change was usually at a four-way intersection where the two legs of the designated route received preference. These segments were removed regardless of the presence of traffic control devices. Several segments were within national and state parks and recreation areas. While that did not merit immediate removal, most of the segments within these park and recreational areas were near on-road services such as toll booths, information booths, boat launches, ranger stations, and recreation vehicle dump stations. These on-road services prevent free-flow operation, and therefore the segments that contained or were near any of these services were removed from the dataset.

## 3.6 Facility Data

The facility data that were collected include curve radius, degree, and class; curve buffer; grade; speed limit; rumble strip presence; lane and shoulder width; driveway density; passing ability; lighting; AADT; and truck percentage. Most of these factors are the variables used in the HSM predictive models while others, such as speed limit and truck percentage, were selected due to their perceived effect on highway safety.

### 3.6.1 Curve Radius, Degree, and Class

Each segment was analyzed to determine the curve radius, curve degree, and curve class. ArcGIS (ESRI 2012) was used to measure the radius (in feet) and the degree of curvature (in degrees) for each curve. The curve class was determined from the definition in the Highway

Performance Monitoring System (HPMS) (FHWA 2014). The classification breakdown is shown in Table 3-1.

Table 3-1: HPMS Curve Classification Breakdown (FHWA 2014)

| **Curve Classification** | **Range of Values** |
|---|---|
| A | Under 3.5 degrees (i.e., 0.061 radians) |
| B | 3.5 - 5.4 degrees (i.e., 0.061 - 0.094 radians) |
| C | 5.5 - 8.4 degrees (i.e., 0.096 - 0.147 radians) |
| D | 8.5 - 13.9 degrees (i.e., 0.148 - 0.243 radians) |
| E | 14.0 -27.9 degrees (i.e., 0.244 - 0.487 radians) |
| F | 28 degrees (i.e., 0.489 radians) or more |

### 3.6.2 Curve Buffer

As discussed previously, it is possible that not all curve-related crashes are recorded between the point of curvature and point of tangency. This could be due to inaccurate recording or simply because the crash occurred just before entering or just after exiting a curve. It was determined that the full length of superelevation runoff and tangent runout should be added to the length of the curve on both ends as a buffer. With this curve buffer, crashes in the immediate vicinity of the curve would be included.

Superelevation runoff is the length required to transition the cross slope of a road from zero percent on the outer lane, (the location of adverse crown) to a superelevated position on a curve. A cross slope of zero percent on the outer lane is a level surface while the inner lane still has the cross slope value. Superelevation runoff includes a portion within the curve and a portion outside the curve, with the PC or PT denoting the start and end of a curve, respectively. The proportion of runoff length in the tangent section varies from 0.6 to 0.8, with most agencies using 0.67 for all street and highway curves (AASHTO 2011). This research is based on the same assumption, and for convenience, a proportion of two-thirds (2/3) was used. Equation 3-1 shows how to calculate superelevation runoff (AASHTO 2011).

$$L_r = \frac{(wn_1)e_d}{\Delta} \tag{3-1}$$

where, $L_r$ = minimum length of superelevation runoff, ft.;

$w$ = width of one traffic lane, ft.;

$n_1$ = number of lanes rotated;

$e_d$ = design superelevation rate, percent; and

$\Delta$ = maximum relative gradient, percent.

Without knowing the design superelevation rate, a value of six percent was assumed. This is the maximum design superelevation rate in the state of Utah, as specified by UDOT (2008). The maximum relative gradient comes from AASHTO (2011) and is shown in Table 3-2. Tangent runout is the length required to transition the cross slope of a road from normal cross slope to zero percent. Equation 3-2 shows how to calculate tangent runout (AASHTO 2011). For this study, a normal cross slope rate of two percent was assumed. The design superelevation rate of six percent was already assumed for the superelevation runoff calculations, and those same calculations determine the final variable of the equation.

Table 3-2: Maximum Relative Gradient (AASHTO 2011)

| Design Speed (mph) | Maximum Relative Gradient (%) | Equivalent Maximum Relative Slope |
|---|---|---|
| 15 | 0.78 | 1:128 |
| 20 | 0.74 | 1:135 |
| 25 | 0.70 | 1:143 |
| 30 | 0.66 | 1:152 |
| 35 | 0.62 | 1:161 |
| 40 | 0.58 | 1:172 |
| 45 | 0.54 | 1:185 |
| 50 | 0.50 | 1:200 |
| 55 | 0.47 | 1:213 |
| 60 | 0.45 | 1:222 |
| 65 | 0.43 | 1:233 |
| 70 | 0.40 | 1:250 |
| 75 | 0.38 | 1:263 |
| 80 | 0.35 | 1:286 |

$$L_t = \frac{e_{NC}}{e_d} L_r \qquad (3\text{-}2)$$

where, $L_t$ = minimum length of tangent runout, ft.;

$e_{NC}$ = normal cross slope rate, percent;

$e_d$ = design superelevation rate, percent; and

$L_r$ = minimum length of superelevation runoff, ft.

### 3.6.3 Grade

All segments were evaluated for grade. The most reliable data came from ArcGIS analysis which included latitude, longitude, and elevation data for each start and end point. The difference in elevation was divided by the length of the segment, as measured in ArcGIS. The absolute value of the quotient became the decimal value for the grade. The grade was converted to a percentage to be consistent with standard reporting.

### 3.6.4 Speed Limit

Speed limit data were obtained from UDOT (2014b) and verified whenever possible via Roadview Explorer (UDOT 2012).

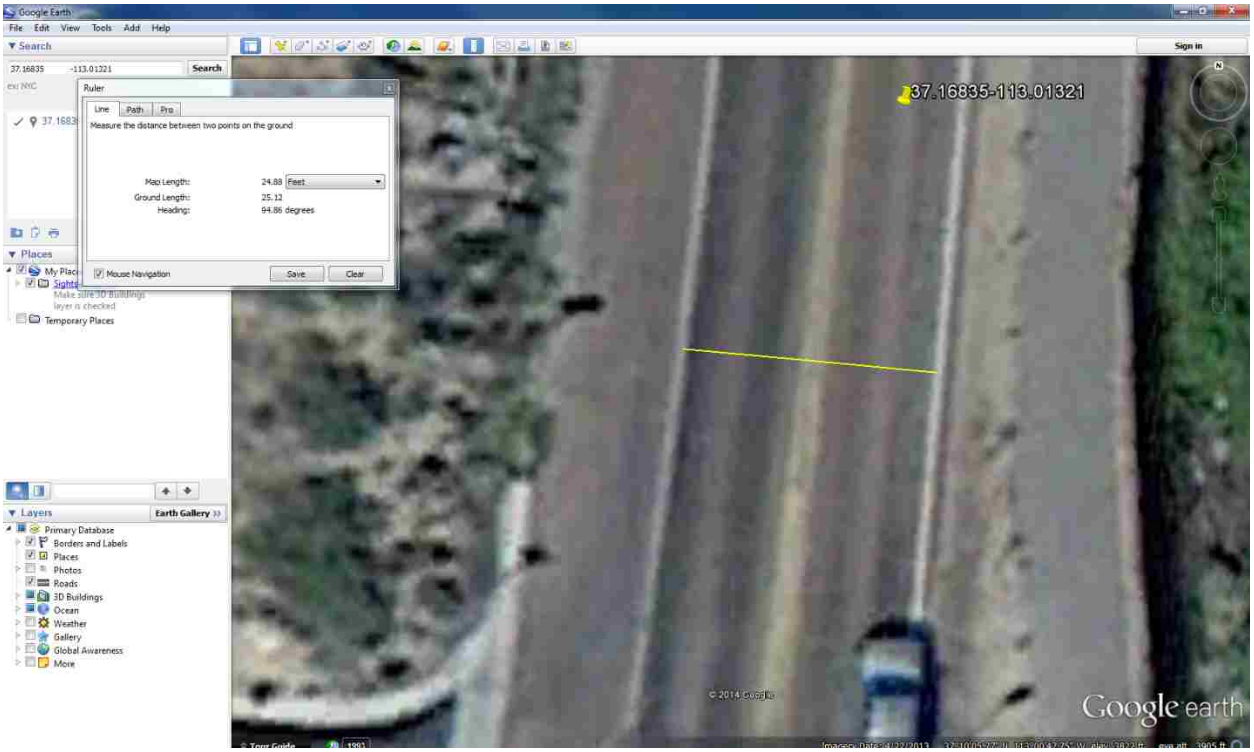
### 3.6.5 Rumble Strip Presence

The rumble strip presence data available from UDOT proved unreliable. It was used as a base for analysis; however, each segment was analyzed based on the information obtained in Roadview Explorer (UDOT 2012) for the actual conditions. Rumble strip presence was recorded for centerline (interior) and shoulder (exterior) implementation.

### 3.6.6 Lane Width

Lane width was measured via Google Earth (Google 2014). Each road was measured from shoulder line to shoulder line and divided by two. Certain roads had asymmetrical arrangements, and in those cases, the average lane width was recorded. Using Google Earth had its limitations, especially when the roadway was adjacent to mountainous or rolling terrain. The software would sometimes assume that the roadway followed the general slope instead of being on a level surface made possible by cut and fill techniques. Even still, the measuring feature on

Google Earth would show the map distance and the ground distance. The map distance is based on latitude and longitude values while ground distance accounts for variations in elevation. The ground distance measurement is almost always larger than the map distance since slope is included in the measurement. The ground distance was always chosen for consistency. Google Earth was the most cost effective means of obtaining the lane width data since the data were not available from UDOT while this study was underway. The HSM predictive model requires all lane widths to be rounded to the nearest whole number. Figure 3-5 shows how the lane width data were collected in Google Earth. The line spanning the width of the highway is the ruler tool within the software. Taking the measurement of the full width allowed an average lane width to be calculated with only one measurement rather than one per lane.



37.16835 -113.01321

Ruler

Line | Path | Pro

Measure the distance between two points on the ground

Map Length: 24.88 Feet

Ground Length: 25.12

Heading: 94.86 degrees

☑ Mouse Navigation

Save | Clear

Figure 3-5: Lane Width Measurement in Google Earth

### 3.6.7 Shoulder Width

Shoulder width data were obtained from UDOT (2014a). However, the shoulder widths were verified on Google Earth (Google 2014) at the same time as the lane width. The measurements were taken from the edge of pavement to the shoulder line. This was done for both sides and an average was calculated.

### 3.6.8 Driveway Density

Driveways were counted using Roadview Explorer (UDOT 2012) and Google Earth (Google 2014). The HSM predictive model specifies that a driveway should only be counted if at least one vehicle uses it per day (AASHTO 2010). This requirement relies on a very subjective evaluation since driveway counts were not available. Residential accesses were always assumed to be used at least once per day, and were thus counted. Farm and other accesses required an evaluation of tire tracks and markings to determine the frequency of use. When tire tracks and markings were plentiful, it was assumed that the driveway in question was used at least once per day. Driveways on opposing sides were counted separately, even if they were aligned like a four-way intersection. Accesses to off-street rest areas were counted. Turnouts and extended shoulders for view areas were treated as one driveway unless there were defined accesses. A driveway that served two or more properties was still treated as one driveway. Side-by-side driveways were treated as separate driveways unless they merged into one driveway before accessing the road. Driveway density was calculated by dividing the number of driveways along the segment by the length of the segment. The units of driveway density are the number of driveways per mile.

### 3.6.9 Passing Ability

The HSM predictive model allows for segments with conventional passing or climbing lanes, provided that the additional lanes are for a limited distance. Similarly, short four-lane sections are allowed under the same stipulation. The data collected specified how many directions had passing ability: zero, one, or two. However, passing ability was also collected on segments with permitted passing zones that did not have additional lanes but instead had a broken centerline. A single yellow broken line was treated as a two-directional passing zone. A double yellow line with one solid and one broken was treated as a one-directional passing zone. A solid double yellow line was treated as a zero-directional passing zone.

### 3.6.10 Lighting

The presence of lighting was observed for each segment. This was done in Roadview Explorer. Overhead street lighting was the only lighting that would qualify. Additionally, the lighting did not need to be present for the majority of the segment like the other geometric attributes—one light would suffice. However, not a single segment in the study had overhead lighting. This is most likely due to the rural location of each segment.

### 3.6.11 AADT

AADT data were obtained from UDOT (UDOT 2011). The data were collected for the years 2008 to 2012—the most recent five years of data available.

### 3.6.12 Truck Percentage

The data for truck percentage are divided into single-unit (single) and combination unit (combo) truck percentages. The definition for single and combo trucks is found in the Traffic

Monitoring Guide from the Federal Highway Administration (FHWA) (2013), and is shown in Table 3-3. The FHWA vehicle category classifications and numbers are shown in Table 3-4.

The vehicle classification data were collected by UDOT at various recording stations across the state and then interpolated throughout UDOT's highway system so that every segment has an associated single and combo truck percentage.

Table 3-3: HPMS/FHWA Vehicle Classes (FHWA 2013)

| **HPMS Summary Table Vehicle Class Group** | **FHWA 13 Vehicle Category Classification Number** |
|---|---|
| Group 1: Motorcycles (MC) | 1 |
| Group 2: Passenger Vehicles equal to or under 102" (PV) | 2 |
| Group 3: Light trucks over 102" (LT) | 3 |
| Group 4: Buses (BS) | 4 |
| Group 5: Single-unit vehicles (SU) | 5, 6, 7 |
| Group 6: Combination Unit (CU) | 8, 9, 10, 11, 12, 13 |

## 3.7 Crash Data

As mentioned previously, crash data were obtained from the UDOT Crash Database for the years 2008 to 2012 (UDOT 2013). Once the segments were randomly selected and the curve buffer was added, the segment data were cross-referenced with the crash data to extract only the crashes that occurred on the segments in the dataset.

Table 3-4: FHWA Vehicle Category Classifications (FHWA 2013)

| Code | Description |
|---|---|
| 1 | Motorcycles (Optional): All two- or three-wheeled motorized vehicles. Typical vehicles in this category have saddle type seats and are steered by handlebars rather than a wheel. This category includes motorcycles, motor scooters, mopeds, motor-powered bicycles, and three-wheeled motorcycles. This vehicle type may be reported at the option of the State, but should not be reported with any other vehicle type. |
| 2 | Passenger Cars: All sedans, coupes, and station wagons manufactured primarily for the purpose of carrying passengers and including those passenger cars pulling recreational or other light trailers. Vehicles registered as passenger cars that are pickups, panels, vans, etc. (described as vehicle type "3") should be reported as vehicle type "3". |
| 3 | Other Two-Axle, Four-Tire, Single-Unit Vehicles: All two-axle, four-tire vehicles, other than passenger cars. Included in this classification are pickups, panels, vans, and other vehicles such as campers, motor homes, ambulances, hearses, and carryalls. Other two-axle, four-tire single-unit vehicles pulling recreational or other light trailers are included in this classification. |
| 4 | Buses: All vehicles manufactured as traditional passenger-carrying buses with two-axles, six-tires and three or more axles. This category includes only traditional buses (including school buses) functioning as passenger-carrying vehicles. All two-axle, four-tire minibuses should be classified as other two-axle, four-tire, single-unit vehicles (type "3"). Modified buses should be considered as trucks and be appropriately classified. |
| 5 | Two-Axle, Six-Tire, Single-Unit Trucks: All vehicles on a single frame including trucks, camping and recreational vehicles, motor homes, etc., having two axles and dual rear wheels. |
| 6 | Three-Axle, Single-Unit Trucks: All vehicles on a single frame including trucks, camping and recreational vehicles, motor homes, etc., having three axles. |
| 7 | Four-or-More Axle, Single-Unit Trucks: All vehicles on a single frame with four or more axles. |
| 8 | Four-or-Less Axle, Single-Trailer Trucks: All vehicles with four or less axles consisting of two units, one of which is a tractor or straight truck power-unit. |
| 9 | Five-Axle, Single-Trailer Trucks: All five-axle vehicles consisting of two units, one of which is a tractor or straight truck power-unit. |
| 10 | Six-or-More Axle, Single-Trailer Trucks: All vehicles with six or more axles consisting of two units, one of which is a tractor or straight truck power-unit. |
| 11 | Five-or-Less Axle, Multi-Trailer Trucks: All vehicles with five or less axles consisting of three or more units, one of which is a tractor or straight truck power-unit. |
| 12 | Six-Axle, Multi-Trailer Trucks: All six-axle vehicles consisting of three or more units, one of which is a tractor or straight truck power-unit. |
| 13 | Seven-or-More Axle, Multi-Trailer Trucks: All vehicles with seven or more axles consisting of three or more units, one of which is a tractor or straight truck power-unit. |

The crash data were tabulated to create totals for the most recent three years (2010-2012) and the most recent five years (2008-2012) on each segment.

## 3.8 Data Preparation Summary

The purpose of data preparation was to randomly select segments that represented a cross section of rural two-way two-lane highways with curves in Utah.  The segments were selected and the facility data pertaining to the selected segments were collected to allow subsequent crash prediction modeling.  The justification for the specific variables collected comes from the HSM predictive model, which outlines the base conditions of any given segment.  With the completion of data preparation, the variables that were required for calibration of the HSM predictive model, along with the additional variables, were used for modeling and further analysis.

# 4 METHODOLOGY

This chapter discusses the process for creating and analyzing the various models that are used in this research. This includes models based on the HSM predictive model, models specific to this research for analyzing both curved and tangent segments, and Utah-specific NB models for curved segments.

## 4.1 HSM Model

This section discusses the development of the predictive models for rural two-lane two-way highways in Utah, as explained in the HSM. The crash prediction model incorporates SPFs, CMFs, and a calibration factor. The calibration factor is what makes the model jurisdiction-specific since it compares the predicted values to the actual values observed on the selected segments within the state.

### 4.1.1 SPF

The HSM outlines process for developing an SPF for rural two-lane two-way highway segments. The SPF predictive model that was introduced in Equation 2-1 is repeated below as Equation 4-1 (AASHTO 2010).

$$N_{spf} = AADT \times L \times 365 \times 10^{-6} \times e^{-0.312} \qquad (4\text{-}1)$$

where, $N_{spf}$ = predicted total crash frequency for roadway segment base conditions

$AADT$ = average annual daily traffic volume (vehicles per day), and

$L$ = length of roadway segment (miles).

As illustrated in Equation 4-1 above, the SPF is based on segment length and AADT, that is, daily VMT. This calculation will stay similar from year to year since the only parameter that fluctuates is AADT. The HSM crash prediction model uses this SPF model to show that the number of crashes on a given segment is directly proportional to the exposure (AADT multiplied by segment length). The multiplier 365 is included to convert AADT from a daily measurement to an annual measurement. The multiplier $10^{-6}$ is to convert the overall units to number of crashes per million VMT. The full model, including the exponential, is based on data from studies performed in the United States (AASHTO 2010).

### 4.1.2 CMFs

As illustrated in the SPF model, the only parameters that vary from segment to segment are AADT and segment length. The SPF equation assumes a base condition for each road segment. The base conditions for rural two-lane two-way highways are shown in Table 4-1.

When a segment does not meet the base condition for any one of the 13 listed in Table 4-1, a CMF must be multiplied to the predicted number of crashes calculated by the SPF model. This will adjust the prediction by incorporating more of the actual parameters. The new prediction model is shown as Equation 4-2 (AASHTO 2010).

Table 4-1: Base Conditions for Rural Two-Lane Two-Way Highways

| | |
|---|---|
| Lane Width | 12 feet |
| Shoulder Width | 6 feet |
| Shoulder Type | Paved |
| Roadside Hazard Rating | 3 |
| Driveway Density | 5 driveways per mile |
| Horizontal Curvature | None |
| Vertical Curvature | None |
| Centerline Rumble Strips | None |
| Passing Lanes | None |
| Two-way left-turn lanes | None |
| Lighting | None |
| Automated Speed Enforcement | None |
| Grade Level | 0% |

$$N = N_{spf} \times CMF_1 \times CMF_2 \times \ldots \times CMF_i \qquad (4\text{-}2)$$

where, $N$ = predicted number of crashes accounting for non-base conditions,

$N_{spf}$ = number of predicted crashes determined for base conditions, and

$CMF_i$ = crash modification factor.

The model in Equation 4-2 shows that CMFs directly affect the predicted number of crashes for a given segment. A CMF with a value greater than 1 will increase the predicted number of crashes, while a CMF with a value less than 1 will decrease the predicted number of crashes. Independent CMFs can be created and multiplied to an SPF; however, there are 12 CMFs that the HSM has identified as the most relevant to crash prediction (shoulder width and type are combined into one CMF).

Deviations from the base conditions assumed in the SPF model are expected on almost all rural two-lane two-way highway segments. One of the frequently unmet base conditions is the zero percent grade specification. Many jurisdictions do not allow roads to be constructed with a zero percent grade as drainage can be compromised (AASHTO 2010). Thus, a CMF will almost always be calculated for grade. The base conditions are not necessarily ideal conditions; rather, they are a starting point for further analysis.

### 4.1.3 Calibration

The HSM predictive model was developed from data that was sourced from several regions in the United States. The use of the HSM predictive model, however, is generally used in a local setting. Overall conditions, such as winter weather and driver behavior, can vary greatly from state to state. For this reason, the HSM predictive model incorporates a calibration factor that jurisdictions may employ to adjust the predicted value to match actual observed crash rates. The full predictive model, including calibration, is shown in Equation 4-3 (AASHTO 2010):

$$N_{pred} = N_{spf} \times C \times CMF_1 \times CMF_2 \times \ldots \times CMF_i \qquad (4\text{-}3)$$

where, $N_{pred}$ = predicted number of crashes,

$N_{spf}$ = number of predicted crashes determined for base conditions,

$C$ = calibration factor, and

$CMF_i$ = crash modification factor.

The calibration factor is found by dividing the actual number of crashes by the predicted number of crashes as shown in Equation 4-4.

$$\mathrm{C} = \frac{\mathrm{N_{actual}}}{\mathrm{N_{pred}}} \qquad (4\text{-}4)$$

where, $C$ = calibration factor,

$N_{actual}$ = actual number of crashes, and

$N_{pred}$ = predicted number of crashes.

Unlike SPFs and CMFs, calibration factors are calculated from an entire set of segments rather than from each segment individually. However, once a single calibration factor has been established, it is used for each segment. A calibration factor greater than 1 indicates that the roadway segments within the set experience more crashes, on average, than the roadways that were used in developing the SPFs (AASHTO 2010). Conversely, a factor less than 1 indicates fewer crashes, on average, than the roadways used in developing the SPFs.

### 4.1.4 HSM Model Summary

The purpose of this subsection was to describe the process by which the HSM predictive model calculates the predicted number of crashes for specified roadway segments. The HSM predictive model uses an SPF calculation to establish a prediction based on vehicle exposure. Beyond exposure, CMFs must be used to account for variations in the prescribed base conditions. These will adjust the prediction for each segment up or down based on facility data. Similarly, the predicted number of crashes for a set of segments can be pooled and compared to the actual number of crashes. This ratio will produce a calibration factor that can be multiplied to the SPFs and CMFs to create the full predictive model.

The HSM predictive model was created to fit any jurisdiction by developing specific calibration factors. However, this research will look into the development of Utah-specific models in the following subsections.

## 4.2 Curve and Tangent Combination

Thus far, this research has focused on the development of SPFs and corresponding calibration factors for curved segments on rural two-lane two-way highways in Utah. Previous research (Saito et al. 2011) has been performed on tangent segments on rural two-lane two-ways highways, and did not include curved segments due to the difficulty in obtaining horizontal curvature data. With the recent acquisition of highway curvature data through UDOT Roadway Imaging/Asset Inventory project (Ellsworth 2013), this research was able to focus on curved segments of rural two-lane two-way highways.

It was hypothesized, however, that a model could be created using both tangent and curved segments. Since almost all highways are a mix of curved and tangent segments, it would be useful to create a way for any segment to be analyzed with one all-encompassing model. This section will present the approach in the HSM for incorporating both curved and tangent segments and also address different methods used for parameterizing horizontal curvature on segments including a simple indicator variable for curve or tangent, a series of indicator variables for curve class, a continuous variable for curve radius, and a continuous variable for the inverse transformation of curve radius.

### 4.2.1 HSM Approach

As described in the previous chapter, the HSM predictive model includes a CMF that specifically adjusts the SPF for horizontal alignment variations. The equation for the CMF is shown below in Equation 4-5 (AASHTO 2010).

$$CMF = \frac{(1.55 \times L_C) + \left(\frac{80.2}{R}\right) - (0.012 \times S)}{(1.55 \times L_C)} \tag{4-5}$$

where, *CMF* = crash modification factor for horizontal alignment;

$L_c$ = length of curve, mi.;

*R* = radius, ft.; and

*S* = 1 if spiral transition curve is present; 0 if spiral transition curve is not present; 0.5 if a spiral transition curve is present at one but not both ends of the horizontal curve.

This model incorporates both the length and radius of a curve as well as making adjustments for spiral transitions. It was assumed in this research that all curves in Utah are not equipped with spiral transitions. While some CMF models allow for a result above and below 1.0, this CMF can only increase from 1.0, which is base value for a tangent segment. The CMF is inversely proportional to the curve radius, which means that the CMF approaches 1 as the radius increases. Thus the sharpest curves have the highest CMFs.

### 4.2.2 Parameterization of Horizontal Curvature

Where the HSM incorporates horizontal curvature into a CMF, a Utah-specific model would need to incorporate horizontal curvature as a parameter in an NB model. Several models were attempted for parameterizing horizontal curvature, including separating curves and tangents by a simple indicator variable; curve class as a series of indicator variables; using curve radius; and using an inverse transformation of curve radius.

#### 4.2.2.1 Simple Indicator Variable

The first proposed model uses an indicator variable for horizontal alignment by assigning 0 for tangent segments and 1 for curved segments. This model creates a simple method for analyzing both curves and tangents with very minimal data collection. The data requirement is the mere identification of horizontal curvature. This is the simplest model for incorporating horizontal curvature into a Utah-specific NB model.

#### 4.2.2.2 Curve Class as a Series of Indicator Variables

The second proposed model allows for more detail than the simple indicator variable model. It assigns each curve a classification based on degree of curvature using the HPMS definitions introduced in Chapter 4. The classification breakdown outlined in Table 3-1 is reprinted in Table 4-2 for convenience.

Since this model involves tangent segments, the definition for curve class A is modified to include curves that have curvature greater than 0.0 degrees up to 3.5 degrees so that a new tangent classification can be introduced. This creates a set of seven variables, each one with possible values of 1 (if the curve in question falls within the range) and 0 (if the curve falls anywhere outside the range). For example, a curve with a C classification would produce a value of 1 for the C class indicator variable, and a 0 for all other indicator variables.

Table 4-2: HPMS Curve Classification Breakdown (FHWA 2014)

| Curve Classification | Degrees |
|---|---|
| A | Under 3.5 degrees (i.e., 0.061 radians) |
| B | 3.5 - 5.4 degrees (i.e., 0.061 - 0.094 radians) |
| C | 5.5 - 8.4 degrees (i.e., 0.096 - 0.147 radians) |
| D | 8.5 - 13.9 degrees (i.e., 0.148 - 0.243 radians) |
| E | 14.0 - 27.9 degrees (i.e., 0.244 - 0.487 radians) |
| F | 28 degrees (i.e., 0.489 radians) or more |

Each class is changed from a letter to a number to allow for parameterization. Tangent becomes 0, class A becomes 1, all the way through class F becoming 6. The modified breakdown is shown in Table 4-3. Each curve class is treated as a separate indicator variable. This model allows for isolation of specific classifications that may correlate better than other classifications.

Table 4-3: Modified Curve Classification Breakdown

| Curve Classification | Degrees |
|---|---|
| 0 | 0.0 degrees (no curvature) |
| 1 | >0.0 - 3.5 degrees (i.e., >0.000 - 0.061 radians) |
| 2 | 3.5 - 5.4 degrees (i.e.,0.061 - 0.094 radians) |
| 3 | 5.5 - 8.4 degrees (i.e., 0.096 - 0.147 radians) |
| 4 | 8.5 - 13.9 degrees (i.e., 0.148 - 0.243 radians) |
| 5 | 14.0 - 27.9 degrees (i.e., 0.244 - 0.487 radians) |
| 6 | 28 degrees (i.e., 0.489 radians) or more |

This model incorporates horizontal alignment as well as groupings for the differing sharpness of curved segments. This is in contrast to the simple indicator variable model, which classifies all curves as the same within the parameter.

#### 4.2.2.3 Curve Radius

The third proposed model uses curve radius, rather than a classification derived from degree of curvature. Also, instead of grouping curves into classification bins, this model uses a continuous variable. The challenge comes in assigning a radius to tangent segment. Since radius increases as a curve becomes shallower, the radius of a tangent would theoretically be infinite. Since infinity is impractical from a modeling standpoint, an arbitrarily high radius of 10 miles is assigned to each tangent segment. This model requires the radius measurement for each curved segment.

#### 4.2.2.4 Inverse Transformation of Curve Radius

The fourth proposed model is very similar to the Curve Radius model, and simply requires an algebraic transformation. The idea for this model is that the value for tangent segments should not be arbitrary. As mentioned previously, radius increases as a curve becomes shallower so a tangent segment would have an infinite radius. By taking the inverse of the radius, the value for tangent segments is 0, with all curves having increasing values as they become sharper. This model creates a better distribution of values that matches the model outlined in the HSM (AASHTO 2010).

### 4.2.3 Curve and Tangent Combination Summary

This section has evaluated several models for addressing horizontal alignment in a crash prediction model. Included in the evaluation were the models laid out in the HSM, as well as models for parameterization of horizontal alignment for use in an NB model. The models of parameterization include using an indicator variable, curve class as a series of indicator variables, curve radius, and inverse transformation of curve radius. Each model has strengths and weaknesses. The results of their use will be discussed in Chapter 5.

## 4.3 Utah-Specific Model

This section discusses the creation of models to predict crashes on rural two-lane two-way highways in Utah. Previous subsections have discussed the predictive model outlined in the HSM and techniques for combining curves and tangents into one model using a variety of parameters to properly account for horizontal alignment. While the previous subsection discussed NB models, the focus was on developing a parameter for modeling horizontal curvature. This section focuses on the overall development of both an NB model and an EB model using all independent variables that are statistically significant.

### 4.3.1 Negative Binomial Development

The development of an NB model was performed using JMP, a statistical software package that is a graphical interface for SAS software (SAS 2013). JMP will create an NB model with any number of independent variables and interactions of variables. JMP will estimate the coefficients for each variable within a model and calculate the p-value for each. A p-value is the probability that a randomized experiment will lead to a test statistic that is as extreme as or more extreme than the one observed (Ramsey and Schafer 2002). Using the p-values, researchers were able to use a backward stepwise technique for identifying which variables are significant and which are not. A backward stepwise technique allows the model to begin with as many variables as are entered. These input variables are then removed one at a time based on their p-value—the variable with the highest p-value is eliminated because the value indicates much less contribution to the integrity of the model. A new model is then created with the remaining variables. This process is continued until all variables have p-values less than 0.05, based on a 95 percent confidence level. This technique allows for every variable to be entered and only the relevant ones will remain after the process is completed. The NB model will take the form shown in Equation 4-6 (Ramsey and Schafer 2002). The equation can be rearranged by exponentiating both sides in order to solve for the number of crashes. The number of independent variables will depend on the results of the backward stepwise technique.

$$\ln(N) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (4\text{-}6)$$

where, $N$ = number of crashes (predicted or observed),

$\beta_0$ = intercept,

$\beta_i$ = coefficient for variable $x_i$,

$x_i$ = independent variable, and

$n$ = number of independent variables.

The input variables for the backward stepwise technique include the same variables that were used for the HSM predictive model and also include additional variables that were hypothesized to have a potential correlation with crash prediction. Table 4-4 shows the preliminary input variables in alphabetical order.

Table 4-4: Preliminary Input Variables for NB Model

| | |
|---|---|
| AADT | Passing Lane Presence |
| Analysis Length | Radius |
| Combo Truck Percentage | Rumble Strip Presence |
| Degree of Curvature | Shoulder Width |
| Driveway Density | Single Truck Percentage |
| Grade | Speed Limit |
| Lane Width | Total Truck Percentage |

The Passing Lane Presence and Rumble Strip Presence variables were simple indicator variables—a value of 1 if the item was present; a value of 0 if the item was not present. Speed Limit was based on posted speed limit for the segment in increments of 5 miles per hour (mph). The Lane Width and Shoulder Width variables followed the same rounding convention discussed in previous chapters, with Lane Width rounding to the nearest foot, and Shoulder Width rounding down to the nearest multiple of two feet. The AADT and Radius variables underwent transformations to create more normal distributions and to reduce the differences in variances.

AADT values ranged from around 300 to over 10,000 vehicles per day (vpd). The NB models were created for both a three-year sample (2010-2012) and a five-year sample (2008-2012), so each AADT value was multiplied by the number of days per year (365) and by the number of years in the sample (3 or 5, depending on the dataset of the model). Once the product of AADT, days per year, and years of data was calculated—now more appropriately called

Vehicle Count—it was determined that a natural log transformation would be the best approach for this variable.

The Radius variable, which has been discussed in previous chapters, also underwent transformations.  The largest radius in the dataset was larger than the smallest radius by three orders of magnitude.  For this reason, it was determined that a log transformation would create a distribution closer to a normal distribution.  Similarly, the idea of an inverse transformation was proposed based on the success observed in the previous chapter.  So in all, radius, radius with natural log transformation, and inverse radius were all included in the model.

Table 4-5 shows the final input variables for the backward stepwise regression technique in JMP.  By performing the backward stepwise regression technique in JMP, each variable can be analyzed individually based on correlation and p-value, and the best variable can be chosen out of potentially overlapping or duplicative variables.

Table 4-5:  Final Input Variable Selection for NB Model

| | |
|---|---|
| Analysis Length | Radius |
| Combo Truck Percentage | Radius with Natural Log Transformation |
| Degree of Curvature | Rumble Strip Presence |
| Driveway Density | Shoulder Width |
| Grade | Single Truck Percentage |
| Inverse Radius | Speed Limit |
| Lane Width | Total Truck Percentage |
| Passing Lane Presence | Vehicle Count |

### 4.3.2 Empirical Bayes Model

The EB model creates a model based on the results of a crash prediction model as well as the actual number of crashes.  It uses a dispersion parameter to create a weight to assign to both the prediction and the actual number of crashes. The dispersion parameter is part of the output data on JMP when creating a NB model. The general equation for the EB model is shown in Equation 4-7 (Hauer 1997).

$$N_{expected} = w \times N_{spf} + (1 - w) \times N_{observed} \tag{4-7}$$

where, $N_{expected}$ = expected number of crashes determined by the EB method,

$w$ = weight (as shown in Equation 4-8),

$N_{spf}$ = number of predicted crashes (previously determined), and

$N_{observed}$ = observed number of crashes at a site.

$$w = \frac{1}{1 + k \times (N_{spf})} \tag{4-8}$$

where, $k$ = dispersion parameter.

Since the EB model uses a combination of predicted and observed, it can more closely approximate the number of crashes. The weight assigned to both the predicted and observed values will change depending on the dispersion of the crash data. With widely dispersed crash data, the EB model will place more weight on the observational data. The EB model is especially appropriate for before-after studies (Srinivasan et al. 2009).

## 4.4 Methodology Summary

This section has discussed both the NB model and the EB model, which both provide estimates of the number of crashes that can be expected on a road segment. The NB model can use a backward stepwise regression technique that will isolate only the significant variables depending on a chosen confidence level. The EB model looks at both predicted crashes and observed crashes to create an expected value for the number of crashes on a given segment. This is done with a dispersion parameter that is given through JMP when an NB model is created. The dispersion determines the weight given to both the predicted values and the observed values. The next chapter will look at the results of the Utah-specific model as well as the HSM predictive model and curve and tangent combination discussed earlier in this chapter.

# 5 RESULTS

This chapter presents the results of the crash prediction modeling efforts, which were discussed in previous chapters. The modeling efforts include the calibration of the HSM predictive model, the development of a model which incorporates both curved and tangent segments, as well as the development of a Utah-specific model. First, a summary of data collection efforts will be presented. Then, the calibration factors that were determined for the HSM predictive model will be discussed. The next section will discuss the results of modeling horizontal alignment by combining curved segments and tangent segments into a single parameter. The NB regression and EB models will then be discussed as part of the Utah-specific model development for curved segments. A summary of the data collection efforts and modeling results will conclude this chapter.

## 5.1 Data Collection Efforts

As discussed in Section 3.6, the data collection took place for each randomly selected segment that was part of the analysis datasets. Gathering sufficient data to develop a comprehensive model proved to be a difficult task. The HSM predictive model states that a dataset should have no fewer than 100 total crashes per year (AASHTO 2010). However, when curves were randomly chosen for analysis, there was no way of knowing the number of actual crashes on each segment until after extensive analysis. This is because the analysis length was based on curve length plus a buffer on both ends of the curve. The calculations for the buffer depended on lane width and speed limit data which needed to be gathered.

Originally, approximately 200 segments were randomly selected for analysis, since the previous research by Saito et al. (2011) had 157 segments with an average of 142 crashes per year. However, the segments for that research were significantly longer since they were tangent segments divided into homogeneous sections. It was not uncommon for a segment to span several miles. Longer segments corresponded with an increased number of crashes, thus allowing the tangent dataset to achieve the 100 crash per year threshold with fewer segments. Curved segments, on the other hand, are significantly shorter on average than tangent segments. Since the number of crashes on any given segment is heavily based on vehicle exposure, it makes sense that more curved segments would need to be included in the dataset to reach the 100 crashes per year threshold.

For the first dataset, the original group of approximately 200 randomly selected segments was combined with two additional groups of randomly selected segments with similar quantities. This combined dataset was evaluated to remove any duplicate or overlapping segments before any statistical analysis was performed. In total, the dataset comprised 579 segments with an average of 112 crashes per year for the three-year sample and 126 crashes per year for the five-year sample. This satisfied the HSM requirement of at least 100 crashes per year.

Two subsequent datasets were assembled for validation purposes, resulting in three total datasets comprised of curved segments only. Dataset 2 had 566 segments with an average of 109 crashes per year for the three-year sample and 113 crashes per year for the five-year sample. Dataset 3 had 608 segments with an average of 140 crashes per year for the three-year sample and 150 crashes per year for the five-year sample. All three datasets met the HSM requirement of at least 100 crashes per year for both the three-year and five-year samples.

With three separate random datasets, this research was also able to look at a combined dataset of all three samples with overlapping and duplicate segments eliminated. This combined dataset had 1,495 segments with an average of 319 crashes per year for the three-year sample and 343 crashes per year for the five-year sample.

## 5.2 Calibration of the HSM Predictive Model

The HSM predictive model is based on an SPF, multiple CMFs, and a calibration factor as explained in Section 4.1. The SPF for the base conditions on rural two-lane two-way roads is shown in Equation 5-1.

$$N_{spf} = AADT \times L \times 365 \times 10^{-6} \times e^{-0.312} \tag{5-1}$$

where, $N_{spf}$ = predicted total crash frequency for roadway segment base conditions,

$AADT$ = average annual daily traffic volume (vehicles per day), and

$L$ = length of roadway segment (miles).

After calculating the number of predicted crashes with the SPF along with the all of the available CMFs outlined in the HSM, the predicted values were compared to the actual values. This allowed the calibration factors to be determined for each dataset along with the combined dataset. The calibration factors are shown in Table 5-1. Equations 5-2 and 5-3 present the HSM

SPFs for rural two-lane two-way highway segments that have been calibrated for Utah. Equation 5-2 is based on the three-year sample (2010-2012) and Equation 5-3 is based on the five-year sample (2008-2012).

Table 5-1: HSM Predictive Model Calibration Factors

| Set | 3-year Sample | 5-year Sample |
|---|---|---|
| 1 | 1.42 | 1.58 |
| 2 | 1.50 | 1.54 |
| 3 | 1.53 | 1.64 |
| Combined | 1.50 | 1.60 |

$$N_{spf3} = 1.50 \times AADT \times L \times 365 \times 10^{-6} \times e^{-0.312} \times CMF_1 \times CMF_2 \times \ldots \times CMF_i \qquad (5\text{-}2)$$

$$N_{spf5} = 1.60 \times AADT \times L \times 365 \times 10^{-6} \times e^{-0.312} \times CMF_1 \times CMF_2 \times \ldots \times CMF_i \qquad (5\text{-}3)$$

where, $N_{spf3}$ = predicted total crash frequency for roadway segment base conditions using a three-year sample,

$N_{spf5}$ = predicted total crash frequency for roadway segment base conditions using a five-year sample,

$AADT$ = average annual daily traffic volume (vehicles per day),

$L$ = length of roadway segment (miles), and

$CMF_i$ = crash modification factor.

Equations 5-2 and 5-3 have been simplified to Equations 5-4 and 5-5.

$$N_{spf3} = AADT \times L \times 4.01 \times 10^{-4} \times CMF_1 \times CMF_2 \times \ldots \times CMF_i \qquad (5\text{-}4)$$

$$N_{spf5} = AADT \times L \times 4.27 \times 10^{-4} \times CMF_1 \times CMF_2 \times \ldots \times CMF_i \qquad (5\text{-}5)$$

As explained above, Equations 5-4 and 5-5 represent the combination of an SPF and the calibration factor. The CMFs still need to be applied to these equations as prescribed by the HSM predictive model for rural two-lane two-way highways. It is interesting to note that the calibration factor for the three-year sample is lower than the calibration factor for the five-year sample. This implies that the overall safety on the sampled highway segments improved in the last three years of the five-year sample. With this implication, the tangent segments evaluated in previous research (Saito et al. 2011) were used to develop calibration factors with more recent crash data. Since the previous research used a three-year sample, the data used for this research was grouped into three separate three-year samples for comparison. The calibration factors for the tangent segments are shown in Table 5-2.

Table 5-2: Calibration Factors for Tangent Segment Sample

| Sample Period | Predicted Crashes | Actual Crashes | Calibration Factor |
|---|---|---|---|
| 2005-2007 | 368 | 426 | 1.16 |
| 2008-2010 | 403 | 415 | 1.03 |
| 2009-2011 | 403 | 374 | 0.93 |
| 2010-2012 | 422 | 354 | 0.84 |

It is important to remember that AADT is the only independent variable that changes between sample periods, based on the HSM predictive model, since geometric features are assumed to be the same. While AADT has increased on almost all segments since the previous research, the actual number of crashes has decreased. A decreasing calibration factor signifies that either the actual number of crashes is decreasing, the predicted number of crashes is increasing, or both. This supports the assumption that overall safety has improved not only since the year 2008, but since at least the year 2005. The improvements in safety can be seen in both the curved segment sample and the tangent segment sample. This also shows that calibration factors need to be updated regularly, as they can change significantly within a few years' time.

## 5.3 Curve and Tangent Combination

This section will discuss the results of the attempts to parameterize horizontal alignment including both curved segments and tangent segments. Four different approaches were attempted to combine segments of different horizontal alignment into the same parameter as outlined in Section 4.2.2. These models included single indicator variable, curve class as a series of indicator variables, curve radius, and inverse transformation of curve radius. A model was created for each approach and for each sample period using the backward stepwise technique outlined in Section 4.3.1. The results of each model are presented in this section along with a discussion of the next steps in this area of research.

### 5.3.1 Single Indicator Variable

The single indicator variable model used an indicator variable for horizontal alignment by assigning 0 for tangent segments and 1 for curved segments. While this is a very simple model with minimal data collection, the results were inconclusive. Since the variable only allowed for two possibilities, all curved segments were treated as equal as were all tangent segments. Any variation in curve radius or degree of curvature was ignored by this model. The parameter estimates are shown in Table 5-3 and 5-4 for the three-year and five-year samples, respectively.

Table 5-3: Parameter Estimates for Three-year Sample Using Single Indicator Variable

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -8.9068 | 0.8798 | 102.5 | < 0.0001 | -10.6311 | -7.1825 |
| Analysis Length (mi) | 0.6817 | 0.1113 | 37.5 | < 0.0001 | 0.4636 | 0.8998 |
| Total Truck Percentage | -0.0165 | 0.0051 | 10.3 | 0.0013 | -0.0265 | -0.0064 |
| Ln(3 year Vehicle Count) | 0.6430 | 0.0580 | 123.1 | < 0.0001 | 0.5294 | 0.7565 |
| Curve Presence [1] | -0.4494 | 0.1594 | 8.0 | 0.0048 | -0.7618 | -0.1370 |
| Dispersion | 0.9360 | 0.1288 | 52.8 | < 0.0001 | 0.6836 | 1.1885 |

Table 5-4: Parameter Estimates for Five-year Sample Using Single Indicator Variable

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -8.1664 | 0.7303 | 125.0 | < 0.0001 | -9.5978 | -6.7350 |
| Analysis Length (mi) | 0.6602 | 0.1033 | 40.8 | < 0.0001 | 0.4576 | 0.8627 |
| Total Truck Percentage | -0.0169 | 0.0043 | 15.8 | < 0.0001 | -0.0253 | -0.0086 |
| Ln(5 year Vehicle Count) | 0.6400 | 0.0483 | 175.2 | < 0.0001 | 0.5424 | 0.7348 |
| Curve Presence [1] | -0.5229 | 0.1382 | 14.3 | 0.0002 | -0.7939 | -0.2520 |
| Dispersion | 0.8525 | 0.0937 | 82.7 | < 0.0001 | 0.6687 | 1.0362 |

Tables 5-3 and 5-4 include the Wald statistic in the fourth column. The Wald statistic is a comparison of the maximum likelihood estimate of the parameter and the proposed value (SAS 2013). The statistic is then compared to a chi-squared distribution to produce a p-value shown in the fifth column. The variables in the shown in Table 5-3 and 5-4 in addition to the Curve Presence variable are the only variables whose p-values were less than 0.05 for these models.

Using curve presence as a single indicator variable is shown to be significant at a 95 percent confidence level, with a p-value of 0.0002. Of particular interest is the sign for the indicator variable estimate. The interpretation of this model is that the presence of curves reduces the overall number of crashes, which is counterintuitive. This idea will be discussed further in Section 5.3.5. This model groups all curves and tangents into two homogeneous classifications. The model, therefore, does not account for sharpness of each curve. Due to this limitation, this model was rejected.

### 5.3.2 Curve Class as a Series of Indicator Variables

This model involved assigning a classification to each curve based on degree of curvature using the HPMS definitions introduced in Chapter 3. Each classification was changed from a letter to a number to allow for parameterization. Tangent became 0, A became 1, B became 2, and so forth, all the way through F becoming 6, as outlined in Table 4-3. Each classification was treated as a separate indicator variable. This created a set of seven variables, each one with possible values of 1 (if the curve in question fell within the range) and 0 (if the curve fell anywhere outside the range). For example, a curve with a C classification produced a value of 1 for the C class indicator variable, and a 0 for all other indicator variables. In the JMP model, the F (or 6) classification became the base group; hence, only six indicator variables are listed in the model output. Tables 5-5 and 5-6 show the parameter estimates for the three-year and five-year samples, respectively.

Table 5-5: Parameter Estimates for Three-year Sample Using Curve Class

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -0.9858 | 1.1417 | 74.6 | < 0.0001 | -12.0959 | -7.6204 |
| Analysis Length (mi) | 0.6895 | 0.1116 | 38.2 | < 0.0001 | 0.4709 | 0.9081 |
| Total Truck Percentage | -0.0158 | 0.0052 | 9.4 | 0.0022 | -0.0260 | 0.0057 |
| Numeric Class [0] | 0.7444 | 0.8334 | 0.8 | 0.3718 | -0.8891 | 2.3779 |
| Numeric Class [1] | 0.2163 | 0.8250 | 0.1 | 0.7932 | -1.4007 | 1.8333 |
| Numeric Class [2] | 0.3398 | 0.8295 | 0.2 | 0.6820 | -1.2859 | 1.9656 |
| Numeric Class [3] | 0.3386 | 0.8368 | 0.2 | 0.6857 | -1.3014 | 1.9786 |
| Numeric Class [4] | 0.6686 | 0.8482 | 0.6 | 0.4306 | -0.9939 | 2.3311 |
| Numeric Class [5] | 0.3046 | 0.9447 | 0.1 | 0.7471 | -1.5469 | 2.1562 |
| ln(3 year Vehicle Count) | 0.6551 | 0.0587 | 124.7 | < 0.0001 | 0.5402 | 0.7701 |
| Dispersion | 0.9260 | 0.1276 | 52.7 | < 0.0001 | 0.6759 | 1.1761 |

Table 5-6: Parameter Estimates for Five-year Sample Using Curve Class

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -9.7737 | 1.0456 | 87.4 | < 0.0001 | -11.8229 | -7.7244 |
| Analysis Length (mi) | 0.6668 | 0.1036 | 41.4 | < 0.0001 | 0.4638 | 0.8699 |
| Total Truck Percentage | -0.0169 | 0.0043 | 15.5 | < 0.0001 | -0.0253 | -0.0085 |
| Numeric Class [0] | 1.4691 | 0.8189 | 3.2 | 0.0728 | -0.1359 | 3.0742 |
| Numeric Class [1] | 0.8807 | 0.8115 | 1.2 | 0.2778 | -0.7097 | 2.4712 |
| Numeric Class [2] | 0.9474 | 0.8150 | 1.4 | 0.2451 | -0.6500 | 2.5447 |
| Numeric Class [3] | 1.0711 | 0.8190 | 1.7 | 0.1909 | -0.5341 | 2.6763 |
| Numeric Class [4] | 1.1930 | 0.8294 | 2.1 | 0.1503 | -0.4326 | 2.8185 |
| Numeric Class [5] | 1.0503 | 0.8895 | 1.4 | 0.2377 | -0.6931 | 2.7937 |
| ln(5 year Vehicle Count) | 0.6488 | 0.0489 | 176.1 | < 0.0001 | 0.5530 | 0.7446 |
| Dispersion | 0.8443 | 0.0929 | 82.6 | < 0.0001 | 0.6622 | 1.0264 |

The results for this model were inconclusive. The model rejected each classification indicator variable based on p-values greater than 0.05. Further analysis observed that the samples did not contain an equal distribution across classifications. This has to do with the actual distribution of all curves within the state. There simply are not as many sharp curves as there are shallow curves. For all of these reasons, this model was rejected.

### 5.3.3 Curve Radius

The curve radius model used curve radius, rather than a classification derived from degree of curvature. The main advantage to this model seemed to be the use of a continuous variable rather than an indicator or a grouping. The challenge, however, came from assigning a radius value to each tangent segment. Since radius increases as a curve becomes shallower, the radius of a tangent would theoretically be infinite. Since infinity is impractical from a modeling standpoint, an arbitrarily high radius was assigned to each tangent segment. A value of 10 miles was assigned as the radius for each tangent segment. Tables 5-7 and 5-8 show the parameter estimates for the three-year and five-year samples, respectively.

Table 5-7: Parameter Estimates for Three-year Sample Using Curve Radius

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -9.3748 | 0.8374 | 125.3 | < 0.0001 | -11.0160 | -7.7336 |
| Analysis Length (mi) | 0.6879 | 0.1118 | 37.8 | < 0.0001 | 0.4687 | 0.9070 |
| Total Truck Percentage | -0.0164 | 0.0580 | 10.2 | 0.0014 | -0.0264 | -0.0063 |
| Ln(3 year Vehicle Count) | 0.6429 | 0.0580 | 122.7 | < 0.0001 | 0.5291 | 0.7567 |
| Radius10 (mi) | 0.0453 | 0.0167 | 7.4 | 0.0066 | 0.0126 | 0.0780 |
| Dispersion | 0.9374 | 0.1290 | 52.8 | < 0.0001 | 0.6846 | 1.1902 |

Table 5-8: Parameter Estimates for Five-year Sample Using Curve Radius

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -8.7093 | 0.6946 | 157.2 | < 0.0001 | -10.0708 | -7.3479 |
| Analysis Length (mi) | 0.6672 | 0.1039 | 41.2 | < 0.0001 | 0.4635 | 0.8709 |
| Total Truck Percentage | -0.0168 | 0.0043 | 15.6 | < 0.0001 | -0.0252 | -0.0085 |
| Ln(5 year Vehicle Count) | 0.6397 | 0.0484 | 174.5 | < 0.0001 | 0.5448 | 0.7346 |
| Radius10 (mi) | 0.0528 | 0.0144 | 13.4 | 0.0003 | 0.0245 | 0.0811 |
| Dispersion | 0.8542 | 0.0939 | 82.8 | < 0.0001 | 0.6902 | 1.0383 |

The variable for Curve Radius had a very low p-value in each model, signifying strong evidence of a relation to crash prediction. The estimate for the Radius variable can be interpreted as an increase in crashes as the radius increases. Similar to the Single Indicator Variable Model results, this is counterintuitive as a sharper curve would be expected to have more crashes than a shallow curve or even a tangent. This will be discussed further in Section 5.3.5.

Assigning an arbitrary radius of 10 miles to each tangent segment resulted in a wide range, with all tangents at the extreme upper end of the range and all curves at the lower end of the range. It was determined that a value would have to be identified as the dividing point, and all segments with radii greater than the dividing point would be classified as tangents and assigned the value of the dividing point. A one-mile radius was preliminarily discussed as a possibility for such a dividing point. However, since attempting to undertake the task of defining what is and isn't a curve will require significantly more research, it was deemed beyond the scope of this research.

### 5.3.4 Inverse Transformation of Curve Radius

This model is very similar to the Curve Radius model, and simply required an algebraic transformation. The impetus for this model was difficulty in assigning an arbitrary radius to tangent segments. As mentioned previously, radius increases as a curve becomes shallower so a tangent segment would have an infinite radius. By taking the inverse of the radius, the value for tangent segments became 0, with all curves having increasing values as they became sharper. This model was the easiest to conceptualize: tangent segments have a value of 0; the shallowest of curves would have values close to 0; and as the curves become sharper, the value would increase. Tables 5-9 and 5-10 show the parameter estimates for the three-year and five-year samples, respectively. Unfortunately, the results were inconclusive as the p-value for the Inverse Radius variable was very high in both the three-year and five-year samples.

Table 5-9: Parameter Estimates for Three-year Sample Using Inverse Curve Radius

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -9.7996 | 0.8376 | 136.9 | < 0.0001 | -11.4412 | -8.1580 |
| Analysis Length (mi) | 0.8769 | 0.1018 | 47.3 | < 0.0001 | 0.6775 | 1.0764 |
| Total Truck Percentage | -0.0137 | 0.0051 | 7.2 | 0.0073 | -0.0238 | -0.0037 |
| Ln(3 year Vehicle Count) | 0.6725 | 0.0577 | 135.9 | < 0.0001 | 0.5594 | 0.7855 |
| Inverse Radius (mi) | -0.0022 | 0.0110 | 0.0 | 0.8385 | -0.0238 | 0.0193 |
| Dispersion | 0.9681 | 0.1314 | 54.3 | < 0.0001 | 0.7105 | 1.2256 |

Table 5-10: Parameter Estimates for Five-year Sample Using Inverse Curve Radius

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -9.1010 | 0.7015 | 168.3 | < 0.0001 | -10.4760 | -7.7260 |
| Analysis Length (mi) | 0.8978 | 0.0970 | 85.6 | < 0.0001 | 0.7076 | 1.0879 |
| Total Truck Percentage | -0.0136 | 0.0042 | 10.5 | 0.0012 | -0.0218 | -0.0058 |
| Inverse Radius (mi) | -0.0085 | 0.0102 | 0.7 | 0.4049 | -0.0285 | 0.0115 |
| Ln(5 year Vehicle Count) | 0.6675 | 0.0486 | 188.5 | < 0.0001 | 0.5722 | 0.7627 |
| Dispersion | 0.8873 | 0.0963 | 84.8 | < 0.0001 | 0.6984 | 1.0761 |

### 5.3.5 Discussion

One of the observations during this process is that most of the tangent segments had significantly longer analysis lengths than the curved segments. This is to be expected since the length of each curved segment was determined by the PC and PT plus a buffer based on superelevation runoff and tangent runout. The segmentation for tangent stretches of roadway was not as limited. In fact, the tangent segments were identified based on sections of homogeneous facility features, such as speed limit, lane width, and rumble strip presence. In some regions, rural highways can go on for several miles without any disruptions in homogeneity. Since segment length is a major factor in vehicle exposure, which is a major factor in the number of crashes on a segment, it is possible that a model would place too much weight on segment length thus assigning an unusually high number of crashes to longer segments. While longer segments should have a higher number of crashes due to higher exposure values, if all tangent segments are longer than curved segments on average, the model may falsely assume that tangency is a strong indication of a high number of crashes.

Due to these observations, the interaction of curve presence and segment length was tested to see if the disparity in segment lengths was affecting the output of the different models. The results the analyses are shown in Tables 5-11 and 5-12.

Table 5-11: Interaction of Curve Presence and Segment Length for Three-year Sample

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -8.6258 | 0.8768 | 96.8 | <.0001 | -10.3443 | -6.9072 |
| Analysis Length (mi) | 0.6054 | 0.1046 | 33.5 | <.0001 | 0.4003 | 0.8104 |
| Total Truck Percentage | -0.0174 | 0.0051 | 11.9 | 0.0006 | -0.0274 | -0.0075 |
| ln_3yr_Veh_count | 0.6313 | 0.0576 | 120.3 | <.0001 | 0.5185 | 0.7441 |
| Curve Presence[1] | -0.9524 | 0.2180 | 19.1 | <.0001 | -1.3796 | -0.5251 |
| Analysis Length (mi) *Curve Presence[1] | 1.8256 | 0.5568 | 10.8 | 0.0010 | 0.7344 | 2.9168 |
| Dispersion | 0.8835 | 0.1247 | 50.2 | <.0001 | 0.6392 | 1.1279 |

Table 5-12: Interaction of Curve Presence and Segment Length for Three-year Sample

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -7.8338 | 0.7260 | 116.4 | <.0001 | -9.2567 | -6.4110 |
| Analysis Length (mi) | 0.5693 | 0.0956 | 35.4 | <.0001 | 0.3819 | 0.7568 |
| Total Truck Percentage | -0.0178 | 0.0042 | 17.9 | <.0001 | -0.0260 | -0.0095 |
| ln_5yr_Veh_count | 0.6252 | 0.0478 | 171.1 | <.0001 | 0.5315 | 0.7189 |
| Curve Presence[1] | -1.0243 | 0.1870 | 30.0 | <.0001 | -1.3909 | -0.6578 |
| Analysis Length (mi) *Curve Presence[1] | 1.7931 | 0.4734 | 14.3 | 0.0002 | 0.8653 | 2.7208 |
| Dispersion | 0.8016 | 0.0907 | 78.1 | <.0001 | 0.6238 | 0.9793 |

The p-value for the interaction term is very low for both the three- and five-year samples. This means that there is strong evidence of a relationship between the analysis length and the presence of a curve. In other words, the curved and tangent samples used in this analysis should not be combined into a single model due to the disparity in average segment lengths. Future research should attempt to set limits on segment lengths, or at least attempt to have samples of tangents and curves that have similar average lengths.

Overall, it is recommended that this idea of a combining curves and tangents in the same model receive further analysis and study. Specifically, a definition for what constitutes a curved segment should be defined based on curve radius and segment length. Using curve radius as a continuous variable that included both curved and tangent segments seems to be the most promising approach for incorporating horizontal alignment. Since this research is focused on curved segments, it was determined that these ideas should not receive further attention in this research. Thus, the modeling continued only for curved segments.

## 5.4 Utah-Specific Model for Curved Segments

This section will address the development of a crash prediction model specifically for rural two-lane two-way highways within the state of Utah. The results of the NB and EB models will be discussed and final models will be presented.

### 5.4.1 Negative Binomial Model

The development of an NB model took place using JMP, a statistical software package that is a graphical interface for SAS software. JMP can create an NB model with any number independent variables and interactions of variables. This research was able to use a backward stepwise technique for identifying which variables were significant and which were not. A backward stepwise technique involves adding as many variables as are available and then removing the variables one at a time based on their p-value—the variable with the highest p-value was eliminated and a new model was created for the remaining variables. This process was continued until all variables had p-values less than 0.05, based on a 95 percent confidence interval. The variables that remained after the backward stepwise technique was performed were segment length, AADT, total truck percentage, and curve radius. This was the case for both the three-year and five-year samples.

This process was performed on the first dataset, and the other two datasets were used for validation purposes. When the model used the data from the second set, it overpredicted the number of crashes by about 5.5 percent for the three-year sample and 13.8 percent for the five-year sample. For the third dataset the model underpredicted the number of crashes by about 6.7 percent for the three-year sample and 3.7 percent for five-year sample.

A chi-squared test for goodness of fit was performed on the second and third datasets for both the three-year and five-year samples. The critical value for the right-tailed chi-square distribution for the second dataset was 670.7. This number was calculated with a 95% probability and 612 degrees of freedom. The degrees of freedom are calculated by taking the difference of the data entries (615) and the estimated parameters (2), and then subtracting by 1. The chi-squared statistic for the three-year sample was 2217.4, and was 2360.4 for the five-year sample. Both of these values were greater than the critical value, indicating that there is no evidence that the distribution of the second dataset approximates a negative binomial distribution. The third dataset also was tested for goodness of fit. The critical value was found to be 663.3 for the third dataset, based on a 95% probability and 605 degrees of freedom. The chi-squared statistic was 1021.8 for the three-year sample and 1277.1 for the five-year sample.

Similar to the second dataset, these values were greater than the critical value, indicating a lack of evidence that the third dataset approximates a negative binomial distribution.

With this information, it was decided that the combined dataset would be used for the NB model. As shown in Figure 5-1, the randomly selected study segments are distributed across the state. The inherent problem is that with only one dataset, there is not a separate dataset for validation. For this reason, the combined dataset was randomly divided into a model set and a validation set, with 75 percent of the segments assigned to create the model and 25 percent of the segments assigned to validate the model. Thus, the modeling proceeded with a random sample of the combined random samples. The NB model takes the form shown in Equation 4-6, repeated in Equation 5-6 (Ramsey and Schafer 2002).

$$\ln(\mathrm{N}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (5\text{-}6)$$

where, $N$ = number of crashes (predicted or observed),

$\beta_0$ = intercept,

$\beta_i$ = coefficient for variable $x_i$,

$x_i$ = independent variable, and

$n$ = number of independent variables.

This is rearranged and shown in Equation 5-7, isolating the predicted number of crashes.

$$\mathrm{N} = \exp\left[\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right] \qquad (5\text{-}7)$$

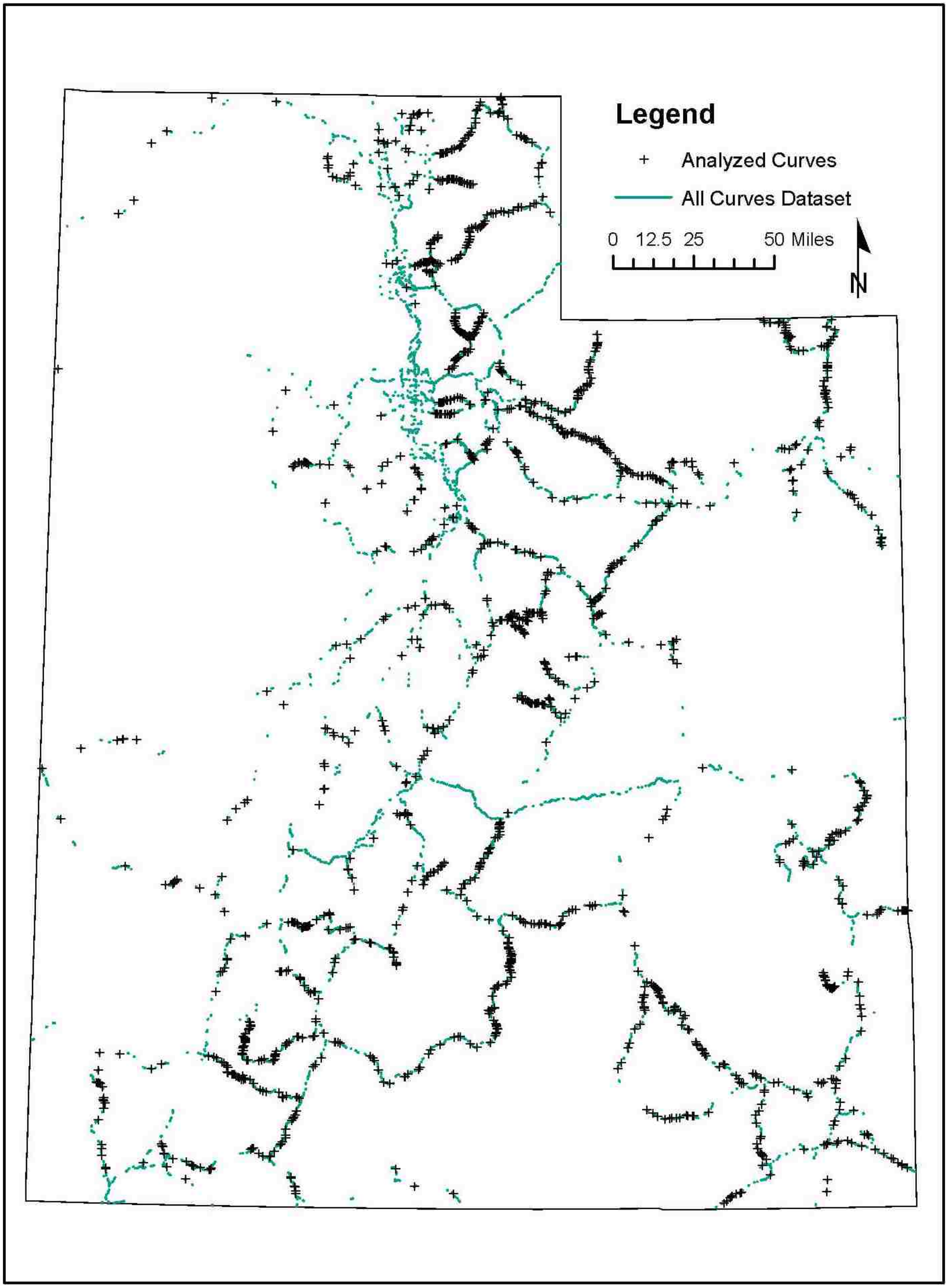


Figure 5-1: Combined Dataset of Curved Segments

The final three NB regression outputs from JMP using the backward stepwise technique are shown in Tables 5-13, 5-14, and 5-15 for the three-year sample. Tables 5-16, 5-17, and 5-18 show the outputs for the five-year sample.

Table 5-13: Third-from-Final Parameter Estimates for Three-year Sample

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -11.9486 | 0.8480 | 198.5 | <.0001 | -13.6106 | -10.2865 |
| Grade | 0.0236 | 0.0201 | 1.4 | 0.24 | -0.0158 | 0.0629 |
| Rounded Shoulder Width | -0.0201 | 0.0223 | 0.8 | 0.3675 | -0.0638 | 0.0236 |
| Analysis Length (mi) | 2.5030 | 0.4101 | 37.2 | <.0001 | 1.6992 | 3.3069 |
| ln 3yr Veh Count | 0.9009 | 0.0518 | 302.4 | <.0001 | 0.7994 | 1.0025 |
| Total Truck % | -0.0120 | 0.0046 | 6.6 | 0.0099 | -0.0211 | -0.0029 |
| Log Radius | -0.2105 | 0.0657 | 10.3 | 0.0014 | -0.3393 | -0.0817 |
| Dispersion | 0.641965 | 0.105336 | 37.142 | <.0001 | 0.435509 | 0.84842 |

Table 5-14: Second-from-Final Parameter Estimates for Three-year Sample

| **Term** | **Estimate** | **Standard Error** | **Wald $\chi^2$** | **Probability $> \chi^2$** | **Lower 95%** | **Upper 95%** |
|---|---|---|---|---|---|---|
| Intercept | -11.8085 | 0.8330 | 201.0 | <.0001 | -13.4411 | -10.1759 |
| Grade | 0.0230 | 0.0201 | 1.3 | 0.25 | -0.0165 | 0.0624 |
| Analysis Length (mi) | 2.4845 | 0.4104 | 36.7 | <.0001 | 1.6802 | 3.2888 |
| ln 3yr Veh Count | 0.8867 | 0.0493 | 323.9 | <.0001 | 0.7902 | 0.9833 |
| Total Truck % | -0.0120 | 0.0047 | 6.6 | 0.0101 | -0.0211 | -0.0028 |
| Log Radius | -0.2092 | 0.0658 | 10.1 | 0.0015 | -0.3381 | -0.0802 |
| Dispersion | 0.6472 | 0.1056 | 37.5 | <.0001 | 0.4402 | 0.8542 |

Table 5-15: Final Parameter Estimates for Three-year Sample

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -11.5570 | 0.8018 | 207.8 | <.0001 | -13.1284 | -9.9855 |
| Analysis Length (mi) | 2.4465 | 0.4089 | 35.8 | <.0001 | 1.6450 | 3.2480 |
| ln 3yr Veh Count | 0.8833 | 0.0491 | 323.0 | <.0001 | 0.7870 | 0.9796 |
| Total Truck % | -0.0127 | 0.0046 | 7.6 | 0.0059 | -0.0218 | -0.0037 |
| Log Radius | -0.2236 | 0.0647 | 11.9 | 0.0006 | -0.3505 | -0.0968 |
| Dispersion | 0.6491 | 0.1058 | 37.7 | <.0001 | 0.4418 | 0.8564 |

Table 5-16: Third-from-Final Parameter Estimates for Five-year Sample

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -12.2296 | 0.8091 | 228.5 | <.0001 | -13.8153 | -10.6439 |
| Rounded Shoulder Width | -0.0242 | 0.0192 | 1.6 | 0.2060 | -0.0618 | 0.0133 |
| Rounded Lane Width | 0.0963 | 0.0437 | 4.8 | 0.0277 | 0.0106 | 0.1820 |
| Analysis Length (mi) | 2.5204 | 0.3427 | 54.1 | <.0001 | 1.8487 | 3.1920 |
| ln 5yr Veh Count | 0.8570 | 0.0431 | 396.2 | <.0001 | 0.7726 | 0.9413 |
| Total Truck % | -0.0146 | 0.0038 | 14.7 | 0.0001 | -0.0220 | -0.0071 |
| Log Radius | -0.2061 | 0.0532 | 15.0 | 0.0001 | -0.3105 | -0.1018 |
| Dispersion | 0.554861 | 0.072403 | 58.72943 | <.0001 | 0.412953 | 0.696768 |

Table 5-17: Second-from-Final Parameter Estimates for Five-year Sample

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -11.9591 | 0.7800 | 235.1 | <.0001 | -13.4878 | -10.4303 |
| Rounded Lane Width | 0.0848 | 0.0430 | 3.9 | 0.0485 | 0.0006 | 0.1690 |
| Analysis Length (mi) | 2.5048 | 0.3440 | 53.0 | <.0001 | 1.8306 | 3.1790 |
| ln 5yr Veh Count | 0.8425 | 0.0415 | 413.0 | <.0001 | 0.7613 | 0.9238 |
| Total Truck % | -0.0147 | 0.0038 | 14.8 | 0.0001 | -0.0222 | -0.0072 |
| Log Radius | -0.2040 | 0.0533 | 14.6 | 0.0001 | -0.3085 | -0.0995 |
| Dispersion | 0.5620 | 0.0727 | 59.8 | <.0001 | 0.4196 | 0.7044 |

Table 5-18: Final Parameter Estimates for Five-year Sample

| Term | Estimate | Standard Error | Wald $\chi^2$ | Probability $> \chi^2$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | -11.2040 | 0.6794 | 271.9 | <.0001 | -12.5357 | -9.8723 |
| Analysis Length (mi) | 2.5753 | 0.3466 | 55.2 | <.0001 | 1.8960 | 3.2545 |
| ln 5yr Veh Count | 0.8606 | 0.0409 | 443.7 | <.0001 | 0.7805 | 0.9407 |
| Total Truck % | -0.0148 | 0.0038 | 15.0 | 0.0001 | -0.0223 | -0.0073 |
| Log Radius | -0.2082 | 0.0534 | 15.2 | <.0001 | -0.3129 | -0.1034 |
| Dispersion | 0.5755 | 0.0734 | 61.4 | <.0001 | 0.4316 | 0.7195 |

A Bayesian Information Criterion (BIC) was given for each model in the JMP output. BIC is a model selection statistic which measures the lack of fit of a model and adds a penalty for the number of terms in the model (Ramsey and Schafer 2002). When multiple models are available, the model with the smallest BIC is chosen. BIC is determined from Equation 5-8 (Ramsey and Schafer 2002).

$$BIC = n \times \ln(RSS) + p \times \ln(n) \tag{5-8}$$

where, $BIC$ = Bayesian information criterion,

$n$ = number of observations,

$RSS$ = sum of squared residuals, and

$p$ = number of independent variables.

The BIC for each of the final three models for both the three-year and five-year samples are presented in Table 5-19.

Table 5-19: BIC Comparison

| Output | Three-year Sample | Five-year Sample |
|---|---|---|
| Third-from-Final | 2129.8 | 2818.1 |
| Second-from-Final | 2123.6 | 2812.7 |
| Final | 2117.9 | 2809.5 |

The final outputs for both the three-year and five-year samples have the lowest BIC, in addition to having variables that are significant at a 95 percent confidence level. Models based on the final output of the backward stepwise technique are represented in Equations 5-9 and 5-10 for the three-year and five-year samples, respectively.

$$N_{3\text{-}year} = \exp[-11.5570 + (2.4465)(\mathrm{L}) + (0.8833)(\ln(\mathrm{VC})) - (0.0127)(\mathrm{TT}) - (0.2236)(\ln(\mathrm{R}))] \qquad (5\text{-}9)$$

$$N_{5\text{-}year} = \exp[-11.2040 + (2.5757)(\mathrm{L}) + (0.8606)(\ln(\mathrm{VC})) - (0.0148)(\mathrm{TT}) - (0.2082)(\ln(\mathrm{R}))] \qquad (5\text{-}10)$$

where, $L$ = length, mi;

$VC$ = vehicle count = (AADT)(365)(number of years in sample);

$TT$ = total truck percentage, percent; and

$R$ = radius, ft.

Simplifying the logarithms and coefficients yields Equations 5-11 and 5-12 for the three-year and five-year samples, respectively.

$$N_{3\text{-}year} = 483.8542 * \text{AADT}^{0.8833} * \text{R}^{-0.2236} * \exp[-11.5570 + (2.4465)(\text{L}) - (0.0127)(\text{TT})] \tag{5-11}$$

$$N_{5\text{-}year} = 640.6824 * \text{AADT}^{0.8606} * \text{R}^{-0.2082} * \exp[-11.2040 + (2.5757)(\text{L}) - (0.0148)(\text{TT})] \tag{5-12}$$

Unlike the HSM predictive model, these models are complete and do not rely on CMFs or any other modification to create a full model. The plots of actual total crashes versus predicted total crashes are shown in Figures 5-2 and 5-3 for three-year and five-year samples, respectively.

The sign for each coefficient shows the general effect of each variable. A positive coefficient means that the predicted number of crashes will increase as the value of the variable increases. A negative coefficient signifies a reduction in the predicted number of crashes as the value of the variable increases. For example, the predicted number of crashes increases as the AADT and segment length increases. This result is expected—more exposure should equate to a higher crash frequency. The predicted number of crashes decreases as the curve radius increases (becomes shallower). This is also expected as sharper curves are perceived as more dangerous. The predicted number of crashes decreases as the total truck percentage increases, which is the same result observed in the previous study performed on tangent segments of rural two-lane two-way highways in Utah (Saito et al. 2011). This could be explained by the fact that truck drivers receive training beyond the average automobile driver and generally have significantly more experience behind the wheel of a vehicle. The increased training and experience of professional truck drivers equate to lower crash frequencies on highway segments with increased truck traffic.

After the model development, both models were used on the validation dataset—the 25 percent of segments that were set aside from the combined dataset. A chi-squared test for goodness of fit was performed on the validation dataset for both the three-year and five-year samples. The critical value for the right-tailed chi-square distribution was 416.9. This number was calculated with a 95% probability and 371 degrees of freedom. The chi-squared statistic for the three-year sample was 370.0 and was 349.2 for the five-year sample. Both of these values were less than the critical value, signifying strong evidence that the distribution of the samples approximate a negative binomial distribution and that the validation dataset was an appropriate sample. The actual number of crashes in the three-year sample of the validation dataset was 204, and 396 crashes took place during the five-year sample. The model predicted 269 crashes for the three-year sample—an overprediction of 32 percent. Also, the model predicted 470 crashes for the five-year sample—an overprediction of 19 percent. Since the values for actual number of crashes are, in fact, actual data, it shows the reality of this type of modeling. Even when the predictions and the actual observed data do not always match, real conditions cannot be ignored. Therefore, these models represent every segment that was evaluated in the research and will stand as the results.

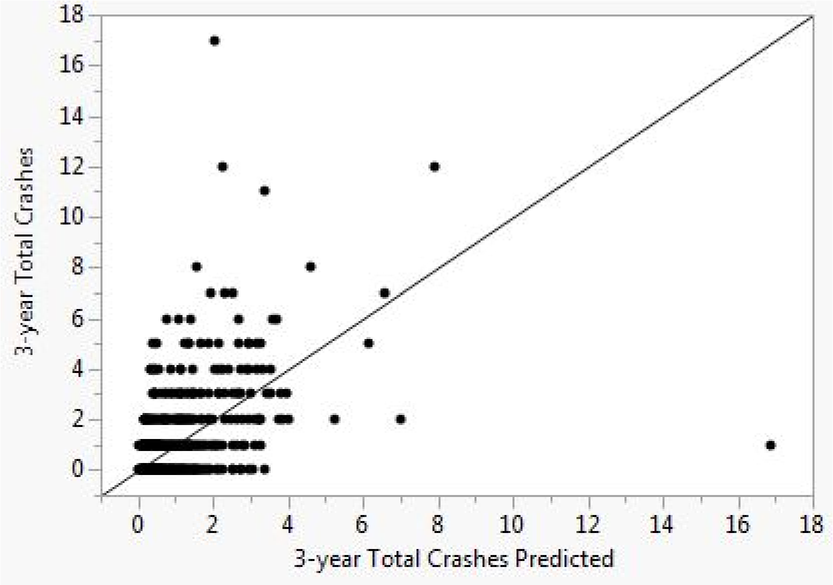


Figure 5-2: Actual vs. Predicted Three-Year Total Crashes

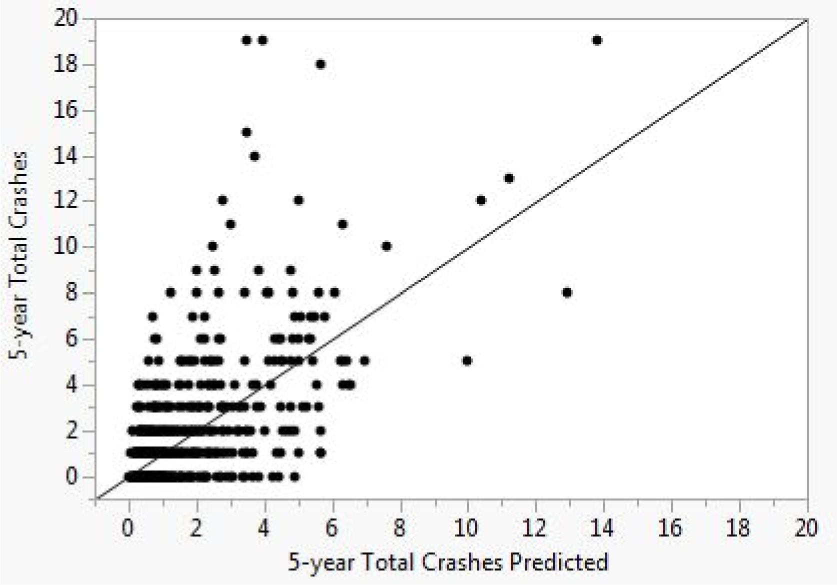


Figure 5-3: Actual vs. Predicted Five-Year Total Crashes

### 5.4.2 Empirical Bayes Model

The EB model was used to estimate the total number of crashes based on predicted values and on actual values. The EB model also uses a dispersion parameter to weight the predicted values. If the predicted values are over-dispersed, then the model gives more weight to the actual number of crashes in the calculations. The dispersion parameter for the three-year sample was found to be 0.6491 and the dispersion parameter for the five-year sample was found to be 0.5755. This means that the three-year sample data were more dispersed than the five-year sample data. The actual weights and models for the EB model vary for each segment. The only values that remain constant for each segment are the dispersion parameters.

The benefit of the EB model is that is more closely approximates the actual number of crashes compared to the NB model, because it incorporates the actual number of crashes into the model. For this reason, the output of the EB model is called the "expected" number of crashes rather than the "predicted" number. The EB model might be more appropriately considered a weighted average of the predicted and actual number of crashes. For example, the combined total of expected crashes on the validation dataset using the EB model was 235 for the three-year sample and 420 for the five-year sample. These numbers fall between the actual and predicted values for the validation dataset, 204 and 269 for the three-year sample and 396 and 470 for the five-year sample, respectively. While the EB model should only be used on a segment by segment basis, these total values illustrate the benefits of the EB model for more closely approximating the expected number of crashes based on actual and predicted values.

## 5.5 Summary of Results

This chapter has discussed the data collection efforts and the three different modeling procedures: HSM predictive model, curve and tangent combination, and the Utah-specific modeling efforts. The calibration factor for the HSM predictive model was found to be 1.50 for the three-year sample of the combined dataset, and 1.60 for the five-year sample of the combined dataset. These values were calculated from the SPF for base conditions, as well as all applicable CMFs. These CMFs included lane width, shoulder width, horizontal alignment, grade, driveway density, rumble strip presence, passing lane presence, and two-way left-turn lane presence. Looking at the decreased calibration factor between the three-year and five-year samples, and comparing it with the decreased calibration factors for the tangent segments over an eight-year period, indicates that overall safety is improving on rural two-lane two-way highways in Utah.

The curve and tangent combination attempt was partially successful in identifying a suitable parameter for variations in horizontal alignment. The simple indicator variable for curve presence was statistically significant at a 95 percent confidence level. However, this approach ignored all variations within the set of curved segments. The use of curve radius as a continuous variable was also statistically significant at a 95 percent confidence level. However, this approach assigned an arbitrary value of 10 miles for curve radius to all tangent segments. The most important observation was the result of testing the interaction between curve presence and segment length. The interaction showed strong evidence of a relationship between curve presence and segment length, signifying that the samples used in the combined models needed to have similar average segment lengths. The widely differing average segment lengths that were seen in the samples used for this research prevented the successful combination of curved and tangent segments into one model. Further research is required to identify the dividing point between a curve and a tangent based on curve radius.

The subsection on the development of Utah-specific models outlined the procedure for identifying significant variables within an NB model. A backward stepwise technique was used to remove insignificant variables. Originally the models were created from the first dataset and validated by the second and third datasets. For the second dataset, the model overpredicted the number of crashes by about 5.5 percent for the three-year sample and 13.8 percent for the five-year sample. For the third dataset, the model underpredicted the number of crashes by about 6.7 percent for the three-year sample and 3.7 percent for five-year sample. So all three datasets were combined into one dataset and 75 percent of the segments were randomly selected to make a new model. The remaining 25 percent were used to validate the new combined model. The model, created from 75 percent of the segments, overpredicted the number of crashes for validation dataset by 32 percent for the three-year sample and 19 percent for the five-year sample.

After the creation of the models, only four variables remained that were significant at a 95 percent confidence level: AADT (modeled as vehicle count, which was the product of AADT, days per year, and years in sample), segment length, curve radius, and total truck percentage. The four variables were significant in both the three-year model and the five-year model.

The Utah-specific NB crash prediction models can benefit from the application of the EB model. This model is used to better approximate the number of expected crashes. The EB model is essentially a weighted average between the predicted number of crashes and the actual number of crashes, using the dispersion factor to determine the weight. Thus, the expected number of crashes from the EB model is closer to the actual number of crashes than the predicted number of crashes from the NB model. A main benefit of the EB model is that it automatically corrects for the regression-to-the-mean effect. The EB model is appropriate for site-specific evaluation; thus EB model results for a combined dataset are not shown.

The main observation of the results is the importance of the four variables identified by the Utah-specific crash prediction models. AADT and segment length were always significantly associated with crash frequencies. More exposure equates to higher crash frequencies. Total truck percentage was found to be a significant variable and an increase in truck traffic is associated with lower crash frequency most likely due to the increased training and experience of professional truck drivers. Radius was indeed significantly associated with crash frequency. Smaller radii—sharper curves—are associated with higher crash frequency.

# 6 CONCLUSION

The purpose of this research was to use historical data to develop crash prediction models for curved segments of rural two-lane two-way highways in Utah. This thesis presents the methodology for developing crash prediction models and reports on the results of the accuracy of the crash prediction modeling effort. The modeling was accomplished by calibrating the HSM crash prediction model as well as by creating Utah-specific models. The data came from 2008-2012 datasets, grouped into a three-year sample from 2010-2012 and a five-year sample from 2008-2012. The HSM predictive model calibration followed the HSM predictive model, including the use of appropriate CMFs as described in the HSM (AASHTO 2010). The Utah-specific models were developed using an NB regression. An EB model was also used to compare the number of crashes predicted by the NB model with the actual number of crashes through weighted average equations.

The main finding of this research is that both the HSM and Utah-specific crash prediction models can incorporate highway curvature as a statistically significant variable. Out of a large list of possible variables, the Utah-specific models resulted in only four statistically significant variables at a 95 percent confidence level. This simplified crash prediction model will be easier to reproduce due to the small amount of data collection required for its use.

This chapter presents the outcomes of the research, recommendations for the use of models, and further research needs.

## 6.1 Outcomes

The calibration of the HSM predictive model for curved segments on rural two-lane two-way roads in Utah was completed for the three-year sample and the five-year sample for comparison. The combined dataset contained 1,495 curved segments throughout the state. The three-year sample had a calibration factor of 1.50 and the five-year sample had a calibration factor of 1.60. The HSM model is underpredicting the number of crashes (i.e. curved segments of Utah's rural two-lane two-way roads have 50 to 60 percent more crashes on average than a national dataset of rural two-lane two-way roads).

The Utah-specific models were developed using NB models. The use of a backward stepwise technique identified only four variables as statistically significant at a 95 percent confidence level. Those four variables were AADT, segment length, curve radius, and total truck percentage.

Where the HSM predictive model uses up to 12 variables in the CMF calculations in addition to the AADT and segment length values used in the SPF calculation, the Utah-specific models require only four variables in total. With the reduced data collection demands, these Utah-specific models may be better suited for crash prediction than the HSM predictive model.

This research also attempted to combine curved and tangent segments into one parameter. This attempt proved inconclusive as some of the models did not pass standard statistical tests for significance for tangent and curve section distinction since the coefficients of the possible variables were not significant at a 95 percent confidence level. The models that did have strong evidence of significance were not evaluated any further as they were either too general or needed further definition. Since the purpose of this research was to create crash prediction models for curved segments, the attempt to incorporate tangent segments did not receive any further attention.

An EB model was also used to determine an expected number of crashes. The EB model relies on a combination of predicted values and actual values. The two values are weighted and added together to provide the overall result. The weight is dependent on a dispersion parameter. These dispersion parameters were obtained during model development so that future analysis can be performed. EB models are meant to be site-specific; therefore, results from the combined datasets were not reported.

## 6.2 Recommended Models

The HSM calibration factors were found to be 1.50 and 1.60 for the three-year sample and the five-year sample, respectively. The Utah-specific crash prediction models can also be used as alternative models for curved segments of rural two-lane two-way highways. These are shown in Equations 5-11 and 5-12, which are repeated here in Equations 6-1 and 6-2, respectively.

$$N_{3\text{-}year} = 483.8542 * AADT^{(0.8833)} * R^{(-0.2236)} * \exp[-11.5570 + (2.4465)(L) - (0.0127)(TT)] \quad (6\text{-}1)$$

$$N_{5\text{-}year} = 640.6824 * AADT^{(0.8606)} * R^{(-0.2082)} * \exp[-11.2040 + (2.5757)(L) - (0.0148)(TT)] \quad (6\text{-}2)$$

where, $AADT$ = average annual daily traffic,

$R$ = radius (ft.),

$L$ = length (mi), and

$TT$ = total truck percentage (percent).

The Utah-specific models use far fewer variables than the HSM models and were developed from segments in Utah rather than across the United States. But simplicity has limitations: the Utah-specific models can only evaluate the effects of the four variables in the models—specifically, improvements on horizontal curvature. The HSM models require more variables, hence are able to evaluate the effects of as many variables as are included in the CMFs.

The EB models discussed in the previous chapter should be used in conjunction with the Utah-specific crash prediction NB models to correct for the regression-to-the-mean effect. Since the EB models use a weighted average of actual and predicted crashes, they are appropriate for before-after analysis where the actual number of crashes is known (Srinivasan et al. 2009).

## 6.3 Future Research Needs

As described earlier, this research was performed for curved segments of rural two-lane two-way highways in Utah. Research has been previously performed on tangent segments of rural two-lane two-way highways in Utah. This research attempted to find a suitable parameter for combining curved and tangent segments into the same model, but no convincing models resulted. Future research should give attention to this possibility. One hypothesis is that a mixed dataset of curved and tangent segments should have similar average segment lengths. Since segment length is a key component of vehicle exposure, which is a key variable in crash prediction modeling, longer average lengths for tangent segments may be falsely correlated with higher crash rates.

As expected, more recent data will help further research on this topic. A comparison of historical predictions versus current crash data would help in the development of more accurate crash prediction models. Also, interactions between variables were not considered in this research—each variable was considered independently from each other. Further research on interactions between variables would shed more light on improving the accuracy of crash prediction models.

# REFERENCES

American Association of State Highway and Transportation Officials (AASHTO). (2010). *Highway Safety Manual, Volume 2*. Washington, D.C.

American Association of State Highway and Transportation Officials (AASHTO). (2011). *A Policy on Geometric Design of Highways and Streets*, 6th ed. Washington, D.C.

Cook, A. A., Saito, M, and Schultz, G. G. (2015). "A Heuristic Approach for Identifying Horizontal Curves and Their Parameters Given LiDAR Point Cloud Data." *Compendium of Papers of the Transportation Research Board 94th Annual Meeting*. Transportation Research Board of the National Academies, Washington, D.C.

Easa, S. M. and You, Q. C. (2009). "Collision Prediction Models for Three-Dimensional Two-Lane Highways: Horizontal Curves." *Transportation Research Record: Journal of the Transportation Research Board,* 2092, 48-56.

Ellsworth P. (2013). "Utah DOT Leveraging LiDAR for Asset Management Leap."  Utah Department of Transportation. <https://www.udot.utah.gov/public/ucon/uconowner.gf?n=8336606666333974> (November 21, 2013).

Environmental Systems Research Institute (ESRI). (2012). "ArcGIS Desktop: Release 10.1," ESRI, Redlands, CA.

Federal Highway Administration (FHWA). (2013). *Traffic Monitoring Guide*. U.S. Department of Transportation, Washington, D.C.

Federal Highway Administration (FHWA). (2014). *Highway Performance Monitoring System Field Manual*. U.S. Department of Transportation, Washington, D.C.

Findley, D. J. (2011). *A Comprehensive Two-Lane, Rural Road Horizontal Curve Study Procedure*. PhD Dissertation, North Carolina State University. <http://repository.lib.ncsu.edu/ir/bitstream/1840.16/7202/1/etd.pdf>. (May 23, 2014).

Findley, D., Zegeer, C., Sundstrom, C., Hummer, J., and Rasdorf, W. (2012). "Applying the Highway Safety Manual to Two-lane Road Curves." *Journal of the Transportation Research Forum,* 51(3), 25-38.

Findley, D., Hummer, J., Rasdorf, W., and Laton, B. (2013). "Collecting Horizontal Curve Data: Mobile Asset Vehicles and Other Techniques." *Journal of Infrastructure Systems*, 19(1), 74–84.

Fitzpatrick, K., Lord, D., and Park, B. (2008). "Accident Modification Factors for Medians on Freeways and Multilane Highways." *Transportation Research Record: Journal of the Transportation Research Board*. 2083, 62-71.

Fitzpatrick, K., Lord, D., and Park, B. (2010). "Horizontal Curve Accident Modification Factor with Consideration of Driveway Density on Rural Four-Lane Highways in Texas." *Journal of Transportation Engineering*, 136(9), 827–835.

Google Earth (Google). (2014). <http://www.google.com/earth/index.html> (May 5, 2014)

Gross, F., Persaud, B., and Lyon, C. (2010). A Guide to Developing Quality Crash Modification Factors, FHWA-SA-10-032. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

Hauer, E. (1997). *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon, Oxford UK.

Hauer, E. (1999). *Safety in Geometric Design Standards*. University of Toronto, Toronto.

Khan, G., Chitturi, M. V., Bill, A. R., and Noyce, D. A. (2012). "Horizontal Curves, Signs, and Safety." *Transportation Research Record: Journal of the Transportation Research Board*, 2279, 124-131.

Labi, S. (2006). *Effects of Geometric Characteristics of Rural Two-Lane Roads on Safety*, FHWA/IN/JTRP-2005/02. Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, IN.

Lord, D. and Persaud, B. (2004). "Estimating the Safety Performance of Urban Road Transportation Networks." *Accident Analysis & Prevention*, 36(4), 609-620.

Lord, D., Kuo, P., and Geedipally, S. R. (2010) "Comparison of Application of Product of Baseline Models and Accident-Modification Factors and Models with Covariates: Predicted Mean Values and Variance." *Transportation Research Record: Journal of the Transportation Research Board,* 2147, 113-122.

National Highway Traffic Safety Administration (NHTSA). (2013). Fatality Rates: Utah, U.S. and Best State. U.S. Department of Transportation, Washington, D.C. < http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/STSI/49_UT/2012/49_UT_2012.htm> (November 10, 2014).

Pradhan, A. and Rasdorf, W. (2009). "GIS and LiDAR Use for Identification of Potential Road Hazard Locations." *Journal of Computing in Civil Engineering*, 125-134.

Ramsey, F., and Schafer, D. (2002). *The Statistical Sleuth.* Duxbury, Pacific Grove, CA.

Rasdorf, W., Cai, H., Tilley, C., Brun, S., and Robson, F. (2004). "Accuracy Assessment of Interstate Highway Length Using Digital Elevation Model." *Journal of Surveying Engineering*, 130(3), 142–150.

Saito, M., Schultz, G. G., 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 & Safety, Research Divisions, Salt Lake City, UT.

SAS Institute, Inc. (SAS). (2013). JMP® Pro 11.2.0. SAS, Cary, NC.

Srinivasan, R., Baek, J., Carter, D., Persaud, B., Lyon, C., Eccles, K., Gross, F., and Lefler, N. (2009). Safety Evaluation of Improved Curve Delineation. FHWA-HRT-09-045. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

Utah Department of Transportation (UDOT). (2008). "Resurfacing, Restoration, and Rehabilitation (3R) Standards for Non-Freeway Systems." <https://www.udot.utah.gov/main/uconowner.gf?n=13756531154159072> (May 5, 2014).

Utah Department of Transportation (UDOT). (2011). Traffic on Utah Highways (AADT). <http://www.udot.utah.gov/main/f?p=100:pg:0::::V,T:,529> (May 5, 2014).

Utah Department of Transportation (UDOT). (2012). "Utah Department of Transportation "Roadview Explorer" Website." <http://www.roadview.udot.utah.gov> (May 5, 2014).

Utah Department of Transportation (UDOT). (2013). Crash Statistics. <http://www.udot.utah.gov/main/f?p=100:pg:::::V,T:,580> (May 5, 2014).

Utah Department of Transportation (UDOT). (2014a). Open Data Guide. <http://udot.uplan.opendata.arcgis.com/> (May 5, 2014).

Utah Department of Transportation (UDOT). (2014b). Traffic Studies. <http://www.udot.utah.gov/main/f?p=100:pg:0:::1:T,V:258,> (May 5, 2014).

Xie, F., Gladhill K., Dixon K. K., and Monsere, C. M. (2011) "Calibrating the Highway Safety Manual Predictive Models for Oregon State Highways." *Proceedings of the Transportation Research Board 90th Annual Meeting*. Transportation Research Board. National Research Council, Washington, D.C.

Zegeer, C. V., Steward, J. R., Council, F. M., Reinfurt, D. W., and Hamilton, E. (1992). "Safety Effects of Geometric Improvements on Horizontal Curves." *Transportation Research Record: Journal of the Transportation Research Board*, 1356, 11–19.