CRASH PREDICTION MODELING FOR CURVED SEGMENTS OF RURAL TWO-LANE TWO-WAY HIGHWAYS IN UTAH

Prepared For:

Utah Department of Transportation
Traffic & Safety, Research Divisions

Submitted By:

Brigham Young University
Dept. of Civil & Environmental Engineering

Authored By:

Mitsuru Saito, Ph.D., P.E.
Professor

Casey S. Knecht, EIT
Graduate Research Assistant

Grant G. Schultz, Ph.D., P.E., PTOE
Associate Professor

Aaron A. Cook, EIT
Undergraduate Research Assistant

Final Report
October 2015

RESEARCH

Utah Department of Transportation - Research Division
4501 South 2700 West - P.O. Box 148410 - SLC,UT 84114-8410

UDOT
Keeping Utah Moving

DISCLAIMER

The authors alone are responsible for the preparation and accuracy of the information, data, analysis, discussions, recommendations, and conclusions presented herein. The contents do not necessarily reflect the views, opinions, endorsements, or policies of the Utah Department of Transportation or the U.S. Department of Transportation. The Utah Department of Transportation makes no representation or warranty of any kind, and assumes no liability therefore.

This report is protected under 23 USC 409.

ACKNOWLEDGMENTS

TECHNICAL REPORT ABSTRACT

1. Report No. UT-15.12	2. Government Accession No. N/A	3. Recipient's Catalog No. N/A
4. Title and Subtitle Crash Prediction Modeling for Curved Segments of Rural Two-Lane Two-Way Highways in Utah		5. Report Date October 2015
		6. Performing Organization Code R0402229
7. Author(s) Mitsuru Saito, Ph.D., P.E.; Casey S. Knecht, EIT; Grant G. Schultz, Ph.D., P.E., PTOE; Aaron A. Cook, EIT		8. Performing Organization Report No.
9. Performing Organization Name and Address Brigham Young University Department of Civil and Environmental Engineering 368 Clyde Building Provo, UT 84602		10. Work Unit No.
		11. Contract or Grant No. 14-8287
12. Sponsoring Agency Name and Address Utah Department of Transportation 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410		13. Type of Report & Period Covered Final Report October 2013 to October 2015
		14. Sponsoring Agency Code PIN No. 10304
15. Supplementary Notes Prepared in cooperation with the Utah Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration		

16. Abstract

This report 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 using negative binomial regression. The data for these models came from 1,495 randomly sampled curved segments in Utah, with crash data coming from years 2008-2012. 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. A negative binomial model was used to develop Utah-specific crash prediction models based on both the three-year and five-year sample periods. 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 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.

17. Key Words Crash prediction, Transportation safety, Curve, Safety performance function (SPF), Crash modification factor (CMF), Utah Crash Prediction Model (UCPM), Highway Safety Manual (HSM), Negative binomial, Empirical Bayes, Horizontal curvature	18. Distribution Statement Not restricted. Available through: UDOT Research Division 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410 www.udot.utah.gov/go/research	23. Registrant's Seal N/A
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 83 22. Price N/A

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AADT   Annual Average Daily Traffic

AASHTO   American Association of State Highway and Transportation Officials

BYU   Brigham Young University

CMF   Crash Modification Factor

DEM   Digital Elevation Model

EB   Empirical Bayes

FHWA   Federal Highway Administration

GIS   Geographic Information System

GPS   Global Positioning System

HAF   Horizontal Alignment Finder

HPMS   Highway Performance Monitoring System

HSM   Highway Safety Manual

LiDAR   Light Detection and Ranging

mph   Miles per Hour

NB   Negative Binomial

PC   Point of Curvature

PT   Point of Tangency

RTM   Regression to the Mean

SPF   Safety Performance Function

TWLTL   Two-way Left Turn Lane

UCPM   Utah Crash Prediction Model

UDOT   Utah Department of Transportation

VMT   Vehicle Miles of Travel

vpd   Vehicle per Day

Highway safety is a top priority for everyone. Understanding and being able to identify crash causes and potential mitigations is paramount. Fatalities from crashes on U.S. highways are far too common. In 2012, there were 33,561 crash-related fatalities in the U.S., 217 of which were in Utah. 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.

This report 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 calibration of the predictive model found in the Highway Safety Manual (HSM) as well as development of Utah-specific models developed using negative binomial regression. Data for these models came from 1,495 randomly sampled curved segments in Utah, with crash data coming from years 2008-2012.

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 HSM predictive model calibration factor was determined to be 1.50 for the three-year period and 1.60 for the five-year period. These factors should 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 (AADT), 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 data collection effort than the HSM predictive model.

1.0 INTRODUCTION

1.1 Problem Statement

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. It could be argued that highway safety affects everyone and the U.S. economy as a whole because crashes create delays and societal costs that impact more than just the people directly involved in crashes. Understanding and being able to identify crash causes and potential mitigations is paramount. To do so, researchers have developed crash prediction models based on historical crash data to estimate the number of future crashes under prevailing conditions. Such models can be used to evaluate the contributions of physical attributes to crash occurrence and severity.

Safety performance functions (SPFs) provide one avenue for crash prediction modeling. 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 highway system 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 the research data were not available for horizontal curvature. Since that research, UDOT performed an inventory of all highway curvature as part of its Light Detection and Ranging (LiDAR) asset management program. With this additional data, UDOT desires to calibrate the HSM predictive model specifically for curved segments of two-lane two-way rural highways. Similarly, UDOT has requested the creation of Utah-specific crash prediction models for two-lane two-way highways exclusive of the HSM predictive model.

1.2 Objectives

The research objective is to develop crash prediction models for curved segments of rural two-lane two-way highways in Utah using historical crash data and roadway attribute data recently collected as part of UDOT's LiDAR asset management program.

1.3 Scope

This research analyzes curved segments of rural two-way two-lane highways managed by UDOT. Ancillary efforts necessary for this analysis are also performed, including development of an algorithm to identify and categorize curved segments given the point cloud data provided by UDOT's LiDAR asset management program.

1.4 Purpose and Need

The purpose of this research is to provide UDOT engineers with crash prediction models for curved segments of rural two-lane two-way highways. These models use historical crash data and roadway attribute data recently collected as part of UDOT's LiDAR asset management program. This is 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 are 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 occurrence and severity. The models identify which factors play the largest role in crash prediction. With this information, UDOT can focus its efforts on 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 government agencies 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., 217 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). Lack of funding is most agencies' main obstacle to safety improvements. Many projects and improvements are shelved or delayed due to insufficient funds. Crash prediction modeling helps agencies focus on the most cost-effective measures to improve highway safety.

1.5 Outline of Report

The executive summary presented an overview of this report and its findings. This chapter presented an overview of the report along with a stated purpose and need for this research. Chapter 3 presents a literature review of topics related to this research. Chapter 4 discusses the data preparation necessary for accurate and complete modeling and Chapter 5 addresses the methodology for crash prediction modeling. Chapter 6 presents and evaluates modeling effort results, followed by Chapter 7, which contains conclusions, recommended use of models, and further research needs.

2.0 LITERATURE REVIEW

2.1 Overview

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, Crash Modification Factors (CMFs), calibration factors, and statistical methods.

2.2 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 has become widely available and 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 image capture that exemplifies the accuracy level of LiDAR compared to the image captured by Roadview Explorer, which is a high-resolution photolog database of all UDOT roads.

Many transportation agencies are utilizing mobile vehicles to collect a wide variety of asset data (Findley et al. 2013). In 2011, UDOT commenced a LiDAR project that would eventually collect highway infrastructure data for every state road in Utah. The data have an average accuracy of +/-3 centimeters (Ellsworth 2013).
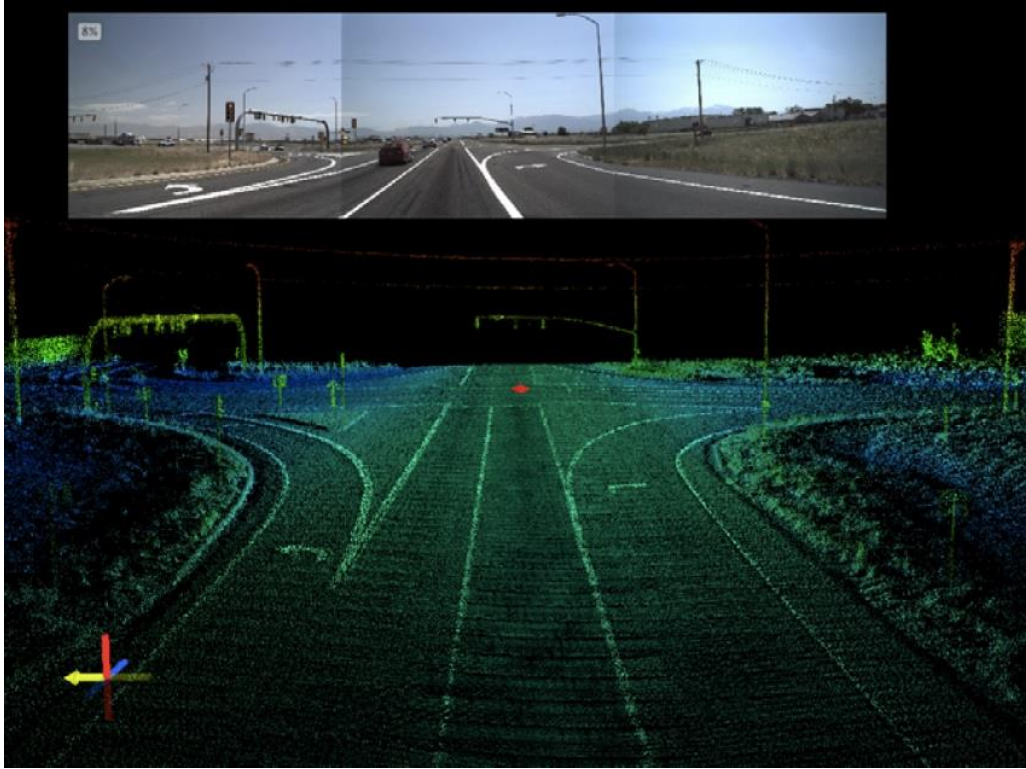
Figure 2.1 LiDAR Capture (Ellsworth 2013)

As technology and machinery become more sophisticated, accuracy naturally improves. 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).

2.3 Highway Curvature

Highway curvature was an important focus of 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. Despite safety design measures, approximately 25 percent of all fatal crashes in the U.S. 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 combine both. Sharpness is the most distinguishing factor of horizontal curves. Sharpness is really a measure of radius or curvature. Although there are more crashes on sharper curves, sharpness is not the only factor. Narrower curve width, lack of spiral transitions, and 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 degrees was a good break point for distinguishing segments that behaved like straight sections versus curved sections. Segments with less than 3.45 degree 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 numbers of curve-related crashes. Ranges of crash reductions for horizontal curve improvements were determined for flattening curves, widening lanes, widening paved shoulders, adding unpaved shoulders, adding spiral transitions, and improving superelevation (Zegeer et al. 1992).

Several factors 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.

Some factors have been shown to have low impact on safety. For highway segments with a degree of curvature greater than 3.45 degrees, the use of advisory signs is not a significant factor (Khan et al. 2012). Hauer (1999) discarded lane width as a factor, specifically 11-foot versus 12-foot lanes. Conventional wisdom says that a larger lane would be safer, but Hauer's research showed more crashes in the 12-foot lanes than the 11-foot lanes.

Crash modeling involves various factors. The most important of these is the human factor, which is very difficult to quantify and describe. The best remedy is collecting more data and creating jurisdiction-specific calibrations of crash prediction models.

2.4 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 with documented calibrations (Fitzpatrick et al. 2008).

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

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

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. 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 to calibrate 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 dataset and making local adjustments for items such as sample sizes for underrepresented facility types. The target number of 100 crashes per year could not be achieved at low-volume intersections. 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 obtain. In certain regions, the locally maintained roads carried more traffic than the state highways. A model was created to estimate AADT on rural highways in order to correct for this problem. 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, 6-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 CMFs, which are multiplied by 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.

2.5 Crash Modification Factors (CMFs)

CMFs represent the relative change in crash frequency due to a change in one specific condition such as geometric design or traffic control (AASHTO 2010). Transportation safety analysts use CMFs because they allow base-line models to be recalibrated for different jurisdictions and conditions.

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 the 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). 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, CMF calibration 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 poor pavement (Labi 2006). Drivers may increase their speed in response to the perception of increased safety, 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.

2.6 Calibration Factors

The HSM predictive model multiplies the SPF crash frequency prediction by a calibration factor to account for differences in the jurisdiction and time period being analyzed versus the data on which the model is based (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 including 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.

2.7 Statistical Models

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, which is the assumption that the crash distribution mean and variance 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 (RTM). The HSM includes the EB model, and has thus established it as the standard method for road safety analysis (AASHTO 2010, Labi 2006).

2.8 Literature Review Summary

Using LiDAR data to account for highway geometry and conditions is a new approach that will be of benefit to transportation agencies. 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 to distinguish curved segments from straight segments.

The HSM model 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.

3.0 DATA PREPARATION

3.1 Overview

The first step of data collection was to randomly select segments representative of rural two-way two-lane highways in Utah. The next step was to gather sufficient data about segment attributes that could affect the predictive power of the crash prediction models (i.e., variables in the HSM predictive models). While the segments did not need to meet the base conditions, attribute identification was necessary to determine the value of an appropriate CMF.

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

3.2 Scope of Data Collection

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 radius, point of curvature (PC) and point of tangency (PT).

3.3 Horizontal Alignment Finder

The HSM crash prediction models consist of a set of variables describing the conditions of the segments selected. Curve radius is one of the most important variables for analyzing rural two-lane two-way highways but it has historically been difficult to collect. 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 is available through UDOT's LiDAR program, it is possible to analyze curved segments in addition to tangent segments. Curved segments have more variation in their attributes, classifications, and accuracy (Findley 2011).

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 curve attributes. The major issue was that the LiDAR dataset segmented more than half of all curves 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 crash analysis results. The HAF algorithm was developed to combine the LiDAR dataset's erroneous curve segmentation into segments more closely in line with true conditions. This method proved to be a crucial ancillary effort to this project.

The HAF combines curve segments at 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 LiDAR dataset 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. 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 that 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 they are closely connected by the mechanism the HAF algorithm uses to associate 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, three constituent segments of a curve would be assigned the same number. This is done by comparing each segment to the segment immediately prior and assigning them the same curve number if they are sufficiently similar. In the combination step, all segments with the same curve number are grouped into one larger segment and their attributes are combined in various ways as discussed below. These two steps are presented visually in Figure 3.2.

14
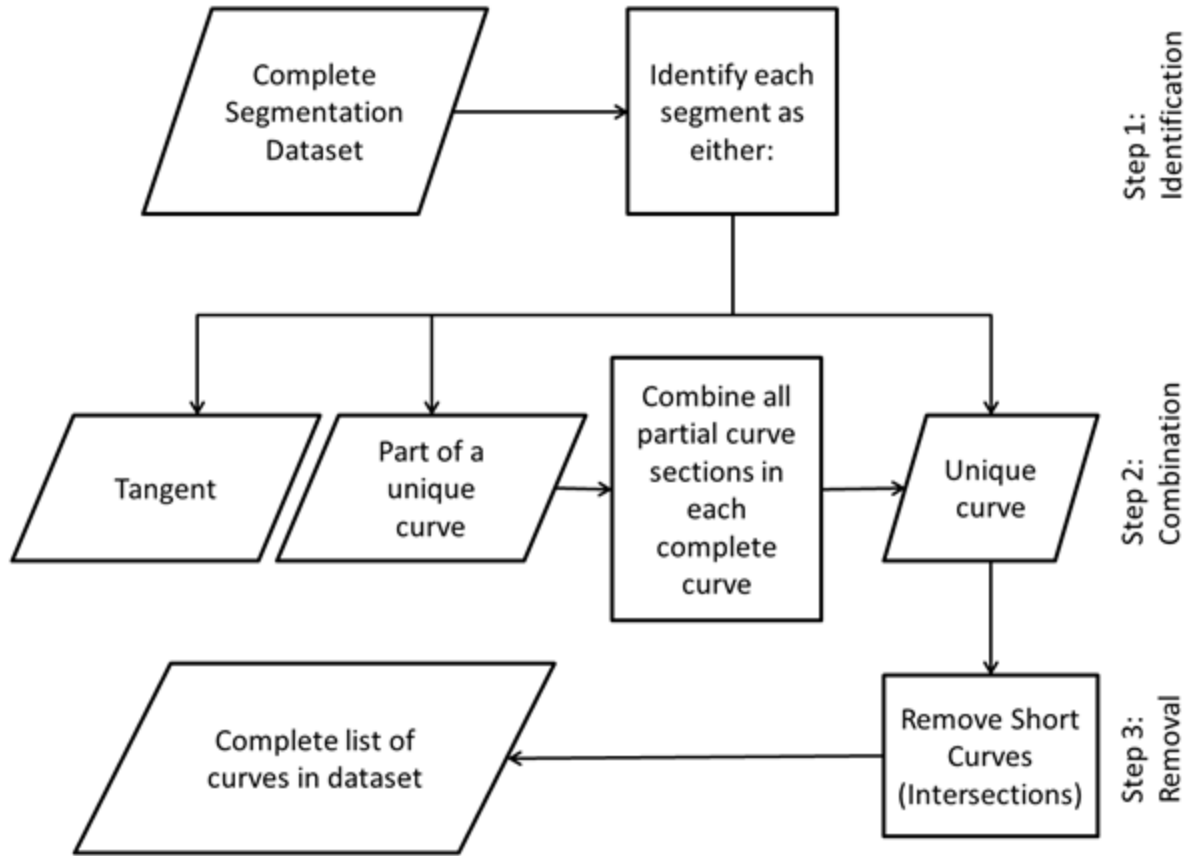
Figure 3.1 The HAF Algorithm Segmentation Steps

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 flowchart of the segment combination process is shown in Figure 3.3.

Four parameters were used to determine if each segment should be classified as a partial curve, a unique curve, or a tangent. The parameters included 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 step. 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 (arbitrarily selected for convenience), including 0. The weighting scale classifies each section as a curve or tangent.

A positive weight classifies the segment as a curve, and a negative classifies it as a tangent. A 0 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 0 weight. The HAF algorithm merely uses the 0 weight as a temporary measure to force the segment to be analyzed without the previous segment's attributes. The segment is then assigned a new number if it is determined to be a curve.

The HAF analysis successfully identified curves approximately 90 percent of the time. It correctly placed the PC and PT at a range between 79 to 90 percent of the time. Although these results are not perfect, they are as good as or better than any other available method. Curve selection randomization and curve buffering to include superelevation runoff and tangent runout likely mitigated the margin of error in most of the problematic curves.

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
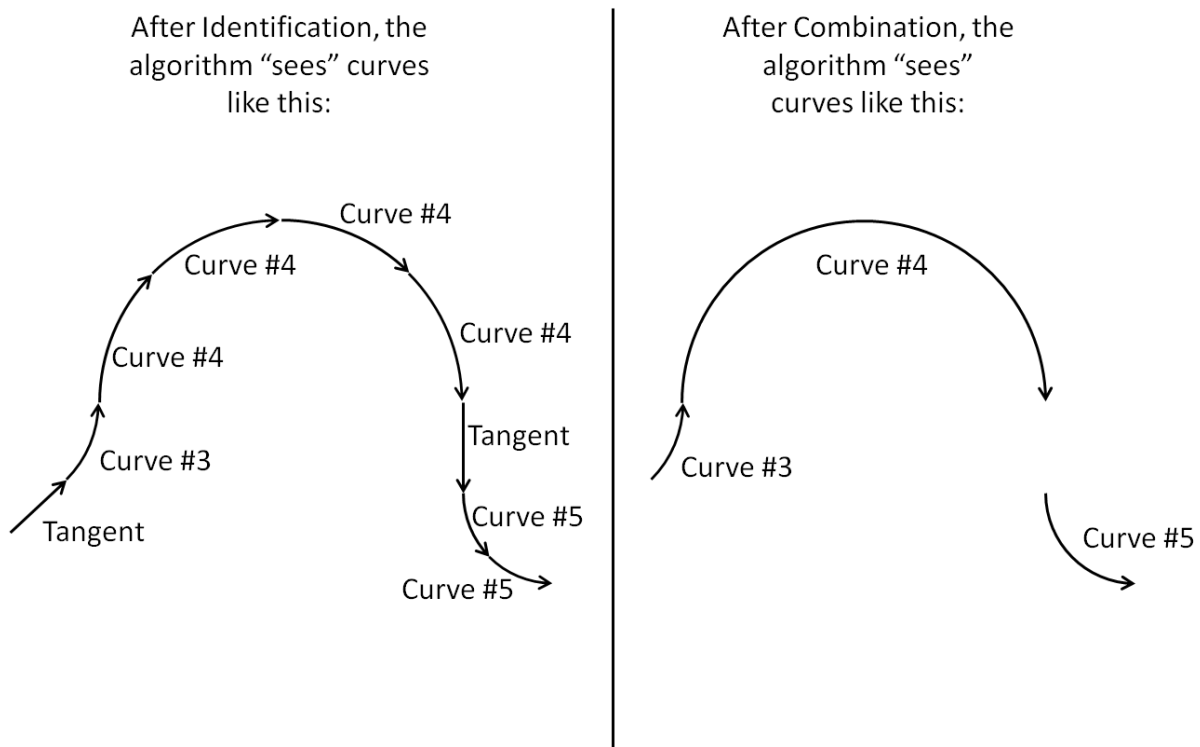
Figure 3.2 Representation of Identification and Combination of the HAF Algorithm

16
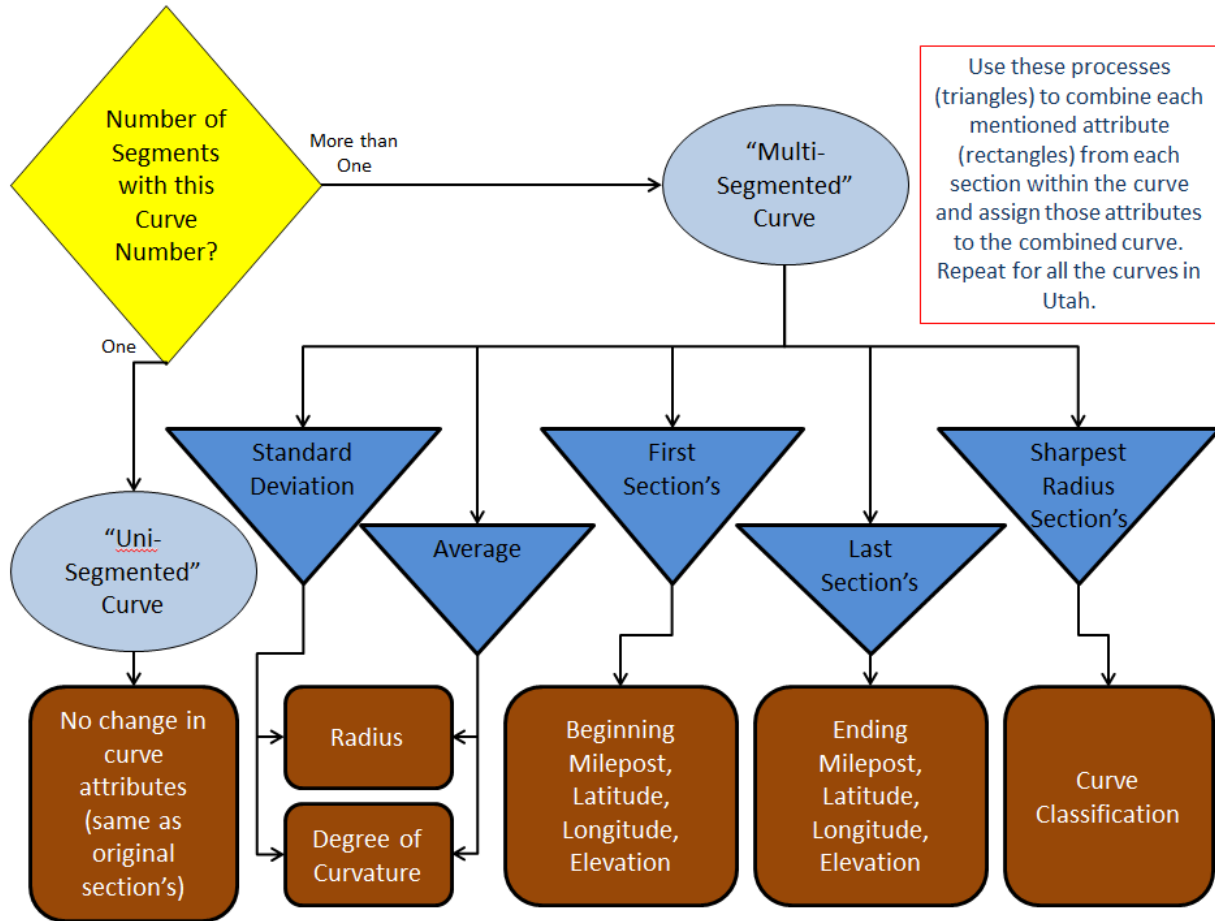
Figure 3.3 Combination Schema for Each Segment

3.4 Resources

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 that allowed for data verification. The resources used in the data collection process included Google Earth, Roadview Explorer, the UDOT Data Portal, and the UDOT Crash Database.

3.4.1 Google Earth

Google Earth (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 are discussed later in this chapter.

3.4.2 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 collected by a vehicle mounted with cameras that drove every state 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.

3.4.3 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 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, which allowed attributes to be tied to the segments.

3.4.4 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.

3.5 Data Limitations

Many crashes in the database are located just before the start of a curve or just past the end of a curve. This could indicate that the entrance or exit of a curve is dangerous in its own right. However, law enforcement reporting often lacks crash location accuracy. Some site investigators use a portable measuring wheel to measure from the nearest milepost. Others use a GPS receiver to pinpoint the site of the crash. 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, the crash may be recorded at the parking location. Superelevation runoff and tangent runout lengths were manually added to both ends of every curve regardless of whether those elements were present on the road. The purpose of doing so was to address possible reporting inconsistencies by allowing crashes to be tied to curve locations where they might otherwise show up in the data a short distance away from curves. The calculations for these lengths are found in Section 3.7.

3.6 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 or residential areas were removed. Segments with speed limits lower than 30 miles per hour (mph) were removed. The reasoning is that 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.
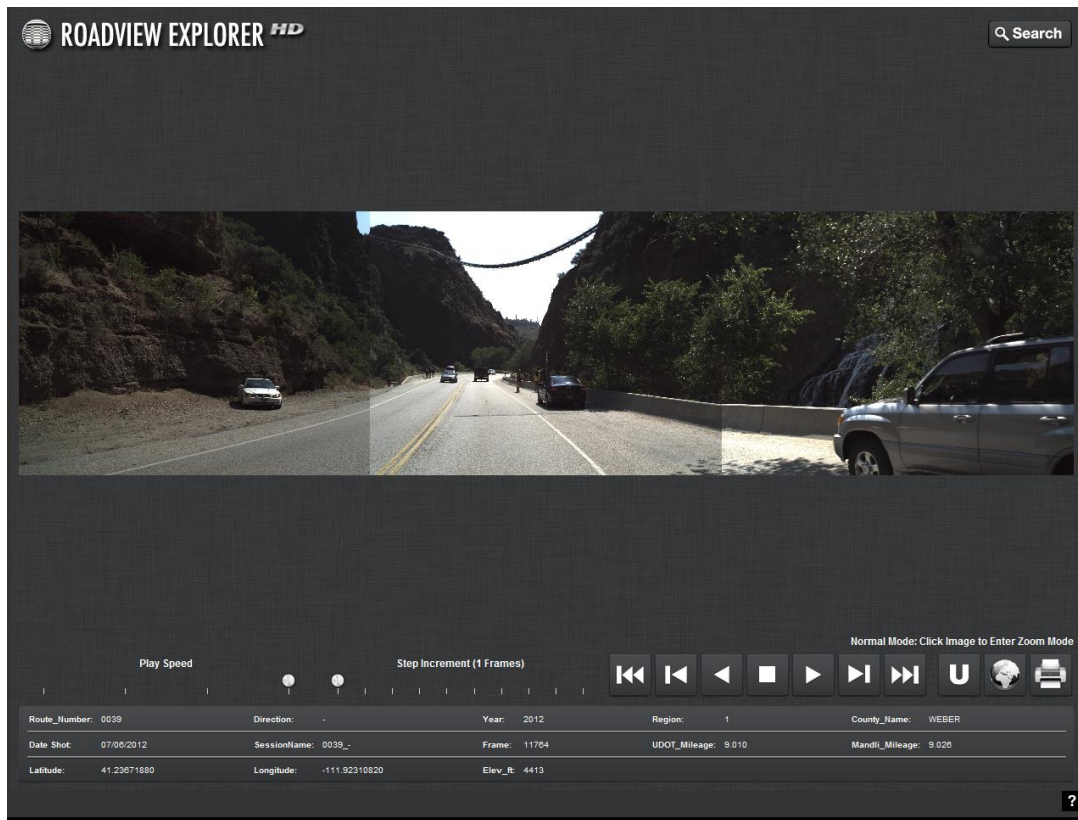
Figure 3.4 Example of Segment with High Pedestrian Traffic (UDOT 2012)

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 degree or near-90 degree 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 traffic flow 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 tollbooths, information booths, boat launches, ranger stations, and recreation vehicle dump stations. These on-road

20

services prevent free-flow operation, and therefore the segments that contained or were near any of these services were removed from the dataset.

3.7 Facility Data

The facility data that were collected include curve radius, degree, 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.

3.7.1 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

3.7.2 Curve Buffer

As discussed previously, it is possible that not all curve-related crashes are recorded between the PC and PT. This could be due to inaccurate recording or simply because the crash occurred just before entering or just after exiting a curve. The full length of superelevation

runoff and tangent runout was added to the length of the curve on both ends as a buffer to compensate for possible inaccuracies and include crashes in the immediate vicinity of the curve.

Superelevation runoff is the length required to transition the cross slope of a road from 0 percent on the outer lane (the location of adverse crown) to a superelevated position on a curve. A cross slope of 0 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).

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

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

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

$\quad\quad\quad\quad\quad\; n_1 \;=\;$ number of lanes rotated;

$\quad\quad\quad\quad\quad\; e_d \;=\;$ design superelevation rate, percent; and

$\quad\quad\quad\quad\quad\; \Delta \;=\;$ maximum relative gradient, percent.

A value of 6 percent design superelevation was assumed due to lack of knowledge about actual conditions. This is UDOT's maximum design superelevation rate (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 0 percent. Equation 3-2 shows how to calculate tangent runout (AASHTO 2011). For this study, a normal cross slope rate of 2 percent was assumed.

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

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

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.

3.7.3 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. Elevation difference was divided by segment length, 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.

3.7.4 Speed Limit

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

3.7.5 Rumble Strip Presence

The rumble strip presence data available from UDOT proved unreliable. It was used as a base for analysis but 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.

3.7.6 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. Nevertheless, 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.
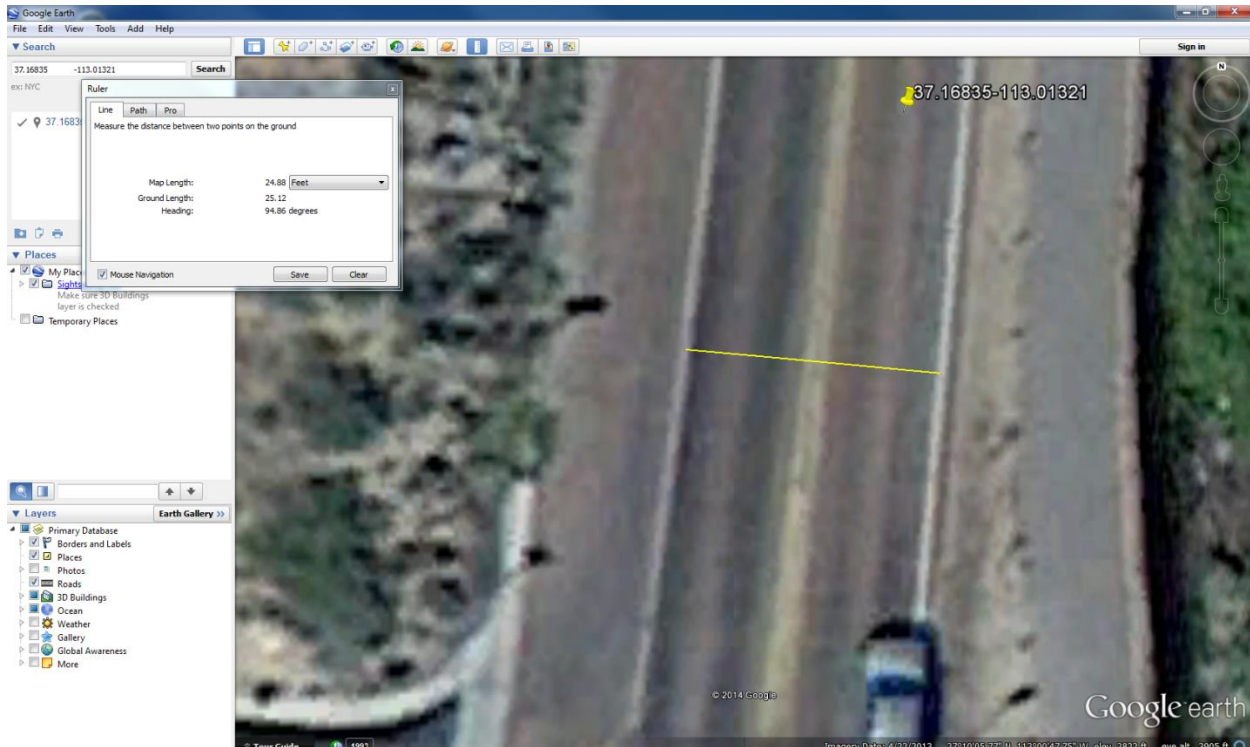
Figure 3.5 Lane Width Measurement in Google Earth

3.7.7 Shoulder Width

Shoulder width data were obtained from UDOT (2014a). However, they were verified on Google Earth (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.

3.7.8 Driveway Density

Driveways were counted using Roadview Explorer (UDOT 2012) and Google Earth (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.

3.7.9 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: 0, 1, or 2. 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.

3.7.10 Lighting

The presence of overhead street lighting was observed for each segment. This was done in Roadview Explorer. The lighting did not need to be present for the majority of the segment like the other geometric attributes—one light would suffice. However, no study segments had overhead lighting. This is most likely due to the rural location of each segment.

3.7.11 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 at the time of the research.

3.7.12 Truck Percentage

Truck percentage data are divided into single-unit (single) and combination unit (combo). 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.

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

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.

Table 3.4 continued

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 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.

3.8 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 buffers were added, the segment data were cross-referenced with the crash data to extract only the crashes that occurred on the segments in the dataset.

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.

3.9 Data Preparation Summary

The purpose of data preparation was to randomly select segments that represented a cross section of rural two-way two-lane highway curves in Utah. 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. Variables required for HSM model calibration, along with the additional variables, were used for modeling and further analysis.

4.0 METHODOLOGY

4.1 Overview

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.

4.2 HSM Predictive Model

The HSM 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.

4.2.1 SPF

The HSM outlines the 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 in this section 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, $Nspf$ = 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, 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 used 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 U.S. (AASHTO 2010).

<u>4.2.2 CMFs</u>

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.

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%

When a segment does not meet one of the 13 base conditions listed in Table 4.1, a CMF must be multiplied by the predicted number of crashes calculated by the SPF model. Doing so adjusts the prediction by incorporating more of the actual parameters. The new prediction model is shown as Equation 4-2 (AASHTO 2010).

$$N = N_{spf} \times CMF_1 \times CMF_2 \times ... \times CMF_i \tag{4-2}$$

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.

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 by 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 0 percent grade specification. Many jurisdictions do not allow roads to be constructed with a 0 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. They are a starting point for further analysis.

4.2.3 Calibration

The HSM predictive model was developed from data sourced from several regions in the U.S. However, it 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.

$$C = \frac{N_{actual}}{N_{pred}} \tag{4-4}$$

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.

4.2.4 HSM Model Summary

The HSM predictive model uses an SPF calculation to establish a baseline prediction based on vehicle exposure. CMFs are used to account for variations in the prescribed base conditions by adjusting the prediction for each segment up or down based on roadway attributes. The predicted number of crashes for a set of segments can be pooled and compared to the actual number of crashes. This ratio produces a calibration factor that can be multiplied by the SPFs and CMFs to create a full predictive model representative of jurisdiction-specific conditions.

4.3 Curve and Tangent Combination

Thus far, this report 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 for tangent segments on rural two-lane two-way 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 the UDOT LiDAR project (Ellsworth 2013), this research was able to focus on curved segments of rural two-lane two-way highways.

Since almost all highways are a mix of curved and tangent segments, it would be useful to create a single model to analyze both types. This section presents the HSM approach for incorporating both curved and tangent segments. It also addresses different methods used for parameterizing horizontal curvature 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.

4.3.1 HSM Approach

As described in Chapter 3, the HSM predictive model includes a CMF that specifically adjusts the SPF for horizontal alignment variations. The equation for the CMF is shown 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.

The horizontal alignment CMF incorporates curve length and radius as well as adjustments for spiral transitions. This research assumed that curves in Utah do not have spiral transitions. This CMF is always greater than 1.0, which is the 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.

4.3.2 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 methods

were attempted for parameterizing horizontal curvature, including separating curves and tangents by a simple indicator variable, representing curve class as a series of indicator variables, using curve radius, and using an inverse transformation of curve radius.

4.3.2.1 Simple Indicator Variable

The first proposed method uses an indicator variable for horizontal alignment by assigning 0 for tangent segments and 1 for curved segments. It 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.

4.3.2.2 Curve Class as a Series of Indicator Variables

The second proposed method 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 3. The classification breakdown outlined in Table 3.1 is reprinted in Table 4.2 for convenience.

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

Since this method 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.

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. The indicator variable method allows for isolation of specific classifications that may correlate better than other classifications. It contrasts with the simple indicator variable method that classifies all curves as the same.

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

4.3.2.3 Curve Radius

The third proposed method uses curve radius rather than a classification derived from degree of curvature. Also, instead of grouping curves into classification bins, this method uses a continuous variable. The challenge comes in assigning a radius to tangent segments. 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 method requires the radius measurement for each curved segment.

4.3.2.4 Inverse Transformation of Curve Radius

The fourth proposed method requires a simple inverse transformation of the curve radius method. The idea for this method is that the value of 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 method creates a better distribution of values to match the HSM model (AASHTO 2010).

4.3.3 Curve and Tangent Combination Summary

This section evaluated several methods for addressing horizontal alignment in a crash prediction model. Included in the evaluation were the method used in the HSM and methods for parameterization of horizontal alignment for use in an NB model. The parameterization methods included 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.

4.4 Utah Specific Model

This section discusses the creation of models to predict crashes on rural two-lane two-way highways in Utah. It focuses on the overall development of NB and EB models using all independent variables that are statistically significant.

4.4.1 Negative Binomial Development

An NB model was developed by using JMP, a statistical software package that is a graphical interface for SAS software (SAS 2013). JMP can create an NB model with any number of independent variables and interactions of variables. It estimates coefficients for each variable within a model and calculates the p-value for each. A p-value is the probability that a randomized experiment will lead to a test statistic that is at least as extreme as the one observed (Ramsey and Schafer 2002).

Using the p-values, researchers were able to use a backward stepwise technique for identifying significant variables. 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 to remain after the process is completed. The NB model takes 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 depends on the results of the backward stepwise technique.

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

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

β_0 =   intercept,

β_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 and a value of 0 if it was not. Speed limit was based on posted speed limit for the segment in increments of 5 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 2 feet. The AADT and radius variables underwent transformations to create more normal distributions and to reduce the differences in variance.

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 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. Ultimately, 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. This technique allows each variable to 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

4.4.2 Empirical Bayes Model

The EB model is based on the results of a crash prediction model and 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. The EB model weights observational data more when crash data are widely dispersed. It is especially appropriate for before-after studies (Srinivasan et al. 2009).

4.5 Methodology Summary

This chapter primarily discussed the NB and EB models that estimate expected crashes on a road segment. The NB model uses a backward stepwise regression technique to isolate only the significant variables depending on a chosen confidence level. The EB model looks at both predicted and observed crashes to create an expected value for the number of crashes on a given segment. This is done with a dispersion parameter calculated by 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 looks at the results of the Utah-specific model, HSM predictive model, and curve and tangent combinations.

5.0 RESULTS

5.1 Overview

This chapter presents the results of the crash prediction modeling efforts. The modeling efforts include calibration of the HSM predictive model, development of a model that incorporates both curved and tangent segments, and development of a Utah-specific model. First, a summary of data collection efforts is presented. Then, the HSM predictive model calibration factors are discussed followed by a discussion of results from modeling horizontal alignment by combining curved segments and tangent segments into a single parameter. Next, the NB regression and EB models are discussed as part of the Utah-specific model development for curved segments. A summary of the data collection efforts and modeling results concludes this chapter.

5.2 Data Collection Efforts

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 that 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

41

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 was comprised of 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.

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.

5.3 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

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

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

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.

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

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

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 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.

5.4 Curve and Tangent Combination

Four different approaches were attempted to combine segments of different horizontal alignments into the same parameter. These approaches included single indicator variable, curve class as a series of indicator variables, curve radius, and inverse transformation of curve radius. A method was created for each approach and for each sample period using the backward stepwise technique. The results of each method are presented in this section along with a discussion of the next steps in this area of research.

5.4.1 Single Indicator Variable

The single indicator variable method assigns 0 for tangent segments and 1 for curved segments. While this is a very simple method 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 Table 5.4 for the three-year and five-year samples, respectively.

Table 5.3 Parameter Estimates for Three-year Sample Using Indicator Variable

Term	Estimate	Standard Error	Wald χ^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 χ^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

Table 5.3 and Table 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-square distribution to produce a p-value shown in the fifth column. The variables shown in Table 5.3 and 5.4 are the only variables whose p-values were less than 0.05. 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 5.4.5. The single indicator variable method groups all curves and tangents into two homogeneous classifications. It therefore does not account for sharpness of each curve and was rejected for this reason.

5.4.2 Curve Class as a Series of Indicator Variables

This method involved assigning a classification to each curve based on degree of curvature using the HPMS definitions introduced in Chapter 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 output. Table 5.5 and Table 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 χ^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 χ^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 method 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.

<u>5.4.3 Curve Radius</u>

The main advantage of the curve radius method 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. Infinity is impractical from a modeling standpoint, so an arbitrarily high radius of 10 miles was assigned to each tangent segment. Table 5.7 and Table 5.8 show the parameter estimates for the three-year and five-year samples, respectively.

The variable for curve radius had a very low p-value in each sample, 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 method 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 5.4.5.

Table 5.7 Parameter Estimates for Three-year Sample Using Curve Radius

Term	Estimate	Standard Error	Wald χ^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 χ^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

Assigning an arbitrary radius of 10 miles to each tangent segment resulted in a wide range, with all tangents at the extreme upper end and all curves at the lower end. A value needed to be identified as a dividing point, with all segments with radii greater than the dividing point 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.

5.4.4 Inverse Transformation of Curve Radius

This method requires a simple inverse transformation of the curve radius method. The impetus for this method was difficulty in assigning an arbitrary radius to tangent segments. Taking the inverse of the radius results, 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, shallow curves have values close to 0, and as the curves become sharper, the value increases. Table 5.9 and Table 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 χ^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 χ^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

5.4.5 Discussion

One of the observations during this process was that most of the tangent segments had significantly longer analysis lengths than the curved segments. This makes intuitive sense because segments were defined by homogeneous roadway attributes and tangents can be homogeneous for many miles, whereas curves are usually much less so. Since segment length is a major factor in vehicle exposure, which in turn is a major factor in the number of crashes on a segment, it is possible that methods placing too much weight on segment length would assign a high number of crashes to longer segments. While longer segments should have a higher number of crashes due to higher exposure values, if tangent segments are longer than curved segments on average, the model may falsely assume that tangency is a strong predictor 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 curvature approaches. The results the analyses are shown in Table 5.11 and Table 5.12.

Table 5.11 Interaction of Curve Presence and Segment Length for Three-year Sample

Term	Estimate	Standard Error	Wald χ^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 χ^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.

The idea of a combining curves and tangents in the same model should 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 includes both curved and tangent segments seems to be the most promising approach for incorporating horizontal alignment. Since this research is focused on curved segments, these ideas did not receive further attention. The modeling continued only for curved segments.

5.5 Utah Specific Model for Curved Segments

This section addresses 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 are discussed and final models are presented.

5.5.1 Negative Binomial Model

An NB model was developed using JMP 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 to identify significant variables. This 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.

The model was run on the first dataset and the other two datasets were used for validation purposes. When the model used data from the second set for the validation comparison, 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. The third dataset underpredicted the number of crashes by about 6.7 percent for the three-year sample and 3.7 percent for five-year sample when it was used for the validation comparison.

A chi-square test for goodness of fit between the actual and estimated crash frequency distributions was performed on the second and third datasets for both the three-year and five-year samples. The estimated crash frequency distribution was determined using the histogram

52

module of the Data Analysis feature of Microsoft Excel. This module rounds segment crash numbers down to the nearest whole number.

The results of the chi-square test for these datasets are presented in Appendix A. The critical values for the chi-square distribution for the second dataset were 7.815 at a 95 percent confidence level and with 3 degrees of freedom for the three-year sample and 11.07 at a 95 percent confidence level and 5 degrees of freedom for the five-year sample. This chi-square test did not assume any mathematical distribution, that is, no parameter was used. Hence, the degree of freedom was the number of cells used minus 1. The chi-square statistics for the three- and five-year samples were 219.156 and 257.697, respectively. Both of these values were greater than the critical value for each sample, indicating no evidence that the second dataset's estimated crash distribution approximated the actual crash distribution.

For the third dataset the critical value was found to be 9.488 at a 95 percent confidence level and 4 degrees of freedom for the three-year sample and 14.07 at a 95 percent confidence level and 7 degrees of freedom for the five-year sample. The chi-square statistics for the three- and five-year samples were 72.629 and 135.897, respectively. Similar to the second dataset, these values were much greater than the critical values, indicating no evidence that the third dataset's crash distribution approximated the actual crash distribution.

With this information, researchers decided to use the combined dataset for the NB model. As shown in Figure 5.1, the randomly selected study segments are distributed across the state. The inherent problem with a single dataset is lack of 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 it. The NB model takes the form shown in Equation 5-6 (Ramsey and Schafer 2002).

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

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

β_0 = intercept,

β_i = coefficient for variable x_i,

x_i = independent variable, and
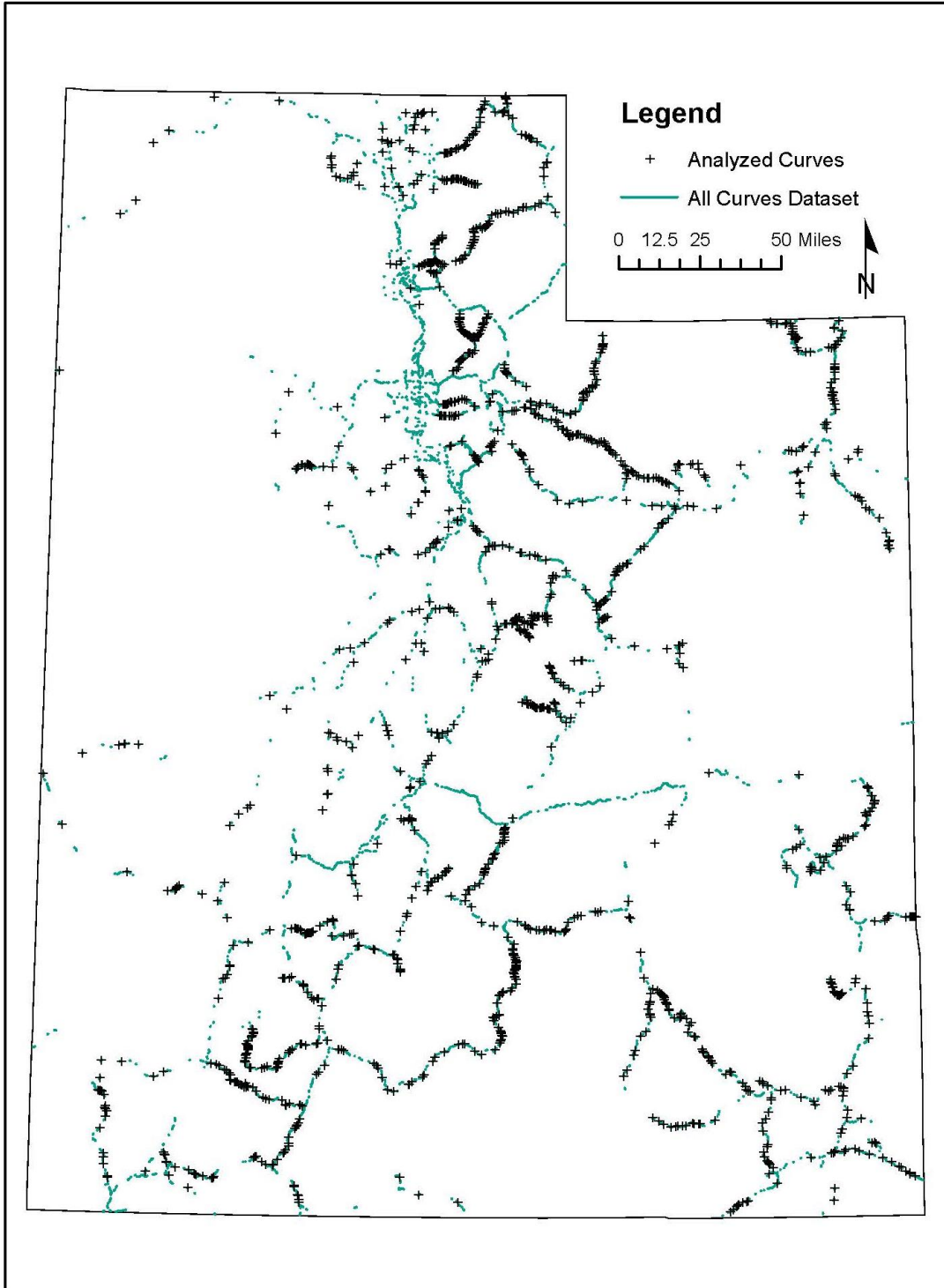
n = number of independent variables.

Figure 5.1 Combined Dataset of Curved Segments

54

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

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

The final three NB regression outputs from JMP using the backward stepwise technique are shown in Table 5.13, Table 5.14, and Table 5.15 for the three-year sample. Table 5.16, Table 5.17, and Table 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 χ^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 χ^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 χ^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 χ^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 χ^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 χ^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 that measures the lack of fit of a model and adds a penalty for it (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).

$$BIC = n \times \ln(RSS) + p \times \ln(n)$$
(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.

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

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

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.

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

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

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 Figure 5.2 and Figure 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-square test for goodness of fit was performed on the validation dataset for both the three-year and five-year samples. The results of the chi-square for the validation dataset are found in Appendix A. The critical values for the chi-square distribution were 7.815 at a 95 percent confidence level and 3 degrees of freedom for the three-year dataset and 11.07 at a 95 percent confidence level with 5 degrees of freedom for the five-year dataset, respectively. The chi-square statistic was 1.610 for the three-year sample and 10.993 for the five-year sample. Both of these values were less than the critical values, indicating that the estimated crash distribution approximated the actual crash distribution.

The actual numbers of crashes in the three- and five-year samples of the validation dataset were 204 and 396, respectively. 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, these values show the reality of this type of modeling. Especially when the predictions and the actual observed data do not always match, real conditions cannot be ignored.
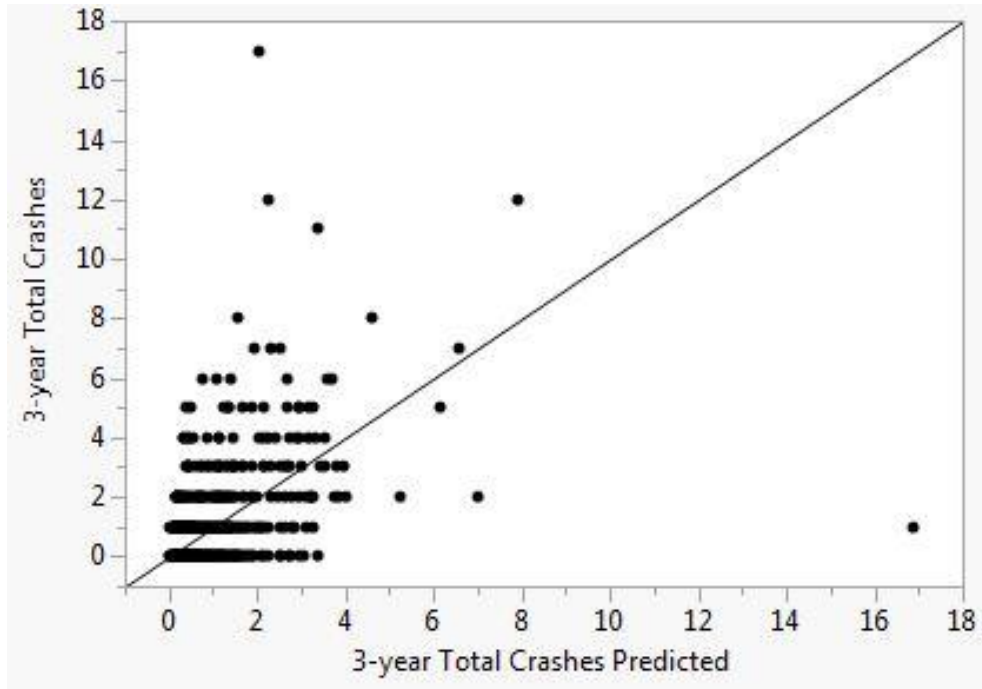
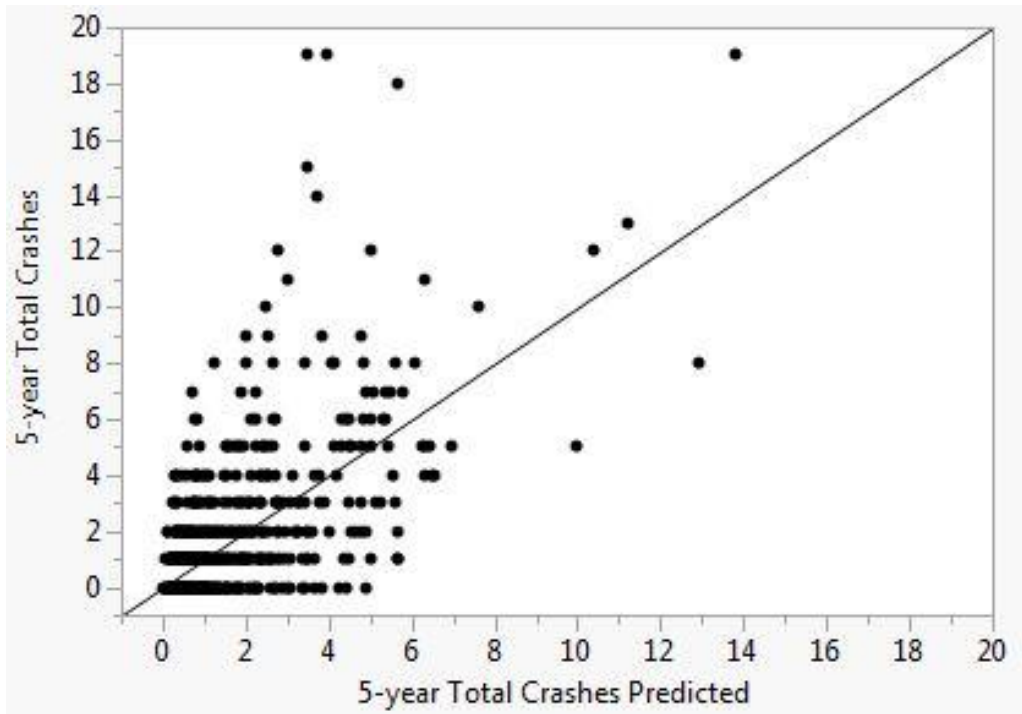**Figure 5.2 Actual vs. Predicted Three-Year Total Crashes**

Figure 5.3 Actual vs. Predicted Five-Year Total Crashes

5.5.2 Empirical Bayes Model

The EB model was used to estimate the total number of crashes based on predicted and actual values. It uses a dispersion parameter to weight the predicted values. If the predicted values are over-dispersed, the model gives more weight to the actual number of crashes in the calculations. The dispersion parameters for the three- and five-year samples were found to be 0.6491 and 0.5755. This means that the three-year sample data were more dispersed than the five-year sample data. The actual weights 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 it more closely approximates the actual number of crashes compared to the NB model because it incorporates the actual number of crashes. 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 its benefits for more closely approximating the expected number of crashes based on actual and predicted values.

5.6 Summary of Results

This chapter discussed the results of the three different modeling procedures: HSM predictive model, curve and tangent combination, and the Utah-specific model. 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 plus 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. Decreasing calibration factors for

the three- and five-year combined segment samples and for tangent segments over an eight-year period provide evidence of improved safety 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 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 describing the 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. 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 the validation dataset by 32 percent for the three-year sample and 19 percent for the five-year sample.

After creation of the models, only four significant variables remained 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. The EB model is used to better approximate the number of expected crashes. It is essentially a weighted average between the predicted and actual numbers 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 RTM bias. 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.

6.0 CONCLUSION

6.1 Overview

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 report presents the methodology for developing crash prediction models and gives results of the models' accuracy. The research was accomplished by calibrating the HSM crash prediction model as well as by creating Utah-specific models. The crash 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 included 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.

6.2 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. The three- and five year samples produced calibration factors of 1.50 and 1.60, respectively, which means that the HSM model underpredicts the number of crashes on curved rural two-lane two way roads in Utah by 50-60 percent as compared to the national dataset used as the basis for the HSM SPF.

The Utah-specific models were developed using NB models. 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.

The HSM predictive model uses up to 12 CMF variables in addition to AADT and segment length. By comparison, the Utah-specific models only require four variables. 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, which are weighted and added together to provide the overall result. The weight is dependent on a dispersion parameter calculated during model development. EB models are meant to be site-specific. Therefore, results from the combined datasets were not reported.

6.3 Recommended Models

The HSM calibration factors were found to be 1.50 and 1.60 for the three- and five-year samples, 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.

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

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

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. Simplicity, however, 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 this report should be used in conjunction with the Utah-specific crash prediction NB models to correct for RTM bias. 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).

6.4 Future Research Needs

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 the possibility of defining a model capable of combining curved and tangent segments.

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 in turn is a key variable in crash prediction modeling, longer average lengths for tangent segments may be falsely correlated with higher crash rates.

More recent data would 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. (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). (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.

Appendix A: Results of Chi-Square Tests

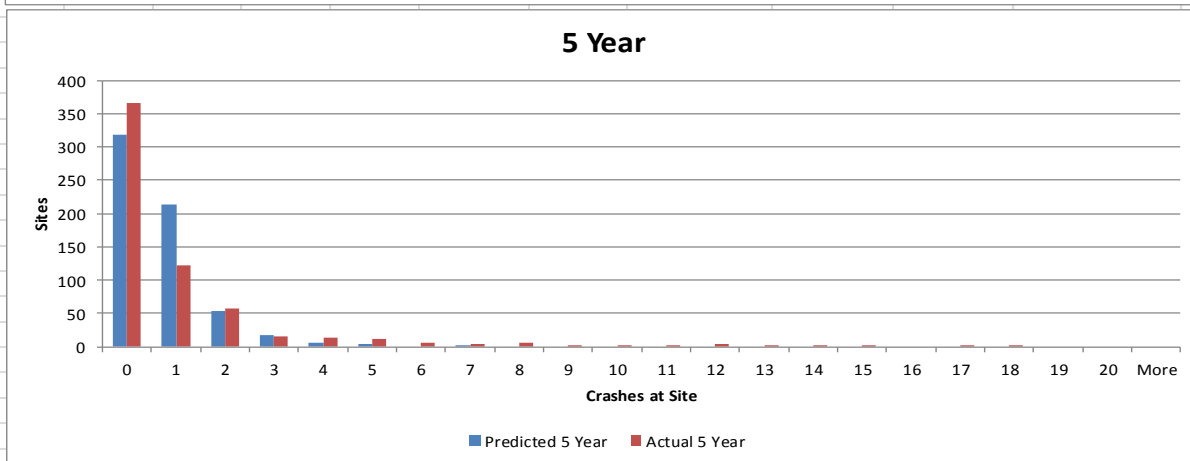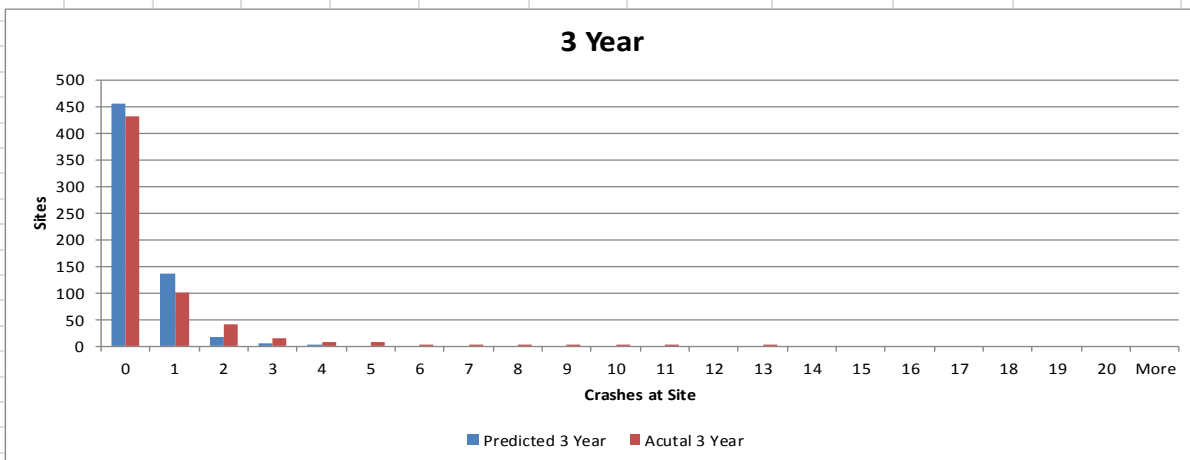| Bin	Predicted 3 Year	Acutal 3 Year	Predicted 5 Year	Actual 5 Year	Combined 3 year predicted	Combined 3 year actual	(o-e)^2/e 3 year	Combined 5 year predicted	Combined 5 year actual	(o-e)^2/e 5 year
0	455	431	318	365	455	431	1.266	318	365	6.947
1	136	101	213	123	136	101	9.007	213	123	38.028
2	17	41	54	57	17	41	33.882	54	57	0.167
3	5	15	18	16	7	42	175.000	18	16	0.222
4	2	7	6	13				6	13	8.167
5	0	7	4	12				6	41	204.167
6	0	4	0	6						
7	0	3	2	4	Σ	219.156		Σ	257.697	
8	0	2	0	6						
9	0	1	0	2	d.f. = 3			d.f. = 5		
10	0	1	0	1	Chi-square critical = 7.815			Chi-square critical = 11.07		
11	0	1	0	1						
12	0	0	0	4						
13	0	1	0	1						
14	0	0	0	1						
15	0	0	0	1						
16	0	0	0	0						
17	0	0	0	1						
18	0	0	0	1						
19	0	0	0	0						
20	0	0	0	0						
More	0	0	0	0						

3 Year

5 Year

Figure A-1. Results of Chi-square Goodness-of-fit Test on Dataset 2

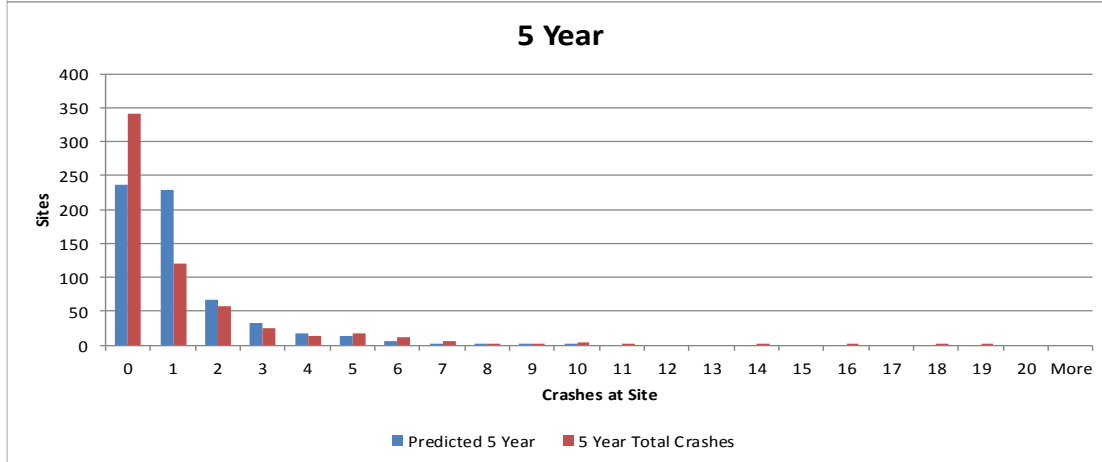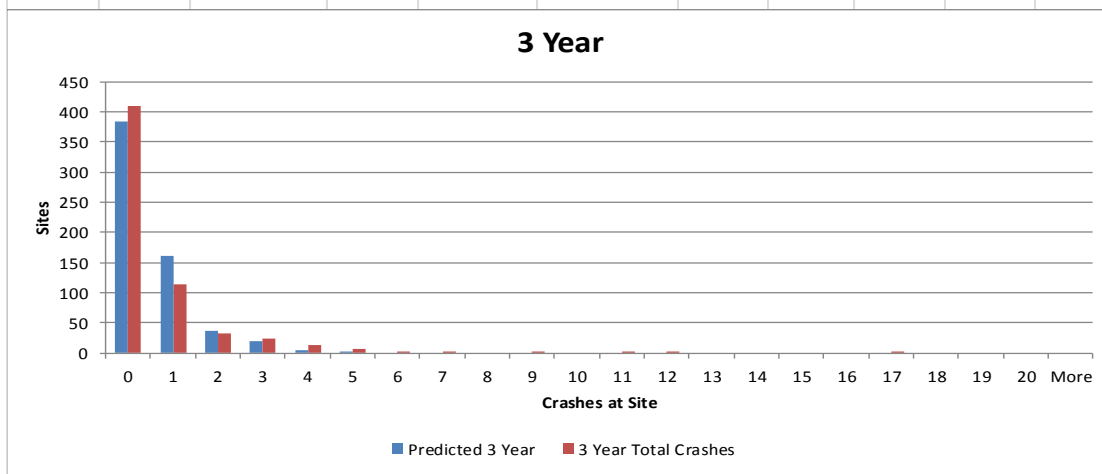| Bin	Predicted 3 Year	3 Year Total Crashes	Predicted 5 Year	5 Year Total Crashes	Combined 3 year Predicted	Combined 3year actual	(o-e)^2/e 3 year	Combined 5 year predicted	Combined 5 year actual	(o-e)^2/e 5 year
0	384	409	237	341	384	409	1.628	237	341	45.637
1	161	113	229	120	161	113	14.311	229	120	51.882
2	36	33	67	57	36	33	0.250	67	57	1.493
3	19	24	32	26	19	24	1.316	32	26	1.125
4	5	13	17	14	8	29	55.125	17	14	0.529
5	3	7	13	17				13	17	1.231
6	0	3	6	12				6	12	6.000
7	0	2	2	6				7	21	28.000
8	0	0	1	1						
9	0	1	2	2		Σ	72.629		Σ	135.897
10	0	0	2	4						
11	0	1	0	3		d.f. = 5 - 1 = 4			d.f. = 8 - 1 = 7	
12	0	1	0	0		Chi-square critical = 9.488			Chi-square critical = 14.07	
13	0	0	0	0						
14	0	0	0	1						
15	0	0	0	0						
16	0	0	0	1						
17	0	1	0	0						
18	0	0	0	1						
19	0	0	0	2						
20	0	0	0	0						
More	0	0	0	0						

Figure A-2. Results of Chi-square Goodness-of-fit Test on Dataset 3

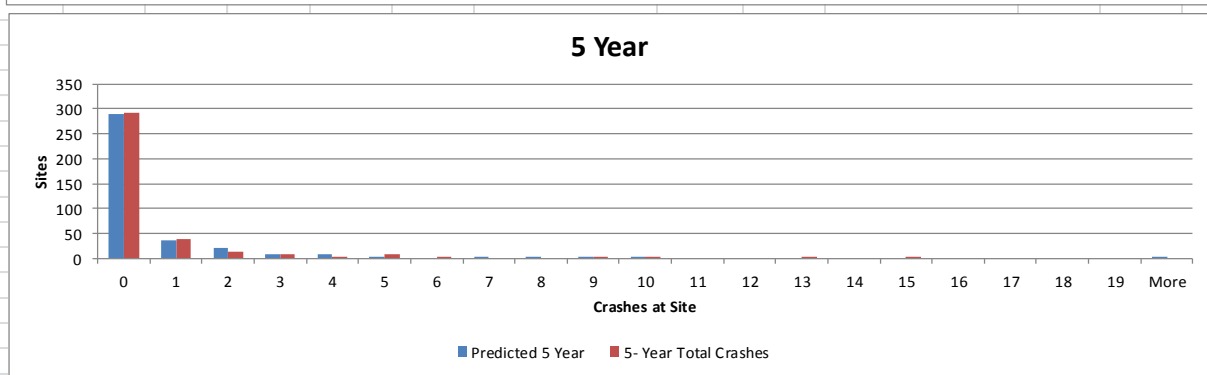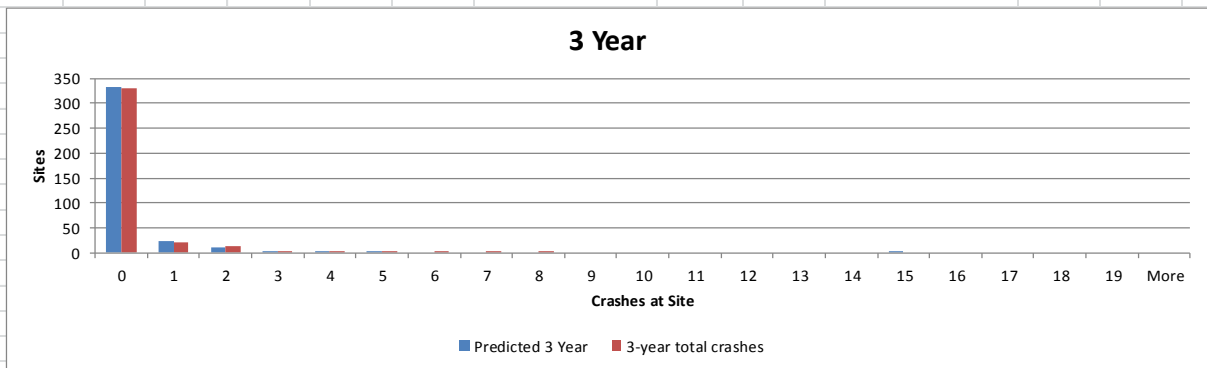| Bin	Predicted 3 Year	3-year total crashes	Predicted 5 Year	5- Year Total Crashes	Combined 3 Year Predicted	Combined 3 year Actual	(o-e)^2/e 3 year	Combined 5 Year Predicted	Combined 5 Year Actual	(o-e)^2/e 3 year
0	332	331	289	293	332	331	0.003	289	293	0.055
1	23	20	37	38	23	20	0.391	37	38	0.027
2	11	12	22	13	11	12	0.091	22	13	3.682
3	1	4	8	7	8	11	1.125	8	7	0.125
4	4	2	7	4				7	4	1.286
5	2	1	4	8				11	19	5.818
6	0	1	0	4						
7	0	1	1	0	Σ =		1.610	Σ =		10.993
8	0	2	3	0						
9	0	0	1	3	d.f. = 4 - 1 = 3			d.f. = 6 - 1 = 5		
10	0	0	1	1	Chi-square critical = 7.815			Chi-square critical = 11.07		
11	0	0	0	0						
12	0	0	0	0						
13	0	0	0	1						
14	0	0	0	0						
15	1	0	0	2						
16	0	0	0	0						
17	0	0	0	0						
18	0	0	0	0						
19	0	0	0	0						
More	0	0	1	0						

Figure A-3. Results of Chi-square Goodness-of-fit Test on Validation Dataset