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

By:

Casey S. Knecht
Access Management Coordinator
Oregon Department of Transportation
455 Airport Rd SE, Bldg B
Salem OR 97301
P: (503) 986-5170
F: (503) 986-2630
casey.knecht@odot.state.or.us

Mitsuru Saito, Ph.D., P.E.
(Corresponding Author)
Professor
Department of Civil & Environmental Engineering
Brigham Young University
368 Clyde Building
Provo, UT 84602
P: (801) 422-6326
F: (801) 422-0159
msaito@byu.edu

Grant G. Schultz, Ph.D., P.E., PTOE
Professor
Department of Civil & Environmental Engineering
Brigham Young University
368 Clyde Building
Provo, UT 84602
P: (801) 422-6332
F: (801) 422-0159
gschultz@byu.edu

Paper # 16-1859
Submitted for publication in Journal of Transportation Research Board
Revised version submitted on November 12, 2015

Word Count: Abstract and Body Text & Tables/Figures = 234 + 5,947 + (2 Figures and 3 Tables) = 7,431
ABSTRACT
Crash prediction models for curved segments of rural two-lane two-way highways in the state of Utah were developed. The modeling effort included the calibration of the predictive model found in the Highway Safety Manual (HSM) as well as the development of Utah-specific models developed using negative binomial regression. The data for these models came from randomly sampled curved segments in Utah, with crash data coming from years 2008-2012. The total number of randomly sampled curved segments was 1,495. 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.
**INTRODUCTION**

Highway safety affects everyone, as a crash on a highway does not only affect the people involved in the crash, but also others who are affected by the ensuing delays and societal costs. Understanding and being able to identify the reasons behind crashes and resolving potential causes are paramount. To do so, researchers have developed crash prediction models that are based on historical crash data to estimate the number of future crashes under prevailing conditions that can be used to evaluate the contributions of physical attributes to crash occurrence.

One of the procedures for crash prediction modeling is using a safety performance function (SPF). SPFs are regression models that estimate average crash frequency for a specific site type as a function of annual average daily traffic (AADT) and segment length \((l, 2)\). 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 \((1)\). 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 \((3, 4)\) developed calibration factors for the Utah Department of Transportation (UDOT). The calibration factors developed at that time were specific to tangent segments of rural two-lane two-way highways in Utah, because at the time of their research no data were available for horizontal curvature. Since that research, UDOT has performed an inventory of all highway curvature within the state of Utah as part of its Light Detection and Ranging (LiDAR) asset management program. With this additional data, UDOT is desirous to calibrate the HSM predictive model specifically for curved segments of rural two-lane two-way highways in Utah. Similarly, UDOT has requested the development of Utah-specific crash prediction models for rural two-lane two-way highways exclusive of the HSM predictive model.

This paper presents the background, purpose and scope of the study, literature review, data preparation, analysis results, and conclusions of the study.

**BACKGROUND**

The Utah Department of Transportation (UDOT) contracted with a private firm to collect LiDAR point cloud data and process this data into segments which would define the point of curvature (PC), point of tangency (PT), radius of curvature, and other important features of the curve geometry. The output dataset from that contract, and the starting point for this method, is a shapefile with attributes for segments along all UDOT maintained roadways with tabulated attributes for each section. This dataset was intended to be used for developing the Utah Crash Prediction Model (UCPM) \((5)\). It contained a unique segment ID number; route designation; direction of travel; mileposts of the segment beginning and end points; classification of the segment according to the Highway Performance Monitoring System (HPMS) Field Manual \((6)\); degree of curvature; radius of curvature; length of curve; and latitude, longitude, and elevation values for the start and end points of each segment. A major reason why the data as provided by the contractor were insufficient for highway safety analysis, and a justification for this research,
was that there were several sections of roadway where one curve was defined as several independent segments. This multi-segmentation of a curve was not purposefully done, but it creates difficulty in defining and cataloguing curves for safety analysis. Tabular data from the contractor-provided dataset were extracted, and the Horizontal Alignment Finder (HAF) algorithm (7) was developed to identify curved segments. A linearly referenced shapefile of all state routes was used to create the final curve shapefile once tabular manipulation was finished.

PURPOSE AND SCOPE OF THE STUDY

The purpose of the study is to develop crash prediction models for curved segments of rural two-lane two-way highways in Utah using historical crash data and facility data recently collected as part of UDOT’s LiDAR asset management program. All crash severities were included, from property damage only through fatal crashes. Ancillary efforts done for the data collection necessary for this analysis are also performed, including the development of an algorithm (the HAF algorithm) to identify curved segments given the point cloud data provided by 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 developing 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 occurrences. The models identify which factors play the largest role in crash prediction. With this information, UDOT can focus its efforts on the improvements that will make the most difference in safety.

LITERATURE REVIEW

This section of the paper provides a brief summary of the literature review conducted for the study on LiDAR, highway curvature, SPFs and CMFs. Knecht (8) contains the entire literature review conducted for the study, containing topics related to highway geometry and safety as well as the acquisition and analysis of data, including LiDAR, highway curvature, SPFs, CMFs, calibration factors, and statistical methods.

Light Detection and Ranging (LiDAR)

LiDAR data are well-suited for transportation applications. LiDAR data are especially useful when combined with geographic information system (GIS) technology to determine accurate 3D surface representations and characteristics (9). 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 (10). LiDAR is capable of providing information at high spatial resolutions and accuracies. Pradhan and Rasdorf (9) discussed the accuracy of LiDAR data, and in 1999, LiDAR data were found to be accurate to +/- 15 centimeters. Figure 1 shows a sample LiDAR capture which exemplifies the accuracy level of LiDAR compared to the image captured by Roadview Explorer.

Many transportation agencies are utilizing mobile vehicles to collect a wide variety of asset data (11). In 2011, UDOT commenced a project that would eventually collect highway infrastructure data for every state road in Utah using LiDAR. The data have an average accuracy of +/-3 centimeters (10). As the technology improves and the machinery becomes more sophisticated, accuracy naturally improves with it. LiDAR employs a significantly higher concentration of data points than surveying or digital elevation model (DEM) (12). Thus less interpolation is required and the points create a redundancy to reduce error.
Highway Curvature

Highway curvature will play an important role in this research. A previous study by Saito et al. (3) 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 curve on SPFs. Approximately 25 percent of all fatal crashes in the United States in 2002 occurred on horizontal curves (13). 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 (14, 15). 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 (16).

Determining segmentation of highway curves can prove difficult. Srinivasan et al. (16) 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 associated with significantly higher numbers of curve-related crashes. Crash reduction rates for horizontal curve improvements were determined for flattening curves, widening lanes, widening paved shoulders, adding unpaved shoulders, adding spiral transitions, and improving superelevation (17).

There are several factors that can affect safety on curves including signage, pavement markings, and roadside hazards (13, 17, 18). In a study on rural two-lane roads in Indiana, Labi (18) 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.

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 \((I, 2)\). SPFs developed in a specific jurisdiction or on a general level can be recalibrated for a different jurisdiction. The HSM contains SPFs and there are documented calibrations that have already been performed \((I9)\).

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

\[
N_{spf} = AADT \times L \times 365 \times 10^{-6} \times e^{-0.312}
\]

where,
\[
N_{spf} = \text{number of predicted annual crashes},
\]
\[
AADT = \text{average annual daily traffic, and}
\]
\[
L = \text{segment length (mi)}.\]

In 2011, a study was performed on calibrating the HSM to predict total crashes on highways in Oregon \((20)\). In this particular study, the guidelines for calibration set forth in the HSM were followed. The study mentioned specifically the difficulty in preparing the data set and the local adjustments made such as adjusting sample sizes for underrepresented facility types. Additionally, the target number of 100 crashes per year could not be achieved at several low-volume intersections.

Any number of variables can be used in a model. The key is choosing variables that are most appropriate and affect the SPF the most. Data collection costs increase as the number of variables increases. The HSM model \((I)\) 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 by incorporating crash modification factors (CMFs) which are multiplied to the number of predicted annual crashes found by the base crash prediction model. The HSM specifies that SPFs should incorporate traffic volume and crash frequency, while geometric design and traffic control features should be incorporated through CMFs.

Crash Modification Factors (CMFs)

CMFs represent the relative change in crash frequency due to a change in one specific condition, estimating the effect of a particular geometric design or traffic control or the effectiveness of a particular treatment or condition \((I)\). CMFs are often preferred by transportation safety analysts because they allow baseline models to be recalibrated for different jurisdictions.

For rural two-lane two-way highway segments, the HSM model has CMFs for 12 design and control features: lane width, shoulder width and type, horizontal curve length and radius, horizontal curve superelevation, grade, driveway density, centerline rumble strips, passing lanes, two-way left-turn lanes, roadside design, lighting, and automatic speed enforcement \((I)\).
When changes are made to highway geometry and/or segmentation, the calibration of CMFs will need to be performed. Hauer (21) claimed that driver behavior can be affected any time there is a change. For example, a road that is repaved may provide an increased sense of safety even if the actual highway geometry is identical. This increased sense of safety is in addition to the actual increase in safety that comes from replacing pavement that is in poor condition (18).

DATA PREPARATION

The first step of data collection was to randomly select curved segments that were representative of rural two-way two-lane highways in Utah. Once the segments were selected for further analysis, the next step was to gather sufficient data on the components of the selected segments that would affect the predictive power of the crash prediction models (i.e., variables in the predictive models included in the HSM). While the segments do not need to meet the base conditions, one must know the values of the components so as to determine the value of an appropriate CMF. In this paper, only data resources, data limitation, and segment sampling are presented. For the complete description of data required for determining CMF values, refer to Knecht (8).

A method was needed to combine the curved segments within the same curve identified by the LiDAR program as curves into one single section. The HAF algorithm provided a method by which the curve segments were combined to a reasonably high success rate (85 percent or better) (7).

Data Resources

The collection of data came from various sources throughout the process. The availability and accessibility of data online allowed for a widespread survey of segments across the state. In many cases, the different resources had redundant features which allowed for verification and validation of different data. The Traffic on Utah Highways (22), the UDOT Data Portal (23), Traffic Studies (24), and the UDOT Crash Database (25) available from UDOT provided raw data for the analysis while Google Earth (26), Roadview Explorer (27), and ArcGIS (28) were used to obtain data not available from the UDOT raw data files but necessary to complete the intended analysis.

Data Limitations

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

Proper Sampling of Study Segments

Some segments initially selected were removed from the dataset at various points in the data compilation process. Duplicate or overlapping segments were removed immediately based on route number and mileposts. Segments that were in urban areas or residential areas were removed.
Segments with speed limits lower than 30 mph were removed. The reasoning for this is because speed limits lower than 30 mph are generally associated with high pedestrian traffic, residential areas, and/or vehicles stopping for roadside services and attractions.

Also, if the lanes were not striped or if they were less than 9 feet wide, the segment was removed. Segments that contained a stop sign, signal, or other traffic control device for the main directions of traffic were removed since their inclusion would be better suited for an intersection analysis. In some areas, the segment included a 90° or near-90° turn from one cardinal direction to another with a very small curve radius. This was encountered where a route traveled on an east/west roadway and then the route designation changed to a north/south roadway. Thus, the change was usually at a four-way intersection where the two legs of the designated route received preference. These segments were removed regardless of the presence of traffic control devices.

Several segments were within national and state parks and recreation areas. While that did not merit immediate removal, most of the segments within these park and recreational areas were near on-road services such as toll booths, information booths, boat launches, ranger stations, and recreation vehicle dump stations. These on-road services prevent free-flow operation, and therefore the segments that contained or were near any of these services were removed from the dataset.

**ANALYSIS RESULTS**

This section presents the results of the calibration of the HSM predictive model and the development of Utah-specific models for curved segments using the NB regression and EB modeling methods.

**Calibration of the HSM Predictive Model**

The HSM predictive model is based on an SPF, multiple CMFs, and a calibration factor as explained previously. The SPF model for the base conditions on rural two-lane two-way roads was shown previously in Equation 1. After calculating the number of predicted crashes with the SPF along with the available CMFs outlined in the HSM, the predicted values were compared to the actual values. The data were separated into three unique sets that were chosen at random, with each set meeting the qualifications set forth in the HSM (1) (see Table 1a). Three data sets were created to test if the calibration factors would be similar to each other or significantly different. Segments chosen by Data Set 1 were excluded when segments for Data Set 2 were selected. Similarly, Data Set 3 did not include segments chosen for Data Set 1 and 2. The calibration factors are shown in Table 1a. 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, implying 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 (3) 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 1b.

It is important to remember that AADT is the only independent variable that consistently changes between sample periods when using the HSM predictive model. While construction projects will almost certainly take place throughout the year, most projects do not affect the variables in the HSM predictive model. While AADT has increased on almost all segments since the previous research, the actual number of crashes has decreased. A decreasing calibration factor implies that either the actual number of crashes is decreasing, the predicted number of crashes is
increasing, or both. This supports the assumption that overall safety has improved not only since the year 2008, but since at least the year 2005. The improvements in safety can be seen in both the curved segment sample and the tangent segment sample. This also shows that calibration factors need to be updated regularly, as they can change significantly within a few years’ time.

Table 1. HSM Calibration Factors

(a) HSM predictive model calibration factors for curved segments

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Number of Samples</th>
<th>3-year Sample (2010 – 2012)</th>
<th>5-year Sample (2008 – 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>321</td>
<td>1.42</td>
<td>1.58</td>
</tr>
<tr>
<td>2</td>
<td>566</td>
<td>1.50</td>
<td>1.54</td>
</tr>
<tr>
<td>3</td>
<td>608</td>
<td>1.53</td>
<td>1.64</td>
</tr>
<tr>
<td>Combined</td>
<td>1,495</td>
<td>1.50</td>
<td>1.60</td>
</tr>
</tbody>
</table>

(b) Calibration factors for tangent segments for the segments used by Saito et al. (3)

<table>
<thead>
<tr>
<th>Sample Period</th>
<th>Predicted Crashes</th>
<th>Actual Crashes</th>
<th>Calibration Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005 - 2007</td>
<td>368</td>
<td>426</td>
<td>1.16</td>
</tr>
<tr>
<td>2008 - 2010</td>
<td>403</td>
<td>415</td>
<td>1.03</td>
</tr>
<tr>
<td>2009 - 2011</td>
<td>403</td>
<td>374</td>
<td>0.93</td>
</tr>
<tr>
<td>2010 - 2012</td>
<td>422</td>
<td>354</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Utah Specific Model for Curved Segments

This section will address the development of a crash prediction model specifically for curved segments of rural two-lane two-way highways within the state of Utah and its results using the NB regression and EB modeling methods.

Negative Binomial (NB) Model

The development of an NB model took place using JMP, a statistical software package that is a graphical interface for SAS software (29). This research used a backward stepwise technique for identifying which variables were significant and which were not—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 level. 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. For other variables evaluated, see Knecht (8).

This process was performed on the first dataset, and the other two datasets were used for validation purposes. When the model used the data from the second set, it overpredicted the number of crashes by about 5.5 percent for the three-year sample and 13.8 percent for the five-
year sample. For the third dataset the model underpredicted the number of crashes by about 6.7 percent for the three-year sample and 3.7 percent for five-year sample. With this information, it was decided that the combined dataset should be used for the NB model. The combined dataset was randomly divided into a model set and a validation set, with 75 percent of the segments assigned to create the model and 25 percent of the segments assigned to validate the model.

The NB model takes the form shown in Equation 2 (30).

\[ \ln(N) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \]  

where, \( N \) = number of crashes (predicted or observed), \( \beta_0 \) = intercept, \( \beta_i \) = coefficient for variable \( x_i \), \( x_i \) = independent variable, and \( n \) = number of independent variables.

This is rearranged and shown in Equation 3, isolating the predicted number of crashes on one side of the equation.

\[ N = \exp[\beta_0 + \sum_{i=1}^{n} \beta_i x_i] \]  

The final parameter estimates of the NB regression outputs for the three and five year data are shown in Table 2 and 3, respectively.

### Table 2 Final Parameter Estimates for Three-year Sample

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald ( \chi^2 )</th>
<th>Probability &gt; ( \chi^2 )</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-11.5570</td>
<td>0.8018</td>
<td>207.8</td>
<td>&lt;.0001</td>
<td>-13.1284</td>
<td>-9.9855</td>
</tr>
<tr>
<td>Analysis Length</td>
<td>2.4465</td>
<td>0.4089</td>
<td>35.8</td>
<td>&lt;.0001</td>
<td>1.6450</td>
<td>3.2480</td>
</tr>
<tr>
<td>ln 3yr Veh Count</td>
<td>0.8833</td>
<td>0.0491</td>
<td>323.0</td>
<td>&lt;.0001</td>
<td>0.7870</td>
<td>0.9796</td>
</tr>
<tr>
<td>Total Truck %</td>
<td>-0.0127</td>
<td>0.0046</td>
<td>7.6</td>
<td>0.0059</td>
<td>-0.0218</td>
<td>-0.0037</td>
</tr>
<tr>
<td>Ln Radius</td>
<td>-0.2236</td>
<td>0.0647</td>
<td>11.9</td>
<td>0.0006</td>
<td>-0.3505</td>
<td>-0.0968</td>
</tr>
<tr>
<td>Dispersion</td>
<td>0.6491</td>
<td>0.1058</td>
<td>37.7</td>
<td>&lt;.0001</td>
<td>0.4418</td>
<td>0.8564</td>
</tr>
</tbody>
</table>
Table 3 Final Parameter Estimates for Five-year Sample

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

Models based on the final output of the backward stepwise technique are represented in Equations 4 and 5 for the three-year and five-year samples, respectively.

$$N_{3\text{-year}} = \exp[-11.5570 + (2.4465)(L) + (0.8833)(\ln(VC)) - (0.0127)(TT) - (0.2236)(\ln(R))]$$  \hspace{1cm} (4)

$$N_{5\text{-year}} =  \exp[-11.2040 + (2.5757)(L) + (0.8606)(\ln(VC)) - (0.0148)(TT) - (0.2082)(\ln(R))]$$ \hspace{1cm} (5)

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 6 and 7 for the three-year and five-year samples, respectively.

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

$$N_{5\text{-year}} = 640.6824 \times \text{AADT}^{0.8606} \times R^{-0.2082} \times \exp[-11.2040 + (2.5757)(L) - (0.0148)(TT)]$$ \hspace{1cm} (7)

The ranges, means and standard deviations of the variables to develop these modes are:

- Average AADT for 3-year model = 20 to 10,227 (Mean = 1,473, Standard Deviation (SD) = 1,835); Average AADT for 5-year model = 21 to 10,196 (Mean = 1,480, SD = 1,824)
- Curve radius (R) = 40 ft to 23,559 ft (Mean = 2,157 ft, SD = 1,607 ft)
- Segment length (L) = 0.0744 miles to 1.5505 miles (Mean = 0.1611 miles; SD = 0.1135 miles)
- Truck percentage (TT) = 3 percent to 85 percent (Mean = 28.0 percent, SD = 10.8 percent)

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 (3). One possible theory for this result is 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 might equate to lower crash frequencies on highway segments with increased truck traffic.

A chi-square test for goodness of fit between the distribution of the frequency of actual number of crashes and the distribution of the frequency of estimated number of crashes was performed on the second and third datasets for both the three-year and five-year samples. The distribution of the frequency of estimated number of crashes was determined using the histogram module of the Data Analysis feature of Microsoft Excel. With this module, when the estimated number of crashes at a segment is equal to or greater than 0 and less than 1, it is categorized as a “0 crash” prediction, while it is equal to and greater than 1 and less than 2, it is categorized as having 1 crash, and so forth. A chi-square test for goodness of fit was performed on the validation dataset for both the three-year and five-year samples. The critical values for the chi-square distribution was 7.815 at a 95 percent confidence level and 3 degrees of freedom for the three-year data set and 11.07 at a 95 percent confidence level with 5 degrees of freedom for the five-year data set, respectively. The chi-square statistic was 1.610 for the three-year sample and 10.993 for the five-year sample, respectively. Both of these values were less than the critical values, indicating that the distribution of the estimated number of crashes approximate the distribution of actual numbers of crashes. Figure 2 shows the distribution of the number of sites against the number of crashes at site.

The actual number of crashes in the three-year sample of the validation dataset was 204 and that for the five-year sample was 396 crashes. 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. Even when the predictions and the actual observed data do not always match, real conditions cannot be ignored.
Empirical Bayes (EB) Model

The EB model was used to estimate the total number of crashes based on predicted values and on actual values. Since actual data are involved, the EB model also uses a dispersion parameter to weight the predicted values. For more discussions on EB models, see HCM (1), Knecht (8) and Hauer (21). If the predicted values are over-dispersed, then the model gives more weight to the actual number of crashes in the calculations. The dispersion parameter for the three-year sample was found to be 0.6491 and the dispersion parameter for the five-year sample was found to be 0.5755 as shown in Table 2 and Table 3. This means that the three-year sample data were more dispersed than the five-year sample data. The actual weights and models for the EB model vary for each segment. The only values that remain constant for each segment are the dispersion parameters.

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

CONCLUSION
The purpose of this research was to use historical data to develop crash prediction models for curved segments of rural two-lane two-way highways in Utah. The modeling was accomplished by calibrating the HSM crash prediction model as well as by developing Utah-specific models. The data came from 2008-2012 datasets, grouped into a three-year sample from 2010-2012 and a five-year sample from 2008-2012. The HSM predictive model calibration followed the HSM predictive model, including the use of appropriate CMFs as described in the HSM (1). 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 calibration of the HSM predictive model for curved segments on rural two-lane two-way highways in Utah was completed for the three-year sample and the five-year sample for comparison. The combined dataset contained 1,495 curved segments throughout the state. The three-year sample had a calibration factor of 1.50 and the five-year sample had a calibration factor of 1.60. The HSM model is under-predicting the number of crashes (i.e., curved segments of Utah’s rural two-lane two-way highways have 50 to 60 percent more crashes on average than a national dataset of rural two-lane two-way highways).

The Utah-specific models were developed using NB models. The use of a backward stepwise technique identified only four variables as statistically significant at a 95 percent confidence level, including AADT, segment length, curve radius, and total truck percentage. Where the HSM predictive model uses up to 12 variables in the CMF calculations in addition to the AADT and segment length, the Utah-specific models require only four variables in total. This simplified crash prediction model will be easier to reproduce due to the small amount of data collection required for its use. However, it is true that the CMF can reflect changes in improvements in those parts of the highways.

An EB model was also used to determine an expected number of crashes. The EB model relies on a combination of predicted values and actual values. The two values are weighted and added together to provide the overall result. The weight is dependent on a dispersion parameter. These dispersion parameters were obtained during model development so that future analysis can be performed.

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

ACKNOWLEDGMENTS
This research was made possible with funding from UDOT. Special thanks to W. Scott Jones, Tim Taylor, Travis Jensen, and the members of the UDOT Traffic & Safety Committee who played key roles as members of the Technical Advisory Committee. The contents of this paper reflect the
views of the authors, who are responsible for the facts and accuracy of the information presented herein, and are not necessarily representative of the sponsoring agency.

REFERENCES


10. Ellsworth P. *Utah DOT Leveraging LiDAR for Asset Management Leap*. Utah Department of Transportation. 


