RESEARCH

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

Report No. UT-12.06

TRAFFIC & SAFETY STATEWIDE MODEL AND GIS MODELING

Prepared For:

Utah Department of Transportation
Traffic & Safety, Research Divisions

Submitted By:

Brigham Young University
Dept. of Civil & Environmental Engineering

Authored By:

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

E. Scott Johnson, EIT
Graduate Research Assistant

Clancy W. Black, EIT
Undergraduate Research Assistant

Devin Francom
Undergraduate Research Assistant

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

July 2012

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 and the US 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.

UDOT RESEARCH REPORT ABSTRACT

1. Report No. **UT-12.06**	**2. Government Accession No.**	**3. Recipients Catalog No.**
4. Title and Subtitle TRAFFIC & SAFETY STATEWIDE MODEL AND GIS MODELING	**5. Report Date** July 2012	
	6. Performing Organization Code:	
7. Author(s) Grant G. Schultz, Ph.D., P.E., PTOE; E. Scott Johnson, EIT; Clancy W. Black, EIT; Devin Francom; Mitsuru Saito, Ph.D., P.E.	**8. Performing Organization Report No.**	
9. Performing Organization Name and Address: Brigham Young University Department of Civil & Environmental Eng. Provo, UT 84602	**10. Work Unit No.**	
	11. Contract No. 12-8471	
12. Sponsoring Agency Name and Address: Utah Department of Transportation Traffic & Safety Division 4501 South 2700 West Salt Lake City, UT 84114-8410	**13. Type of Report and Period Covered** **DRAFT**	
	14. Sponsoring Agency Code Project No. F-ST99(158) PIN No. 10304	

15. Supplementary Notes:
Prepared in cooperation with the Utah Department of Transportation

16. Abstract

Several steps have been taken over the past two years to advance the Utah Department of Transportation (UDOT) safety initiative. Previous research projects began the development of a hierarchical Bayesian model to analyze crashes on Utah roadways. Development of this statistical model continued in this project to analyze all state roadways by functional classification and consider additional variables including annual average daily traffic (AADT), vehicle miles traveled (VMT), speed limit, percent trucks, number of lanes, road geometry, road surface type, land use, and crash type. The model analyzes roadway segments and determines a posterior predictive distribution, or a distribution of the number of crashes that would be expected for that segment based on the number of crashes reported on other segments with the same characteristics (e.g., functional classification). The actual number of crashes for each segment is compared to the predictive distribution by calculating a percentile. A high percentile indicates more crashes than would be expected and a low percentile indicates less. The actual numbers of crashes are also compared to the mean of the predictive distribution to illustrate how many crashes above or below the estimate have occurred on that segment.

In addition to the statistical model a Geographic Information System (GIS) framework was developed to facilitate the analysis. The GIS framework has the capability to format the raw data obtained from UDOT such that it can be read into the statistical model. The GIS framework also displays the numerical data output by the statistical model spatially, allowing for an easy and intuitive analysis by UDOT staff.

17. Key Words Transportation safety, Bayesian models, safety mitigation, Geographic Information Systems (GIS)	**18. Distribution Statement** Available from: Utah Department of Transportation Research Division Box 148410 Salt Lake City, Utah 84114-8410 Brigham Young University Department of Civil & Environmental Engineering 368 Clyde Building Provo, Utah 84602

19. Security Classification (Of this report) Unclassified	**20. Security Classification (Of this page)** Unclassified	**21. No. of Pages** 162	**22. Price**

ACKNOWLEDGMENTS

This research was made possible with funding from the Utah Department of Transportation (UDOT) and Brigham Young University (BYU). Special thanks to the several professionals at UDOT and BYU who played key roles as members of the Technical Advisory Committee (TAC). The members of the TAC were:

- Robert Hull – UDOT Traffic & Safety;
- W. Scott Jones – UDOT Traffic & Safety;
- Tim Taylor – WCEC/UDOT Traffic & Safety;
- Travis Jensen – WCEC/UDOT Traffic & Safety;
- Members of the UDOT Traffic & Safety Committee;
- Grant Schultz – BYU Associate Professor;
- Mitsuru Saito – BYU Professor;
- E. Scott Johnson – BYU Graduate Student;
- Clancy W. Black – BYU Undergraduate Student;
- Devin Francom – BYU Undergraduate Student; and
- C. Shane Reese – BYU Professor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Several steps have been taken over the past several years to advance the Utah Department of Transportation (UDOT) safety initiative. Previous research projects evaluated and introduced some of the shortcomings of traditional before and after analysis methods commonly used to analyze automobile crashes by noting that these methods are limited in that they do not account for the mean and variance of the data not being equal (as in a Poisson regression) or they do not account for regression to the mean (RTM) bias. The previous research reports outlined the advantages of Bayesian methods, including the Empirical Bayes (EB) method, which accounts for both the difference between the mean and variance and RTM, but can be complicated to employ and has limitations of its own. A second Bayesian method, the hierarchical Bayesian method was then utilized to develop a model to analyze crashes on Utah roadways.

The hierarchical Bayes model developed in the previous research projects was advanced as part of this research to analyze all state roadways by functional classification and consider additional variables that may contribute to safety including annual average daily traffic (AADT), vehicle miles traveled (VMT), speed limit, percent trucks, number of lanes, road geometry, road surface type, land use, and crash type. The model was developed to analyze roadway segments and determine a posterior predictive distribution, or a distribution of the number of crashes that would be expected for that segment based on the number of crashes reported on other segments with the same characteristics (e.g., functional classification). The actual number of crashes for each segment is compared to the predictive distribution by calculating a percentile. A high percentile indicates more crashes than would be expected and a low percentile indicates fewer crashes than would be expected. The actual numbers of crashes are also compared to the mean of the predictive distribution to illustrate how many crashes above or below the estimate have occurred on that segment.

In addition to the statistical model, a Geographic Information System (GIS) framework was developed as part of this research to facilitate the analysis. The GIS framework has the

capability to format the raw data obtained from UDOT such that it can be read into the statistical model. The GIS framework also displays the numerical data output by the statistical model spatially, allowing for an easy and intuitive analysis by UDOT staff. The GIS framework was developed to prepare the data for analysis by creating segments based on three characteristics: 1) functional classification; 2) AADT (converted to VMT to account for segment length); and 3) speed limit. After the segmented data were analyzed by the model, the GIS framework provided a method to display the results for each segment on a color scaled map allowing for easy identification of hotspots using contrasting colors as illustrated in Figure ES-1. A sample analysis was presented to demonstrate how the method could be applied for a safety study on hotspots. This will allow staff at UDOT to accurately evaluate the safety needs of roadways in the state.

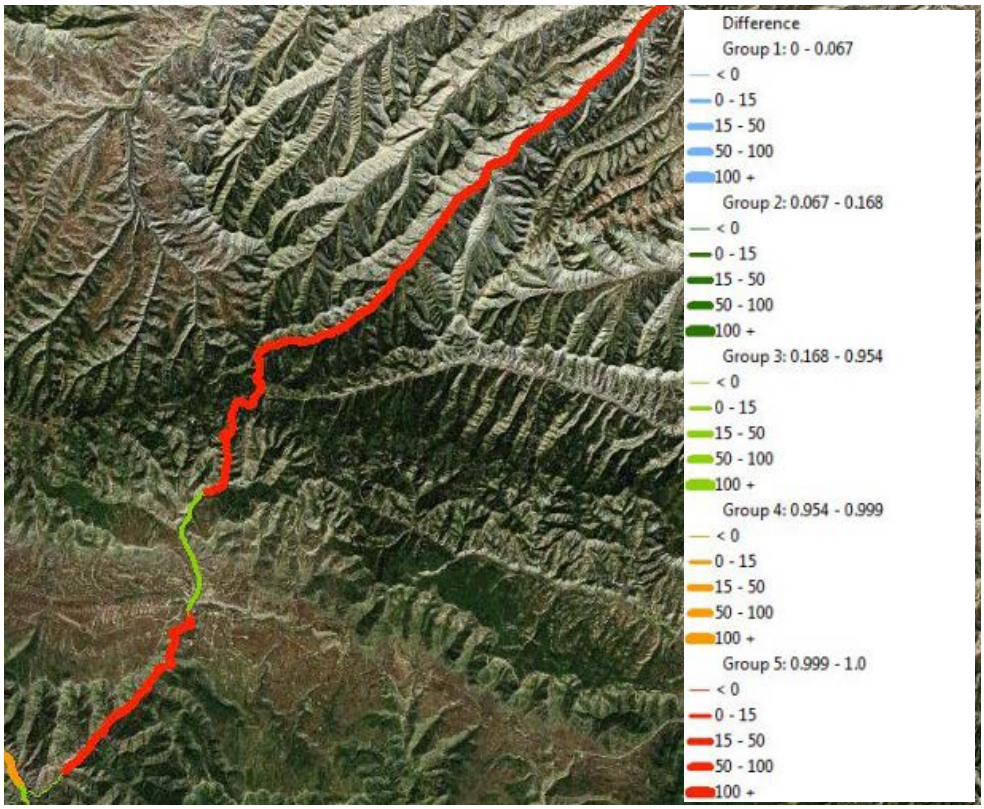


Figure ES-1. Sample model output.

1 INTRODUCTION

The Utah Department of Transportation (UDOT) Traffic & Safety Division continues to advance the safety of roadway sections throughout the state. UDOT has continually placed safety at the forefront of their priorities and continues to develop and publicize the "Zero Fatalities: A Goal We Can All Live With™" campaign to increase awareness of the importance of highway safety. UDOT has also continued at the forefront of research and education through their active participation and membership in the Transportation Research Board (TRB) Highway Safety Performance Committee and their willingness to invest in safety research. The Federal Highway Administration (FHWA) and the American Association of State Highway and Transportation Officials (AASHTO) are also continually working to aid states in safety analysis, primarily with the release of the AASHTO *Highway Safety Manual* (HSM) (AASHTO 2010) to aid in the analysis of transportation safety data.

1.1 Background

Several steps have been taken over the past two years to advance the UDOT safety initiative, which has included contracting for a research projects entitled *Analyzing the Effectiveness of Safety Measures using Bayesian Methods*, hereafter referred to as Volume 1 (Schultz et al. 2010); *Calibration of the Highway Safety Manual and Development of New Safety Performance Functions*, hereafter referred to as Volume 2 (Saito et al. 2011); and *Framework for Highway Safety Mitigation and Workforce Development*, hereafter referred to as Volume 3 (Schultz et al. 2011). These three projects included tasks to: evaluate existing traffic safety data and analysis (Volume 1); calibrate HSM models (focused on two-lane, two-way roadways) (Volume 2); develop a framework for safety mitigation (HSM based) (Volume 3); develop a basis for low-cost safety improvement prioritization to set the stage for future research (Volume 3); and establish background for safety workforce development training (Volume 3). This

research was established specifically to explore ways to utilize crash data more effectively, to develop a methodology for crash data analysis using advanced statistical methods (e.g., hierarchical Bayes models), to begin the process of calibrating roadways consistent with the HSM, and to establish a basis for future studies.

The research conducted in the Volume 1 through 3 reports was followed up in 2011 with two specific work task orders to help meet the need of more accurately and efficiently analyzing the safety of Utah's roadways. The first was developed to expand upon the previous research to develop a statistical model of traffic crashes and safety by functional classification on UDOT roadways statewide. The second was to begin the process of combining the traffic safety models with Geographic Information System (GIS) capabilities as a first phase approach to scope out the possibilities of GIS modeling and to develop a framework and prototype for future development. This analysis will help to ensure that crash mitigation efforts are implemented where the need is greatest so that public funds are used where they will be of the most benefit.

1.2 Objectives

The objective of this research is to advance the level of safety research in the state of Utah by building upon preliminary research to provide UDOT with the necessary tools to go beyond today and address the future of the system and the needs for tomorrow. This research builds upon previous research begun to develop a statewide model and to complete the first phase scoping of a GIS framework for safety modeling.

The statewide model developed through this research will continue to advance the analytical model developed for safety analysis in the state by expanding the coverage to include all state roadways by functional classification, while also exploring additional variables for analysis, including annual average daily traffic (AADT), vehicle miles traveled (VMT), speed limit, percent trucks, number of lanes, road geometry, road surface type, land use, and crash type.

The GIS framework developed in this research is part of a continuing effort to explore alternatives in combining traffic safety reporting and GIS capabilities to better visualize safety in the state. A geographic representation of crashes, their locations, and the safety aspects of the crashes will allow the Traffic & Safety Division to focus their efforts on those areas most affected by crashes.

1.3 Organization

This report is organized into the following chapters: 1) Introduction; 2) Literature Review of Traffic Safety Analysis; 3) Literature Review of GIS; 4) Data; 5) Statistical Model; 6) Framework for GIS-based Crash Data Analysis; 7) Conclusion. A list of references, list of acronyms, and Appendices follow the indicated chapters.

Chapter 2 is a literature review outlining the state of the practice in traffic safety analysis including defining traffic safety and techniques for analyzing and mitigating crashes.

Chapter 3 is a literature review outlining GIS tools as well as adaptations of GIS utilized by other states and organizations to analyze automobile crashes. GIS tools include linear referencing and dynamic segmentation, data visualization methods, and other analysis tools. The GIS adaptations discussed were developed by various universities and government agencies to meet each one's specific needs.

Chapter 4 discusses the data used in this analysis. The methods used to collect and store roadway data are discussed as well as necessary steps taken to prepare the data for analysis and the importance of data uniformity.

Chapter 5 discusses the statistical model developed to analyze the data. A framework for the statistical model had been developed as part of a previous research project. The model code was modified to analyze more variables and to accommodate greater ease of use. The theoretical background, basis for development, and an explanation of the model output are discussed.

Chapter 6 discusses the development of the GIS framework. Using ArcGIS, the results from the statistical model can be interpreted graphically and displayed spatially allowing for an improved interpretation of the results.

Chapter 7 presents the conclusions of this research as well as recommendations for future research in this area.

2 LITERATURE REVIEW OF TRAFFIC SAFETY ANALYSIS

A literature review has been performed on methods used to analyze traffic safety. This chapter gives the reader background on traffic safety, and a summary of common crash analysis techniques. A summary section is also provided at the end of the chapter. This literature review is provided as a general introduction to traffic safety analysis and does not include the details of traffic safety analysis that are included in the Volume 1 through Volume 3 series. For a more detailed approach to traffic safety analysis, the reader is referred to the literature reviews of these three reports (Saito et al. 2011, Schultz et al. 2010, Schultz et al. 2011).

2.1 Traffic Safety

This section discusses why traffic safety is important, how traffic safety is defined, and how it is measured.

2.1.1 The Importance of Traffic Safety

The goal of the transportation industry is to promote the movement of people and goods in a safe and efficient manner. Researchers have developed ways to evaluate safety on our roadways and identify ways to mitigate unsafe conditions. In the United States, each state is required by the FHWA as part of the Highway Safety Improvement Program (HSIP) to submit an annual report describing not less than 5 percent of their highway locations exhibiting the most severe safety needs, as well as develop a plan to remedy those hazardous locations, all in an effort to provide a safe and efficient transportation network (FHWA 2011).

The HSM provides recommendations for states to evaluate and take steps to increase highway safety. States can also develop their own methods to evaluate safety on their roadways. All of this is an effort to promote safety and reduce the number of fatalities and serious injuries on the nation's roadways.

2.1.2 Defining Safety

Traffic safety is defined by the HSM as "the crash frequency or crash severity, or both, and collision type for a specific time period, a given location, and a given set of geometric and operational conditions" (AASHTO 2010, p. 3-1). The HSM also defines crashes as "a set of events not under human control that results in injury or property damage due to the collision of at least one motorized vehicle and may involve collision with another motorized vehicle, a bicyclist, a pedestrian or an object" (AASHTO 2010, p. 3-3).

Safety is often quantified or measured by tracking the raw number (frequency) of fatalities, injuries, or crashes, or by calculating injury, fatality, or crash rates which are crash frequencies normalized for exposure as shown in Equation 2-1. Each method has its advantages and disadvantages that depend on the intended use and audience of the data (Herbel et al. 2010). Crash rates are most often measured by the number of crashes occurring per million vehicle miles traveled (MVMT) for roadway segments, or crashes per million entering vehicles (MEV) for intersections (AASHTO 2010). Equation 2-2 shows the crash rate equation for a section of roadway (Roess et al. 2004).

$$Crash\ Rate = \frac{Average\ Crash\ Frequency\ in\ Period}{Exposure\ During\ Period} \quad (2\text{-}1)$$

$$CR_{sec} = \frac{N}{V_{sec} \times 365 \times L} \times 10^6 \quad (2\text{-}2)$$

where: CR_{sec} = crash rate for section (in crashes per MVMT),
N = number of crashes per year,
V_{sec} = average annual daily traffic (AADT) of road section, and
L = length of section (in miles)

Equation 2-3 shows the crash rate equation for intersections (Roess et al. 2004).

$$CR_{int} = \frac{N}{V_{int} \times 365} \times 10^6 \quad (2\text{-}3)$$

where: CR_{int} = crash rate for intersection (in crashes per MEV), and
V_{int} = sum of average daily approach volumes of intersection.

2.2 Highway Safety Manual

The HSM was developed as a tool to help practitioners build, analyze, and modify roadways to increase traffic safety by reducing the number and severity of crashes that occur (AASHTO 2010). Over the years standards have been developed for roadway design but these standards do not always meet the needs of every roadway network. Roads, their environments, and the people that use them vary from region to region, as do the factors that can best be modified to reduce the number of crashes that occur on those roadways. The HSM uses three statistical tools to analyze traffic safety: 1) safety performance functions (SPFs); 2) crash modification factors (CMFs), and 3) calibration factors.

2.2.1 Safety Performance Functions

SPFs are statistical base models that are used to estimate the average crash frequency for a facility type with certain base conditions as a function of AADT and segment length (in the case of roadway segments). Base conditions may include factors such as lane width, presence or absence of lighting, presence of turn lanes, etc. An example of an SPF (for roadway segments on rural two-lane highways) is shown in Equation 2-4 (AASHTO 2010).

$$N_{SPF} = (AADT) \times (L) \times (365) \times 10^{-6} \times e^{-0.4865} \quad (2\text{-}4)$$

where: N_{SPF} = estimated crash frequency for the given conditions (crashes/year),
$AADT$ = average annual daily traffic (AADT) of road section, and
L = length of section (in miles)

The SPFs developed in the HSM can be calibrated to imitate local conditions or agencies with sufficient expertise may develop SPFs unique to their jurisdiction (Saito et al. 2011).

2.2.2 Crash Modification Factors

The CMF is a ratio of the expected crash frequencies associated with two different conditions and may serve as an estimate of the effectiveness of a specific type of design, control feature, or treatment. The calculation of a CMF is illustrated in Equation 2-5. "Condition a"

represents the set of base conditions for a particular site. "Condition b" represents the conditions of the same site but where one characteristic or condition of the site has been modified or a treatment has been applied. CMFs represent the relative change in crash frequency due to a change in one specific condition when all other conditions and site characteristics remain constant (AASHTO 2010).

$$CMF = \frac{Expected\ Average\ Crash\ Frequency\ with\ Site\ Condition\ b}{Expected\ Average\ Crash\ Frequency\ with\ Site\ Condition\ a} \quad (2\text{-}5)$$

Under base conditions the value of a CMF is 1.00. CMF values less than 1.00 indicate that the alternative treatment reduces the estimated crash frequency for the facility when compared to the base conditions. Conversely, a CMF value greater than 1.00 indicates that the alternative treatment increases the estimated crash frequency when compared to the base conditions. A CMF value equal to 1.00 indicates that the treatment or modification had no effect on the average crash frequency (AASHTO 2010).

The CMF can also be used to determine the expected percentage reduction (or increase) in crash frequency using Equation 2-6 (AASHTO 2010):

$$Percent\ Reduction\ in\ Crashes = 100 \times (1.0 - CMF) \quad (2\text{-}6)$$

2.2.3 Calibration Factors

Crash rates, even under very similar conditions, can vary from region to region. SPFs can be calibrated to reflect the differing crash frequencies in different locations. Calibration can be undertaken for a single state or for smaller regions within a state where appropriate. This is done by multiplying the SPF by a calibration factor. Calibration factors can be calculated using Equation 2-7 (AASHTO 2010).

$$C_i = \frac{\sum observed\ crashes}{\sum predicted\ crashes} \quad (2\text{-}7)$$

where, C_i = local calibration factor for site type *i*.

2.3 Crash Analysis Techniques

This section discusses before and after techniques for crash analysis; regression to the mean (RTM) bias and how not accounting for RTM can lead to incorrect estimates of traffic safety; the limitations of ordinary least squares (OLS) regression and Poisson estimation as crash prediction methods; and the use of Bayesian methods in predicting crashes.

2.3.1 Before and After Studies

As discussed in Volume 1, traditional before and after studies have historically been used to analyze crash statistics. In a before and after study the count of crashes before a treatment is compared to the count of crashes after the treatment is applied. The effectiveness of the treatment is measured by the difference between crashes predicted by the before count and the actual number of crashes after the treatment. The underlying assumption in before and after studies is that no other influence aside from the applied treatment had an effect on the number of crashes at that location. In reality, many other factors including conditions that change naturally over time (traffic, weather, road user behavior), other treatments that may have been implemented during the before or after periods, or changes in crash reporting requirements could be associated with a change in crash counts (Hauer 1997).

2.3.2 Regression to the Mean

Crashes by nature are random events and crash frequencies naturally fluctuate over time. Fluctuations in crash frequency make it difficult to determine whether a change in the number of crashes is the result of a specific treatment or the result of natural fluctuations due to the random nature of crashes. This is especially true when studying crashes over a short period of time due to the fact that it is impossible to know if the short observations accurately depict the long term behavior of the site. These fluctuations in crash counts make it difficult to determine whether a reduction in crashes is a result of a particular treatment, changes in site conditions, or because of fluctuations over time due to the random nature of crashes. This is referred to as RTM bias (AASHTO 2010).

The RTM phenomenon expects that a value that is determined to be extreme will tend to regress to the long-term average over time as illustrated in Figure 2-1. This means that it is

statistically probable that a period of high crash frequency at a site will be followed by a period of low crash frequency (Hauer 1997). RTM bias refers to the selection of a site as a result of the short-term trend it exhibits, thus not taking into account the RTM.

In Figure 2-1 the observed crash frequencies for a site are plotted over several years. The long term average line represents the actual behavior of the site. The short term average lines represent the perceived actual behavior of the site if short term windows are used in the estimation. This illustrates how the crash frequency estimation could be higher or lower than the actual average crash frequency because of the time period selected for analysis (Schultz et al 2010).

Not accounting for RTM bias could result in an overestimation or underestimation of the effectiveness of a treatment. This makes it impossible to know if an observed reduction in crashes is due to a specific treatment or if it is simply a natural fluctuation in crash frequency over time (AASHTO 2010). Figure 2-2 shows the difference between the perceived reduction in crashes when RTM bias is not accounted for and the actual reduction in crashes when RTM is accounted for. Failing to account for RTM bias is one of the primary limitations with many current safety analysis practices (Schultz et al. 2010).

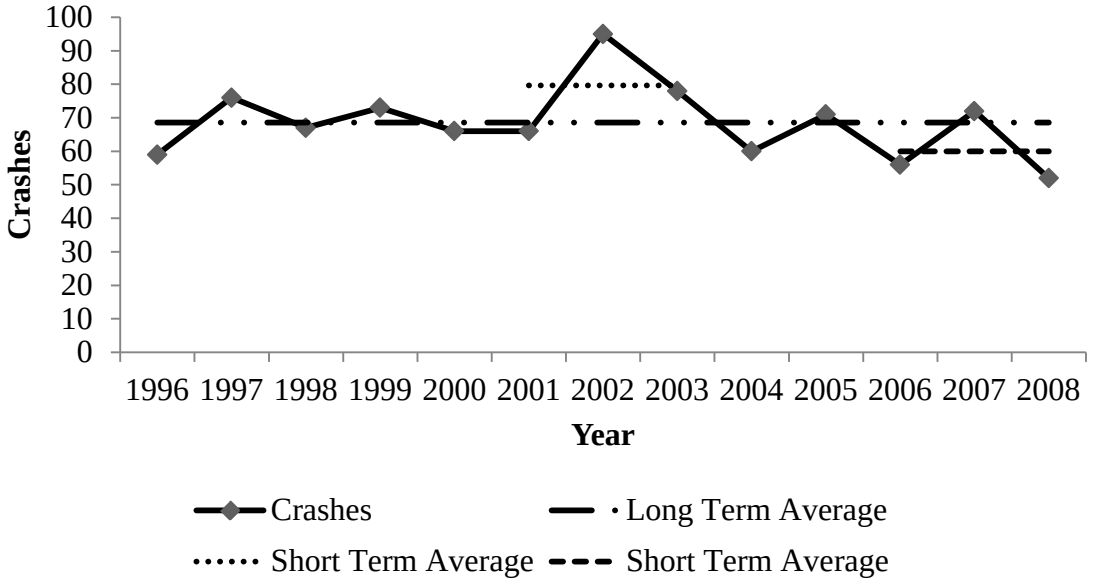


Figure 2-1. Variation in short-term and long-term crash frequency (Schultz et al. 2010).

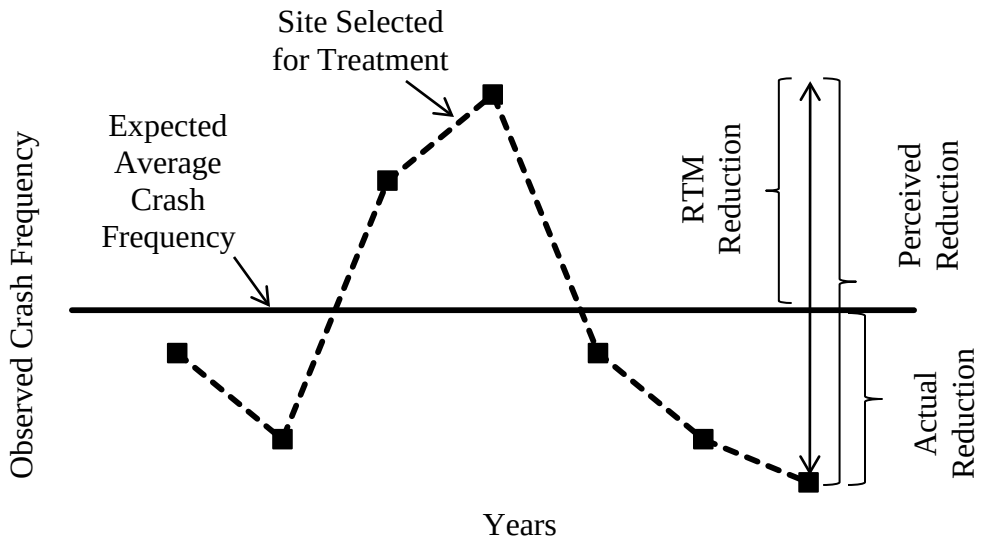


Figure 2-2. Perceived vs. actual reduction (adapted from AASHTO 2010).

2.3.3 Ordinary Least Squares Regression and Poisson Estimation

Another limitation of many current safety analysis practices is overdispersion. Two methods that have historically been used for crash prediction are OLS regression and Poisson estimation (Hadi et al. 1995, Strathman et al. 2001). An underlying assumption of OLS estimation is that crash frequency is normally distributed. This assumption is rarely satisfied as crash frequencies are typically skewed toward zero corresponding to a Poisson distribution (Jovanis and Chang 1986).

Poisson estimation requires the mean and variance of the crash frequency to be equal. Often times the variance will exceed the mean which is referred to as overdispersion. Even though Poisson estimation is still unbiased when crash frequencies are overdispersed, the errors of the parameter estimates tend to be understated. When this happens parameters may be interpreted as statistically significant when they are not (Strathman et al. 2001). To overcome the obstacle of overdispersion researchers in Connecticut employed Poisson regression models using quasi-likelihood estimation techniques. Quasi-likelihood estimation, as implemented in the S-Plus statistical package, accounts for over or underdispersion in the count observations by estimating the over or underdispersion parameter as part of the process (Ivan et al. 2000).

2.3.4 Empirical Bayes Method

The effects of overdispersion and RTM were well documented by Hauer, who derived an Empirical Bayes (EB) approach for estimating the true mean crash rate for a location (Hauer 1997). The EB method estimates the true mean crash rate through a mathematical combination of the predicted and observed crash frequencies accounting for both overdispersion and RTM. Equation 2-8 is used to estimate the expected crash frequency. The weight, *w*, used in Equation 2-8, is calculated by combining the model's overdispersion parameter (φ) and N_{spf}, (discussed in section 2.2.1) as shown in Equation 2-9 (Saito et al. 2011).

$$N_{expected} = w \times N_{spf} + (1 - w) \times N_{observed} \quad (2\text{-}8)$$

where, $N_{expected}$ = expected number of crashes determined by the EB method,
w = weight determined by Equation 2-9, and
$N_{observed}$ = observed number of crashes at a site.

$$w = \frac{1}{1+\varphi\times(N_{spf})} \quad (2\text{-}9)$$

where, φ = overdispersion parameter.

2.3.5 Hierarchical Bayes Method

Qin et al. (2005) note that Tunaru improved upon Hauer's method using a hierarchical Bayesian generalized linear modeling approach for multiple crash response at a location. A major criticism of the EB approach is that it is unable to incorporate uncertainties in the model parameters. The EB method assumes that the parameters are error free and can be replaced easily by their posterior analysis estimates (Schultz et al. 2010). These limitations can be overcome by using the flexible modeling associated with the hierarchical Bayesian method (Sloboda 2009).

In a hierarchical Bayesian analysis prior information and all available data are integrated into posterior distributions from which inferences can be made thus accounting for all uncertainties in the analysis. The hierarchical Bayesian method may be less costly to implement

and may result in safety estimates that have more realistic standard errors. The hierarchical Bayesian approach has several advantages over the EB approach such that it is believed to require less data for untreated reference sites, it better accounts for uncertainty in crash data, it provides more detailed casual inferences, and it offers more flexibility in selecting crash count distributions (Schultz et al. 2010). The hierarchical Bayesian method will be discussed in more detail in Chapter 5.

2.4 Chapter Summary

States are required by law to analyze traffic safety and take steps to mitigate crashes. Safety is quantified by measuring the number and/or severity of at a given location. The HSM was developed to assist states in analyzing traffic safety and selecting appropriate countermeasures. States are encouraged to customize the methods set forth in the HSM to meet local needs.

Traditional before and after studies used to analyze crash statistics tend to produce skewed results due to the fact that they do not account for RTM. Crashes by nature are random events and crash rates naturally fluctuate over time. RTM phenomenon expects that a period experiencing high frequency will be followed by a period of low crash frequency.

OLS and Poisson regression are two common methods used in crash predictions. These methods rely on assumptions about the distribution of the crash frequency that are rarely met. Hauer developed an EB approach to crash analysis that is robust to these assumptions and also accounts for RTM producing a more accurate estimate of the true mean crash rate for a location. Hierarchical Bayesian is considered to be an improvement over the EB approach because it requires less data, better accounts for uncertainty in the data, provides more flexibility in selecting crash count distributions than the EB approach, and thus may result in more realistic estimates of safety conditions.

3 LITERATURE REVIEW OF GEOGRAPHIC INFORMATION SYSTEMS

This literature review has been prepared as a summary of existing research and progress regarding the application of GIS in crash data and roadway safety analysis. Research has included the full range of available literature including peer-reviewed journals, technical publications, public agency reports, and examination of GIS technologies and programs used in crash data analysis. This chapter will focus on Esri ArcGIS software, linear referencing and dynamic segmentation methods in GIS, crash data visualization tools, data analysis tools, a discussion of existing GIS-enabled crash data analysis programs, and a summary of key features in GIS programs.

3.1 Software

ArcInfo Version 10.0, Esri's ArcGIS Desktop software package, was used as the primary platform for this literature review and research. ArcInfo is the highest license level available from Esri. The two products used for this research were ArcMap and ArcCatalog (Esri 2012a).

ArcMap is Esri's base GIS platform, and is generally accepted as the standard for geographic data processing and analysis. ArcMap offers the user a display of geographic data in a map setting along with the ability to view the data in tabular format. ArcMap also includes hundreds of tools for performing analysis of geographic data including spatial analyst, spatial statistics, proximity, overlay, linear referencing, and network analyst. ArcMap also allows creation of attractive map layouts for presenting data and results. Many different features and tools in ArcMap will be used and discussed in this research (Esri 2012a).

ArcCatalog is a data management application, similar to Explorer in Windows operating systems. ArcCatalog facilitates data storage and organization and allows files to be transferred to different formats. ArcCatalog was used primarily to upload raw data and prepare it for use in ArcMap (Esri 2012a).

Esri is very active in developing new applications of its software. Many of the programs discussed in this chapter are based on custom tools or application available through Esri.

3.2 Linear Referencing and Dynamic Segmentation

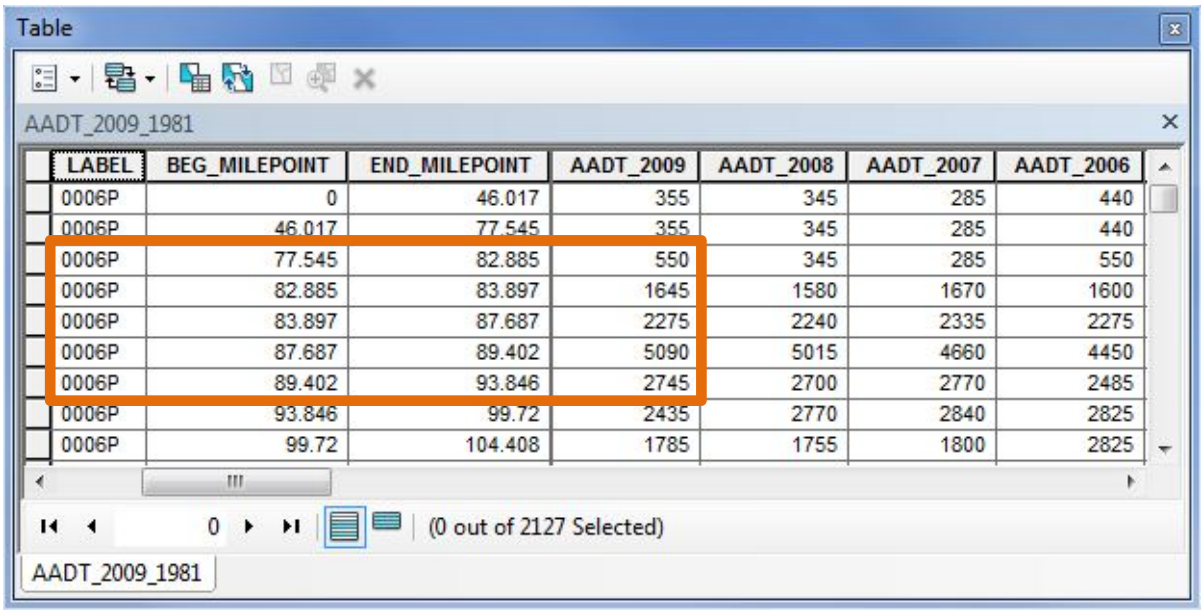
Linear referencing tools are an extension available with ArcInfo for use in ArcMap. They are designed specifically for transportation or other linearly-based data such as pipelines or transmission lines. These tools are invaluable to crash data analysis and were used throughout the research. This section will explain their use and how they apply to crash data analysis. Dynamic segmentation, a useful tool when working with linear referencing data, will also be described.

3.2.1 Linear Referencing Systems

Linear referencing (referred to as "Linear Referencing System" or LRS) is a GIS data management system built around the route-milepoint system used by most transportation agencies. Instead of features being located by a latitude-longitude (lat/long) or other coordinate system reference, features are located by specifying a distance (milepoint) along a measured line (route). A linear referenced line is a polyline that has measures associated with it. The line will have a beginning milepoint (usually zero) and an ending milepoint. All locations on the line between those points can then be identified by an intermediate milepoint value that is somewhere between the beginning and ending milepoint. Non-LRS lines in GIS can be measured, but they do not have intermediate measures associated with any point on the line, the measure is a static value (Esri 2011c).

Linear referencing is useful in crash data analysis because it can represent point and line features of the roadway in the same geographic context. Linear attributes such as number of lanes, AADT, roadway width, speed limit, curvature, functional classification, and others can all be represented in a table with milepoints and the appropriate variable value for that segment. Figure 3-1 shows an example of an attribute table for linear referenced data in ArcMap. This table contains AADT information for state routes. The "LABEL" field serves as the route reference and the "BEG_MILEPOINT" and "END_MILEPOINT" fields indicate where that segment begins and ends. The AADT variables are then stored for that segment. When these data

are mapped in ArcMap the system will find the line that represents route 0006P then "draw" a line on it from milepoint 0 to milepoint 46.017 for the first record in the table. That line will then be associated with the AADT values shown. For the second record a line will be drawn from milepoint 46.017 to milepoint 77.545 and so on with each record after that. Figure 3-2 shows the base linear referenced routes representing the state highway system and Figure 3-3 shows the same routes after the AADT attribute file has been mapped using the records in the table in Figure 3-1. Each individual segment is represented by a unique color and is labeled with its respective AADT. Point data such as crash location is represented in LRS the same way except that there is only one milepoint listed in the table rather than two. When data are mapped using LRS the data is referred to as a "Route Event," signifying that it is an event that occurred on that route (crash) or a characteristic of that route (AADT). In Figure 3-3 the AADT route events have been mapped. LRS also facilitates adding new data relatively easily. A new record with attributes can be added to any Route Event Layer table and it will instantly be digitized on the map. No manual drawing or processing is necessary.



LABEL	BEG_MILEPOINT	END_MILEPOINT	AADT_2009	AADT_2008	AADT_2007	AADT_2006
0006P	0	46.017	355	345	285	440
0006P	46.017	77.545	355	345	285	440
0006P	77.545	82.885	550	345	285	550
0006P	82.885	83.897	1645	1580	1670	1600
0006P	83.897	87.687	2275	2240	2335	2275
0006P	87.687	89.402	5090	5015	4660	4450
0006P	89.402	93.846	2745	2700	2770	2485
0006P	93.846	99.72	2435	2770	2840	2825
0006P	99.72	104.408	1785	1755	1800	2825

Figure 3-1. An attribute table for LRS data (Boxed section represents data shown in Figure 3-3).

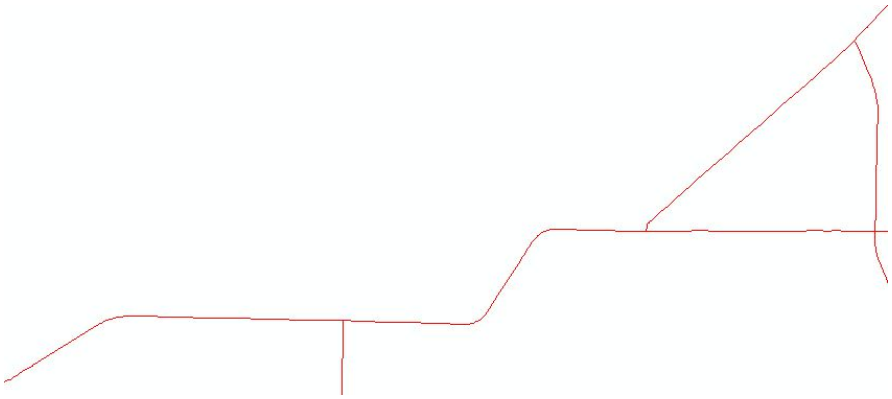


Figure 3-2. Linear referenced routes before attribute mapping.

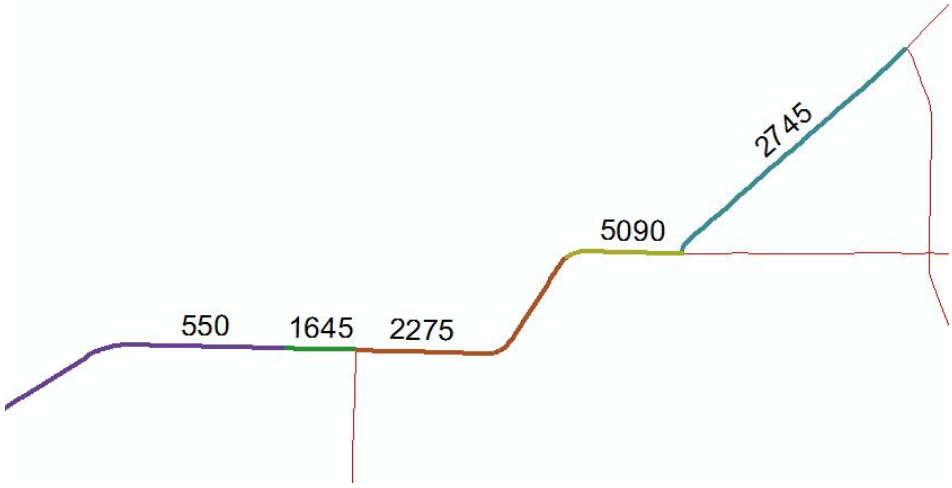


Figure 3-3. Linear referenced routes after mapping AADT data for a portion of the roadway.

3.2.2 Dynamic Segmentation

Dynamic segmentation is an LRS process that allows data to be overlaid in order to merge or collect attributes, a very useful operation in crash data analysis (Mitra 2009). Dynamic segmentation is a process that takes two different route event tables or layers and merges them into one file with milepoint locations adjusted accordingly. In addition, the attributes of each layer will both be present in the new file. Table 3-1 through Table 3-3 show a simple example of how dynamic segmentation works. The speed data and AADT data are combined and the output has new milepoint breaks and the appropriate data for each segment. The final product in Table 3-3 is an output dataset of segments that are each homogeneous with respect to the input data. This feature of LRS is extremely useful in crash analysis because it allows the creation of

segments with homogeneous characteristics. These segments and the associated characteristics can then be input into statistical models for further analysis. Overlay's can also be done with points and lines. The output is a point file that has added the attributes of the line to its attribute table, but there is no change in the milepoints. This is useful for determining crash variables that are not recorded in the crash record but are available in an LRS line file within GIS.

Table 3-1. LRS Speed Limit Data

BEG_MILEPOINT	END_MILEPOINT	SPEED
0	5	25
5	10	35

Table 3-2. LRS AADT Data

BEG_MILEPOINT	END_MILEPOINT	AADT
0	2.5	1000
2.5	10	1500

Table 3-3. Overlay of Speed and AADT Data (Dynamically Segmented)

BEG_MILEPOINT	END_MILEPOINT	SPEED	AADT
0	2.5	25	1000
2.5	5	25	1500
5	10	35	1500

Several other LRS tools are available in ArcMap. These include tools to create new routes (after new route construction), to calibrate or adjust route measures (after a route realignment), to transform a route from one reference to another, and to locate the milepoint or measure value of non-LRS data. These tools were not used extensively in the research and will not be described in this report (Esri 2012a).

3.3 Visualization of Crash and Safety Data Using GIS

GIS brings a spatial dimension to crash data, which helps analysts understand the crash in context of the roadway and environment. Adding the spatial reference to the crash can improve understanding of the factors involved in the incident and will help determine the most appropriate countermeasures (Miller 1999, Mitra 2009). This chapter will describe visualization techniques that have proven effective at representing crash data in GIS. This includes both basic and advanced visualization techniques along with the use of aerial and street-level imagery. Work done as part of this project will be used along with examples from the literature.

3.3.1 Basic Visualization

Visualization of crash data has a two-fold benefit. First, it offers unique ways of viewing data, providing information that may otherwise be unnoticed. Second, it serves as an effective tool for presenting information regarding crash safety to the public and decision makers in a way that it can be understood by someone who is otherwise untrained in safety engineering (Harkey 1999).

Basic visualization practices involve using colors and shapes to represent features based on the data (Graettinger et al. 2005; Roche 2000). Figure 3-4 shows how points representing crashes can be scaled to signify a feature of the crash. This method is often used to show crash severity, as is done in Figure 3-4, with larger points representing higher severity crashes. Other data commonly shown with this method may include number of occupants, vehicles involved in a crash, or crash cost. One advantage to scaled points is that it can show features that overlap, however it still does not show the total number of features because some of the points overlap. Linear features such as AADT and number of lanes can be represented using scaled lines as shown in Figure 3-5.

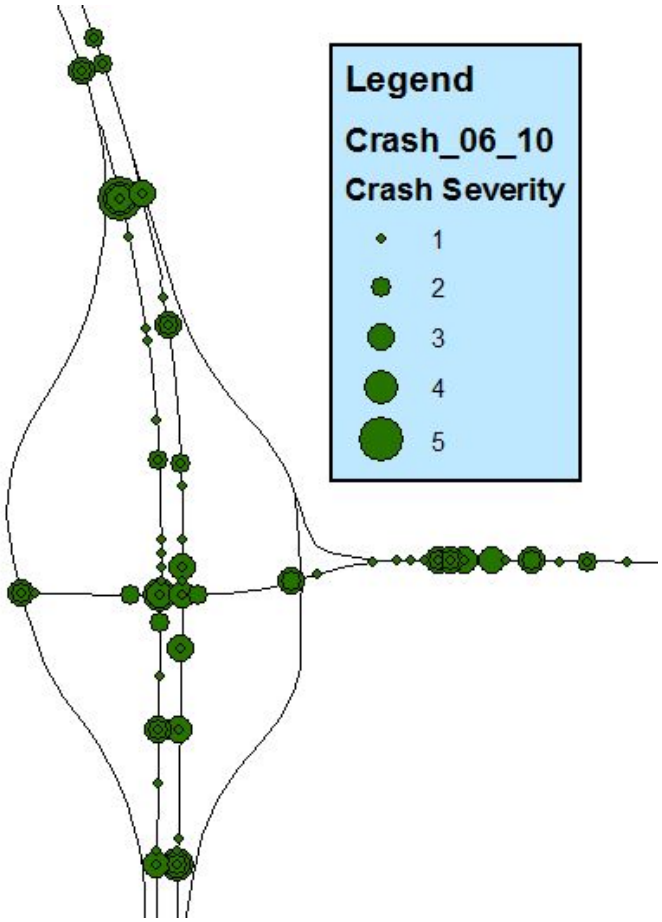


Figure 3-4. Crash severity symbolized by scaled points.

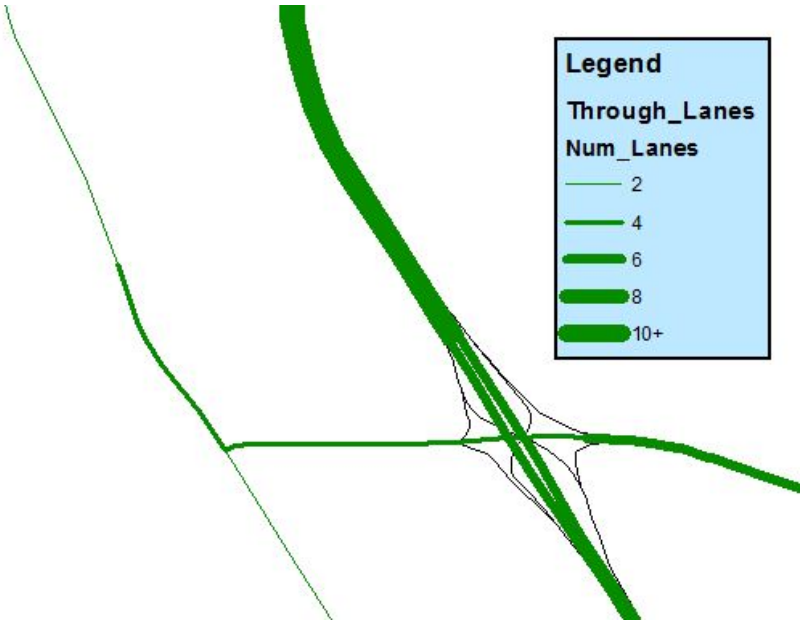


Figure 3-5. Number of lanes symbolized by scaled lines.

Color symbolization of points and lines is also used to represent relative values of crash and roadway data. Colors tend to be more effective at showing risk or impact comparisons because the contrast is more distinct than with scaled features. Figure 3-6 shows how colors have been used to show crash severity. This could also show crash cost, number of injuries, or speed at which crash occurred. Color-coding on roadway segments is commonly used to show relative risk or crash rates on segments as shown in Figure 3-7.

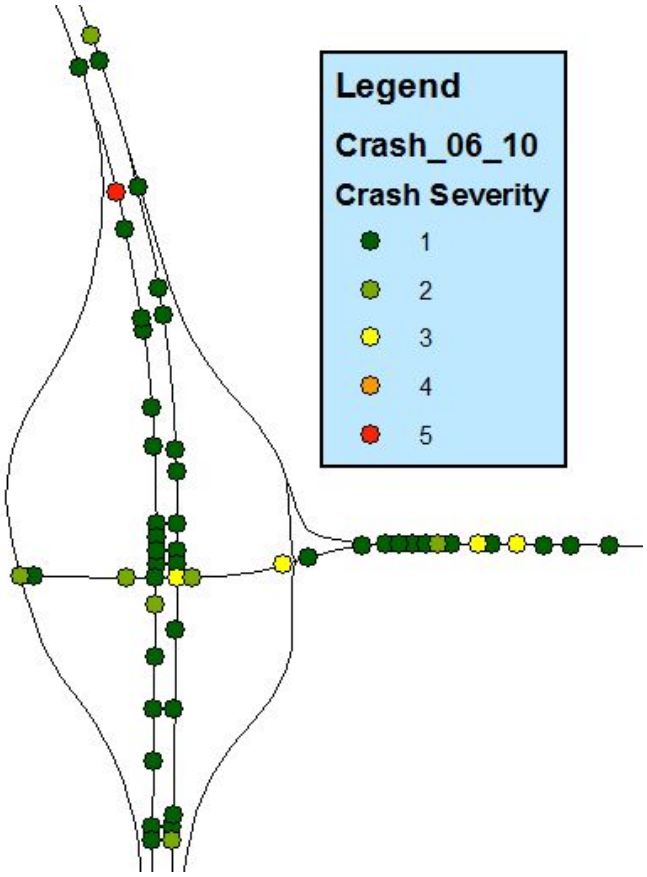


Figure 3-6. Crash severity symbolized by colored points.

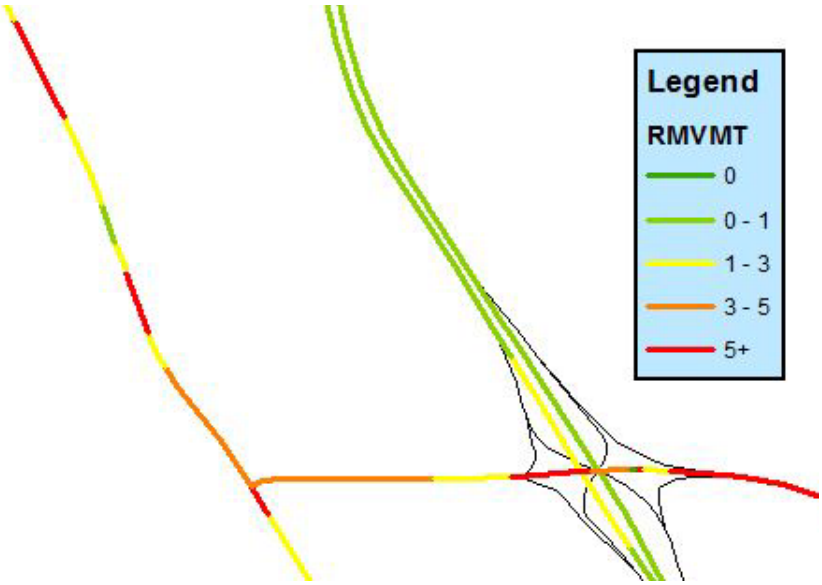


Figure 3-7. Crash rate in RMVMT symbolized by colored lines.

Figure 3-8 shows how color-coding and scaling can be used effectively together to show attributes of crashes and a roadway for comparison (Qin and Wellner 2011). In this example the narrow inner line represents analysis results using one set of segments, and the large outer line represents analysis results of another set of segments. This allows the analyst to see changes in road conditions and to compare different results or methods.

Figure 3-9 shows another way to compare data using offsets. When working with linear referenced data the data can be mapped offset from its actual location as shown by the colored lines, allowing the analyst to compare information from multiple datasets simultaneously. In Figure 3-9 two datasets, AADT and Speed Limit, have been mapped on either side of the roadway, with crashes shown on the centerline.

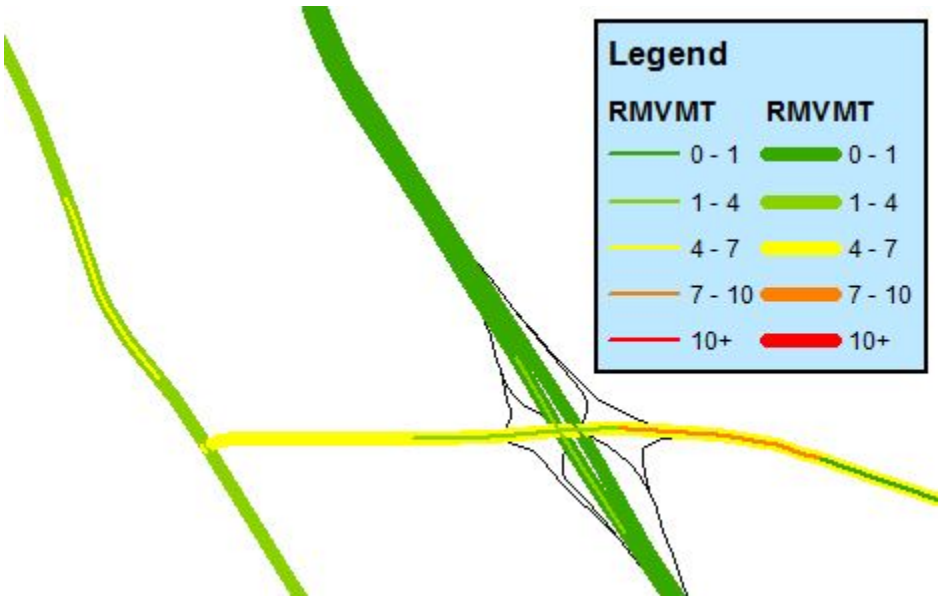


Figure 3-8. Example of how coloring and scaling can be used to show results of two analyses with different segments.

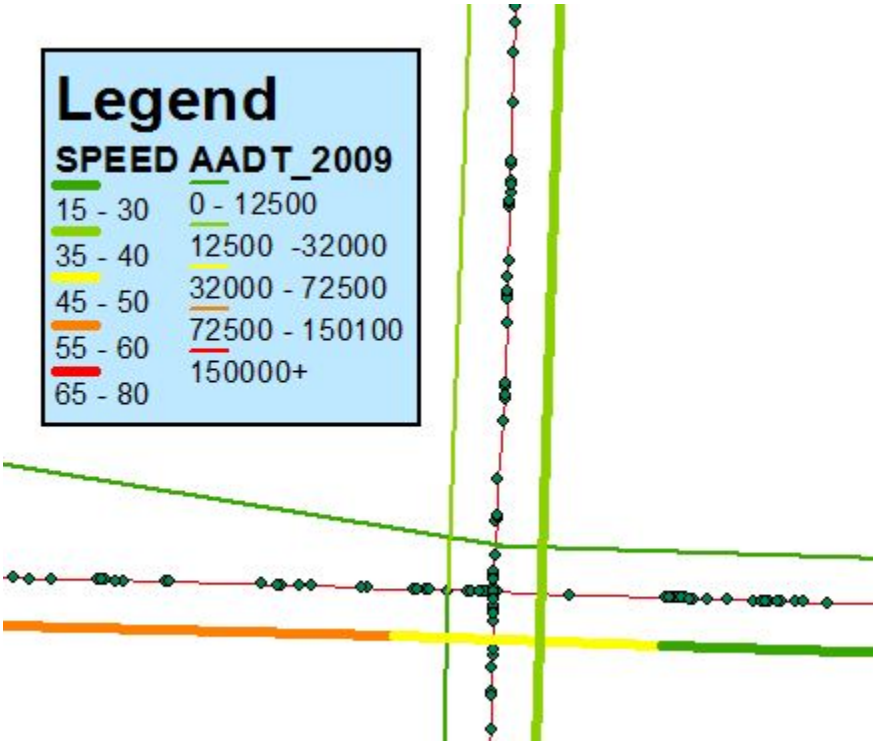


Figure 3-9. Offsets are used to show changes in roadway characteristics.

The data mapped in Figure 3-9 shows how these visualization methods help an analyst identify patterns, trends, and potential causes in crash data. For example, at the intersection in Figure 3-9 there are only crashes on the eastbound approach leg. From the speed limit line (thicker line below the road) the reader can see that there is a high approach speed with a speed limit reduction before the intersection, which could be a contributing factor to the crashes on the eastbound leg. A cluster of crashes is also apparent when the speed drops again from the 45-50 mph range to the 15-30 mph range. This suggests that a more gradual speed reduction or improved signage could reduce the number of crashes. In this example the AADT values (thinner line in the image) do not appear to be correlated with the crash because the AADT is fairly consistent throughout the segments. For the north and south legs of the intersection there does not appear to be any obvious patterns or correlations in the crash and roadway data. This doesn't guarantee that this is not a high risk location for those approaches; it only suggests that the north-south leg is not likely to be a candidate for safety improvements. After noticing these trends in the data a statistical analysis should be done to see if the observations are valid and if improvements can be made in this area.

3.3.2 Advanced Visualization

Several advanced GIS visualization techniques are useful in crash data analysis. Stacked bar charts can be used to represent areas where multiple crashes occur at the same point. Figure 3-10 shows how this was done in a program developed by Iowa State University for the Iowa Department of Transportation (IDOT 2012). This program will be discussed in depth in Section 3.5.5 of this report. The stacked crashes allow easier identification of high crash locations versus the scaled or colored points display methods shown previously. Several others have used a similar method (Roche 2000).

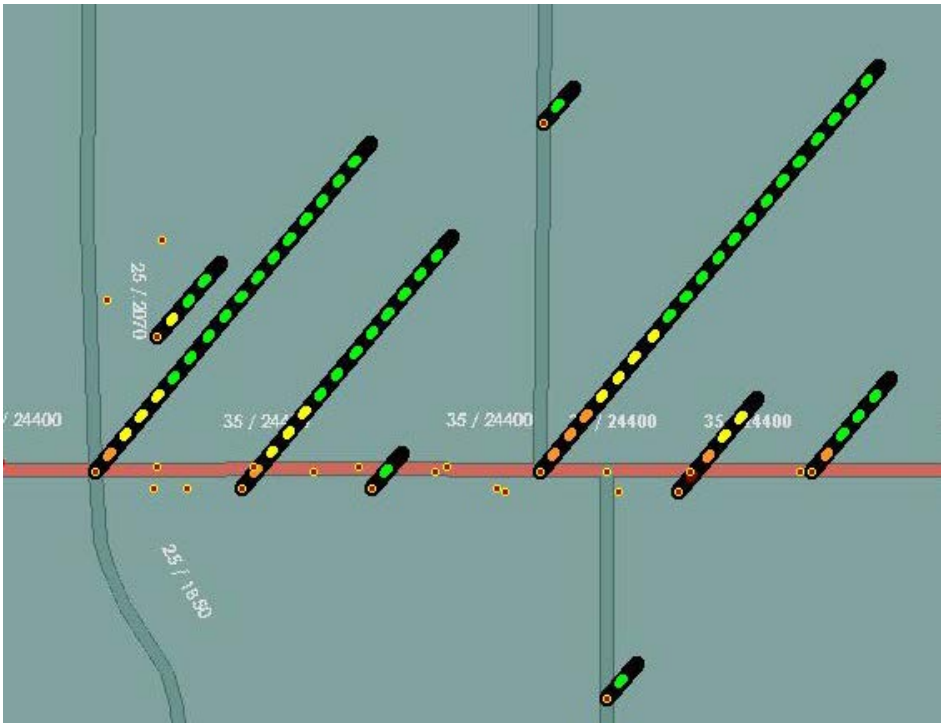


Figure 3-10. Stacked crashes help identify high crash locations where points overlap (IDOT 2012).

Another way to show overlapping incidents has been developed by Google Earth (Google Earth 2011). When the cursor hovers over a certain point and is clicked, if there are overlapping features they will expand or "pop out" of the page to allow viewing of every feature as shown in Figure 3-11. This could be beneficial when looking for individual crashes or when examining a site that has already been identified as a high risk location. Crashes will often overlap so having them expand would help in viewing and analyzing individual incidents.

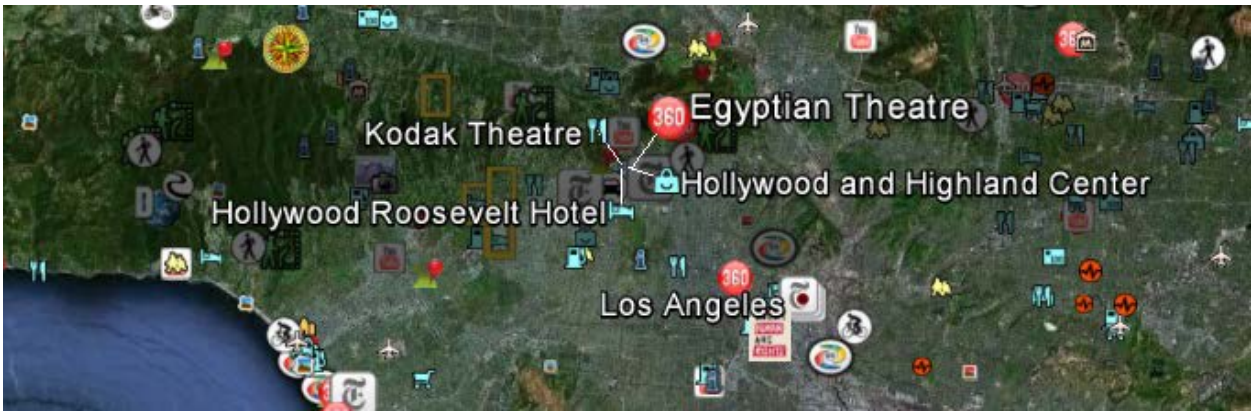


Figure 3-11. Overlapping features "pop out" to show each location (Google Earth 2011).

Three-dimensional (3-D) GIS tools can provide a stacked perspective to crash data and are more informative than a colored two-dimensional (2-D) line. Figure 3-12 provides an example of how 3-D tools can be used to provide this perspective. The figure shows how segments can be represented as a column. With 3-D columns the actual value is represented by the height of the column, so the graphic provides more information than just a colored line. Colors can still be used to help group segments with similar characteristics or risk factors as shown in Figure 3-12 (Li and Zhang 2007).

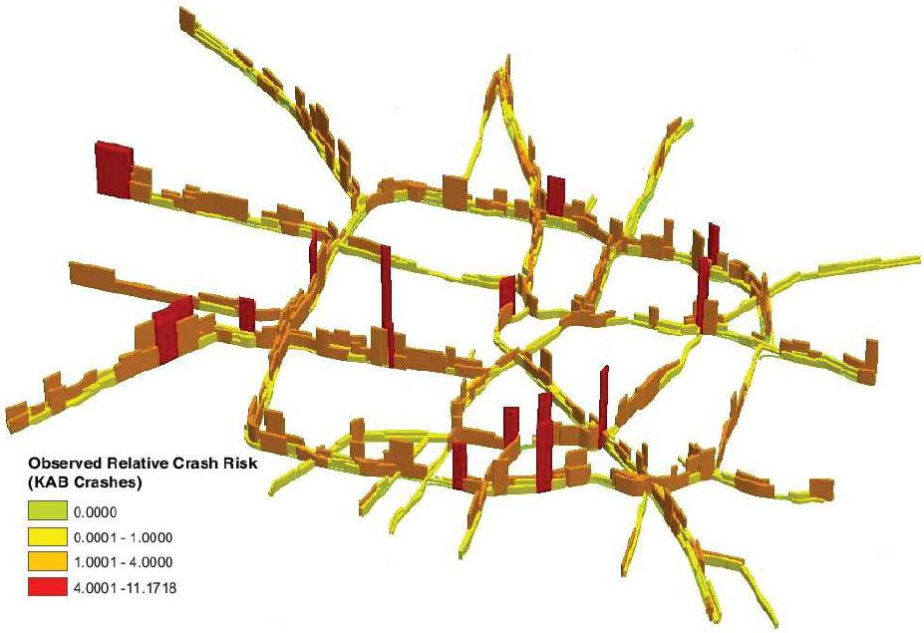


Figure 3-12. Segments shown as 3-D columns to show crash risk and grouping (Li and Zhang 2007).

Gradient or density representations are useful in showing crash risk. Because crash data are point locations and roadways are linear, continuous raster or polygon gradient representations are not completely applicable, but still provide interesting presentation of results. “Heat” maps, as they are referred to by saferoadmaps.org (SafeRoadMaps 2012) and shown in Figure 3-13, show crash risk across an entire area. The problem is that this infers risk across the entire map area, which doesn’t make sense given that crashes can’t occur everywhere and not all roadways are connected or related. For example, the hotspot at the center of Figure 3-13, contains five different crashes, none of which occur on the same road, but according to the image they are part of a hotspot or high crash location. This makes it difficult to use the results to identify any

potential causal factors or countermeasures because they are five crashes on different types of roads (Breyer 2000).

A more useful perspective of density or gradient representation is shown in Figure 3-14. The same idea of a crash density was used, but the gradient was only shown across the linear segments of roadway. This method assures that influence will not spread to non-adjacent routes and matches actual conditions of the individual roadways more accurately. The benefit of gradient maps is that the segments are not so discrete and defined. It is easier to see how influence or risk spreads across an entire roadway rather than the risk breaking at specific points within the segment as seen in previous maps (Esri 2012b).

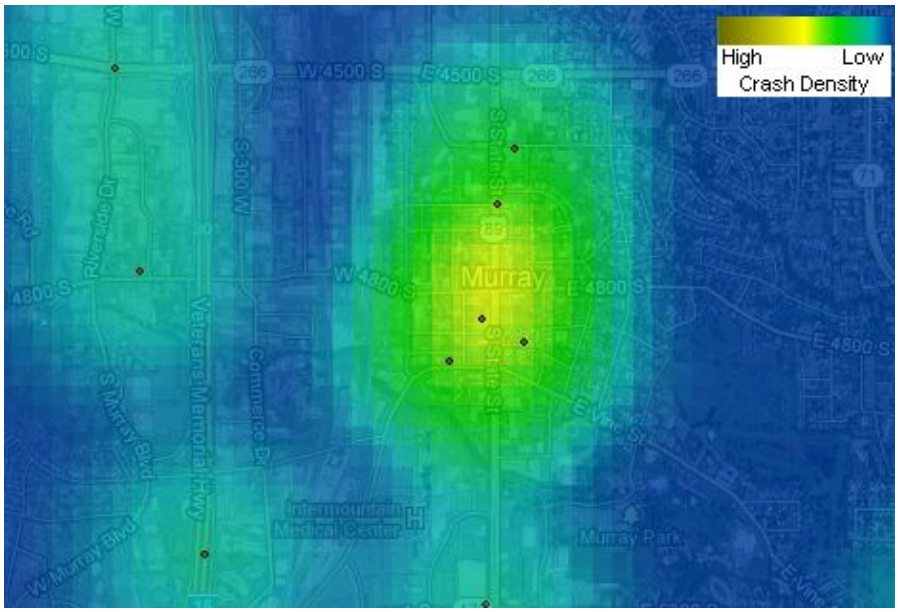


Figure 3-13. "Heat" density map identifying high crash areas using saferoadmaps.org.

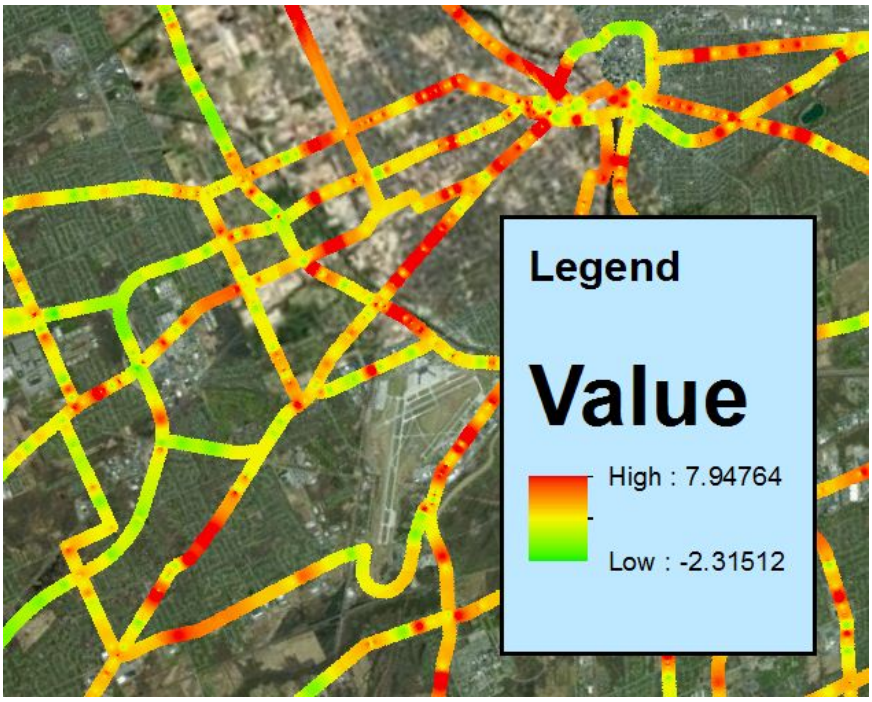


Figure 3-14. Linear crash density map (Esri 2012b).

3.3.3 Imagery

Aerial imagery quality and availability has improved significantly over the past several years. This has made GIS technology increasingly viable and powerful in crash data analysis. Having imagery for locations during crash analysis provides an even richer context for the analyst to see and understand the factors involved in the incident. The high-level perspective of aerial imagery provides information that would be very difficult or impossible to discern from data alone. With this, the analyst needs to be careful to not let the image weigh too much in the analysis; any conclusions should be backed by statistical evidence. The imagery should serve as information to help interpret and understand the data and statistics.

Figure 3-15 and Figure 3-16 provide an interesting example of how imagery can be beneficial in data analysis. Figure 3-15 shows a location that has a high crash rate according to the analysis completed. With no imagery there is very little information that can be gathered from the map alone. It does seem unusual that the high rate section is away from the intersection, where crashes are more common. When the imagery is added as shown in Figure 3-16 the cause for the high crash rate seems clearer. A very large queue exists in the left turn lane, and based on the angle of vehicles at the back of the queue the analyst can see that drivers have had to make

sudden maneuvers to enter the left turn lane, not realizing that the queue was so long. Looking at the collision types, the analyst can see that nearly all the crashes on that high-rate segment are two vehicle angle collisions, which would fit the hypothesis regarding left turning vehicles. From there the data could be analyzed further in a statistical program or a more detailed study could be performed. Even though all crash site imagery will not have such useful information as this example, this example shows how imagery can substitute for or enhance a site visit. The issue with left-turn vehicles would be easy to observe if the site was visited. Other situations where imagery can be beneficial include viewing land uses, access densities and locations, signal spacing, and other roadway or roadside characteristics.

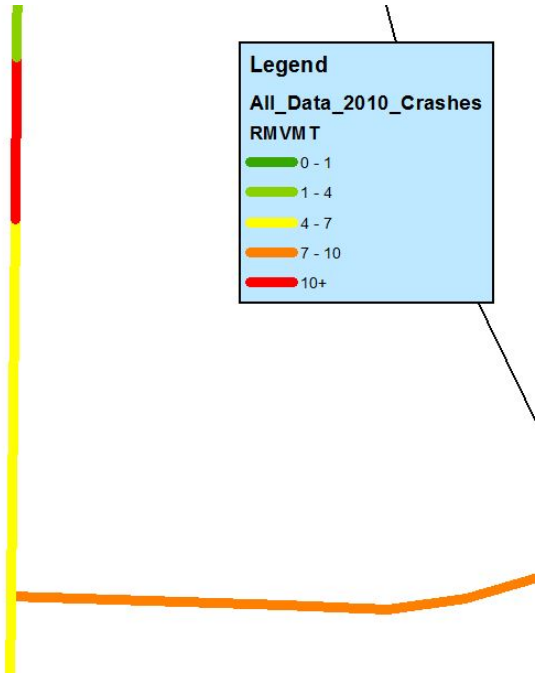


Figure 3-15. Crash rates without imagery.

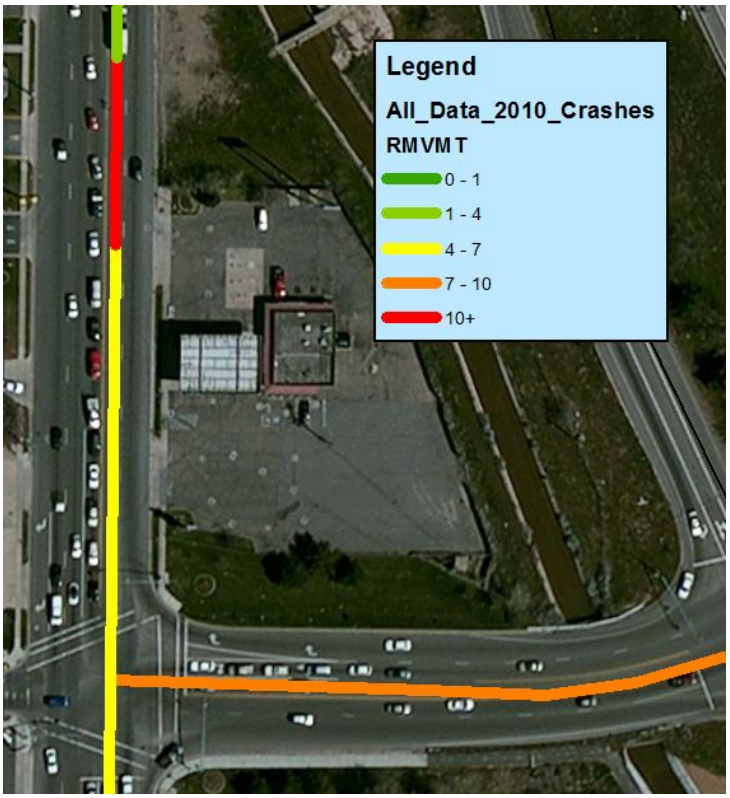


Figure 3-16. Crash rates with imagery.

Street view images can provide additional information when analyzing a specific site. Issues like poor signage, road striping, or site distance can be seen using a street view image but not with aerial imagery. These images are not available directly within ArcMap but can be found on Google Maps and through the UDOT Roadviewer (UDOT 2011a). Several programs developed for crash data analysis also incorporate this feature as will be discussed later in this report (see Section 3.5.3 and 3.5.9). Figure 3-17 shows a street view pop-up window used in saferoadmaps.org. When a crash is selected this window immediately shows the site. CARE, developed by the University of Alabama uses a similar street view interface for analyzing crashes (State of Alabama-CARE 2009).

Figure 3-17. Street view pop-up window used in saferoadmaps.org (SafeRoadMaps 2012).

3.4 Analysis of Crash and Safety Data

GIS analysis of crash and safety data involves processing data to enable interpretation and performing statistical analysis. Most analyses performed in GIS are very similar to how they would be done outside of a GIS environment, but the added spatial context adds functionality, gives results that are easy to connect to actual locations, and can make the analysis much easier. This section will explain query and attribute selection tools, analysis tools developed by Esri for use in ArcMap, and statistical analysis methods.

3.4.1 Query and Attribute Selection Tools

The ability to analyze data is dependent on the ability to query and select records of interest, something GIS is well-suited for. ArcMap has several built in tools that allow this. Data can be selected by both location and attributes. Location selection is used for queries such as 'Select crashes that occurred within Davis County,' or 'Select crashes within 2 miles of the Interstate.' Attribute queries are focused on the type of route or crash. Examples include, 'Select crashes that involved 3 or more vehicles,' or 'Select crashes with high severity that involve alcohol,' or 'Select roadway segments with crash rates higher than 10 crashes per MVMT.' More

complex queries that include both locational and attribute features are common in data analysis and greatly simplified by GIS. Examples include, 'Select crashes of high severity that occurred within 100 yards of a road segment with a radius of 200 feet or less,' or 'Select crashes that occurred near intersections while traveling west between 5 and 8 pm.' These queries would be difficult to complete using only a tabular database, because the database has limited locational information, and the query requires information that may be stored in several different datasets. GIS provides a framework for analysis that allows attribute and locational querying use tabular and spatial data of several different datasets simultaneously (Esri 2012a).

GIS allows the user to input datasets and combine data to allow queries like the examples given. Roadway data such as AADT, speed limit, or curvature can be added to a crash record, using join operations. This allows easy selection of crashes that fit a certain criteria. Without the ability to join these data together the analyst would be dependent on crash comments/ descriptions or checking each individual location by hand. Buffer tools make proximity queries easier, and overlay or intersect tools allow datasets that are not compatible to be transformed into datasets that can "communicate" with each other. These same operations could possibly be done without GIS, but using GIS is far more efficient and produces results that have an accurate spatial connection to the real world.

Query methods are especially useful when mining crash data or preparing it for statistical analysis. Data mining is aided with GIS because the results are seen visually and patterns can be more apparent. Many statistical models also require very exact inputs and these query and selection tools facilitate creating those data sets.

Querying data also allows analysts to apply the treatability concept. The treatability concept is the idea that analysts should only look for crashes and areas where the problem can be solved with engineering measures. Because human error is so prevalent in crash occurrence, not all high-risk locations can be fixed with engineering solutions. Recognizing this, analysts should only address crash risk areas where the issue can actually be resolved. This concept is still somewhat disputed but is a suggestion that bears consideration (Miaou and Song 2005). Should this concept be applied GIS is the tool to make it happen. GIS could be used to find all candidate sites for the specified treatment, and then statistical analysis can determine which location needs the treatment most urgently.

3.4.2 ArcMap Analysis Tools

Esri has developed three standard tools for crash analysis in ArcMap. These include the spot, strip, and sliding scale analysis. The following sections will explain each tool and its use in crash data analysis. (Esri 2012c, USDOT 1999)

3.4.2.1 Spot Analysis

Spot analysis helps analyze crashes within a specified distance of selected points. The points identify locations of interest for analysis (typically intersections) and the distance represents a buffer around those points. The tool will output a dataset containing all crashes occurring within the specified buffers.

Figure 3-18 shows a sample output of the spot analysis tool. For this example, points were placed at each of the three interchanges and a one quarter mile buffer chosen. This tool is a preliminary analysis tool that helps subset the data before analysis. The ArcMap spot analysis tool allows the user to quickly place points at an intersection and then run the tool. An additional benefit is that once an intersection point file is created placing the points is not necessary a second time. Instead a query could select the intersections of interest then run the analysis on those locations.

Figure 3-18. Sample output of Spot Analysis Tool.

3.4.2.2 Strip Analysis

Strip analysis helps analyze roadway segments. The tool breaks the input roadway system into segments that contain a user-specified number of crashes. A window length is selected along

with a minimum crash threshold. The tool works by laying the window over the roadway end to end and counting the number of crashes for each window. Any window that has at least the specified number of crashes will be in the output file.

Figure 3-19 shows a sample output of a strip analysis. The roadway has been broken into quarter mile segments with crashes counted for each segment. Post-processing was then used to calculate the rate per MVMT and display it by color. In the example, using the strip analysis has helped the analyst identify the sections with the highest crash rate, without being limited by pre-defined segments. It is likely that in a typical analysis the roadway shown would be only one or two segments instead of the six or more seen, greatly diluting the crash rates. Another benefit is found at the intersections. With pre-defined segments each intersection will typically be split in the center, creating four different segments that include the intersection crashes and diluting the effects of the intersection. Strip Analysis has a greater chance of including both intersection approaches in the same segment as can be seen in Figure 3-19. This tool also serves as a method to analyze only segments that have a reasonable number of crashes. Pre-defined segments often lead to segments with very high crash rates but a low number of occurrences, making analysis difficult and countermeasures not financially viable. Another benefit of the strip analysis, as shown in Figure 3-19 by the thin red lines, is that some segments were not analyzed because they did not have the minimum number of crashes. This tool allows the analyst to ignore these segments and focus on areas where countermeasures can have a significant impact.

Figure 3-19. Sample output of Strip Analysis Tool.

3.4.2.3 Sliding Scale Analysis

Sliding scale is the final tool. Sliding scale is very similar to the strip analysis except that it moves the window along the routes in an increment rather than end-to-end. This can be even more beneficial than strip because there is more flexibility in where the segment begins and ends. The only concern with the version of the tool available from Esri is that it aggregates all adjacent candidate segments into one long segment, meaning this tool is only useful on a corridor level. Figure 3-20 shows how this tool can be useful for identifying high crash corridors, which could then be studied further with different analysis methods.

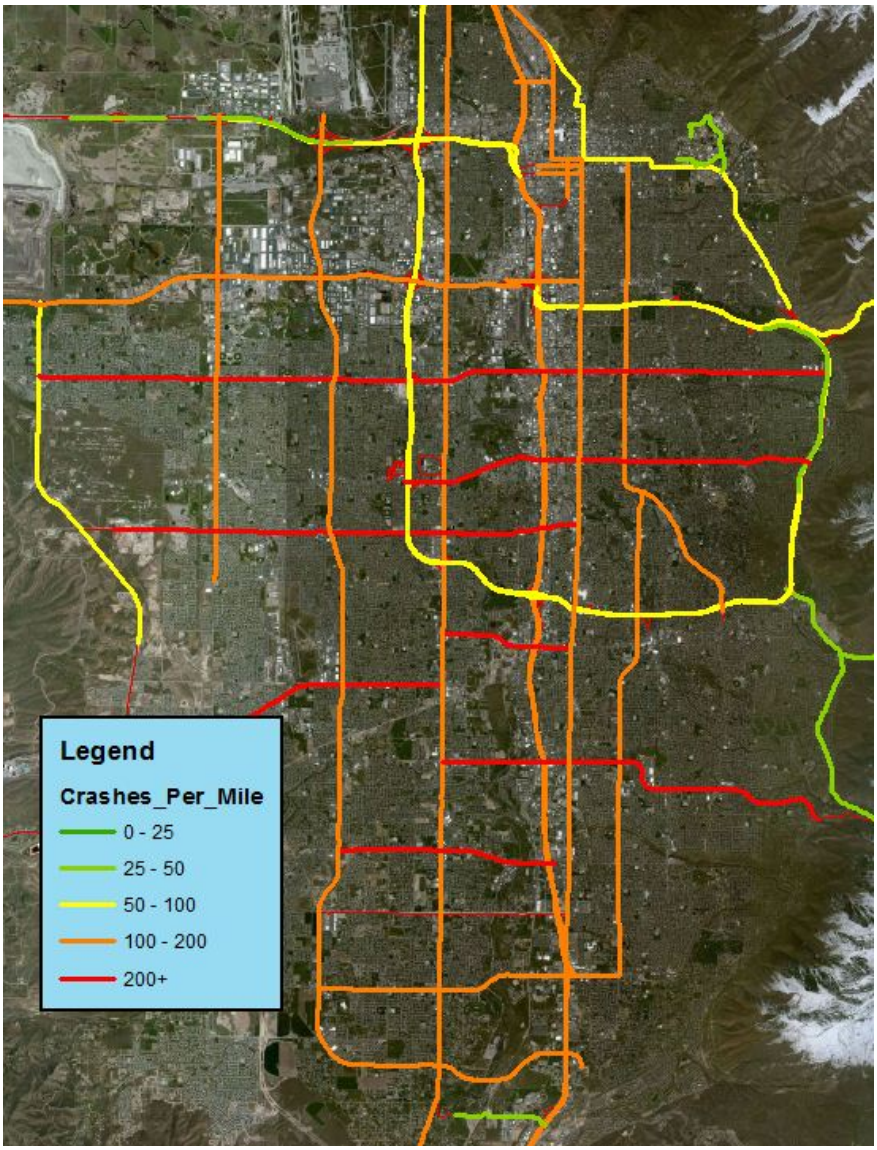


Figure 3-20. Sample output from Sliding Scale Tool.

A variation sliding scale was created for a study conducted by South Dakota State University. It performs the same task but doesn't aggregate the segments to one linear feature. Instead only segments of the specified window length will be output for analysis. The concern with this method is that the segments will overlap. This is better for accurately identifying short segments that need safety improvements, but still is limited for use at a statewide scale because of how much data it creates and processes. In the South Dakota study analysts had to break the state into small regions to reduce the processing time (Qin and Wellner 2011).

3.4.3 Statistical Analysis

GIS has improved the ability of an analyst to perform advanced statistical analysis of crash data (Miller 2000). In general, GIS aids statistical analysis in two ways. First, it allows more careful and accurate data selection, screening, reduction, and spatial analysis of the results in pre-and post-processing of results. Second, GIS has allowed the development of spatial statistics that rely on geographically referenced data and software platforms that can handle processing of that data. This section will only provide a brief summary of these analysis methods and refer the reader to additional sources.

Bayesian statistics have become the "gold standard" of crash data analysis. A primary reason for this is the ability of Bayesian models to account for the "small area problem" that is prevalent in crash data (Davis et al. 2009, Li and Zhang 2007, MacNab 2004, Miaou and Song 2005, Mitra 2009). The small area problem is the uncertainty seen in crash data when using low volumes to calculate exposure, rates, or very short segment lengths. Low volumes can make crash rates appear very high, and short segments can do the same, especially when there are a lot of crashes within the short segment, such as within an intersection. GIS improves the ability of Bayesian models account for the small area problem and incorporate spatial correlation into the model by providing the spatial connections in the data. Li and Zhang (2007) used GIS to create relative crash risks within roadway segment classes and then generated an "adjacency matrix" to allow the model to borrow strength from roadways of similar characteristics and adjacent locations. Such a process would not be possible without GIS and the ability to incorporate spatial data. Li and Zhang's (2007) study showed how GIS and advanced statistical models can be used. The Bayesian analysis was performed outside of any GIS platforms. However, all of the data preparation, segmentation, and screening were performed in GIS. This approach is effective

because GIS allows the user to work with the data in both a spatial and tabular context. It also allows the user to export and then import data from many different formats. It is possible to complete the analysis without ever viewing data or results outside of GIS. Data can be exported as a CSV or many other formats, imported to a statistical model, then the output can be placed right back in GIS and evaluated. This basic workflow is simple, adaptable to many situations, and produces accurate results. Figure 3-12 showed the output produced by the Li and Zhang (2007) model. Sections 2.3.4, 2.3.5, and 5.1 of this report include more details about Bayesian modeling.

The Getis-Ord Statistic is a spatial statistic that has found application in many fields, including crash data (Getis and Ord 1992, Khan et al. 2008, Ord and Getis 1995). In many situations this statistic has also been used with the Moran's I, another spatial statistic (Anselin 1995, Getis and Ord 1992, Khan et al. 2009). Both of these are aimed at using distances and spatial correlation to identify high-risk or high occurrence locations. The statistics were created before GIS was commonly used but have since been integrated because of ability to analyze the distances, locations, and spatial information needed for data inputs (Anselin 1995, Khan et al. 2008). The Getis-Ord Statistic and Moran's I and both now available in ArcMap as standard tools for data analysis.

Other uses of statistics in GIS have been examined in the literature. Harkey (1999) provides a good example of how GIS can improve data preparation for statistical analysis. GIS tools were used to identify truck crashes for the analysis. GIS was then used to select routes that trucks are legally allowed to drive on using buffers and network tools. The results were analyzed with weighted rate statistics to determine high risk locations (Harkey 1999). Khan et al. (2009) used GIS to analyze crashes across the entire state of Wisconsin using Network K-functions and Lattice Data Analysis. Finally, Pulugarth et al. (2007) used a Kernel Density method to analyze high pedestrian accident locations.

3.5 Existing GIS-Based Crash Data Analysis Programs

Several programs have been developed around the country for crash data analysis. Each program has developed new ideas and applications for how GIS can be used effectively in crash data analysis. This section will examine several of these programs and discuss the best attributes

of each along with elements that could be improved to satisfy the specific needs of UDOT. Analysis of each program is from the perspective of the research team writing this report with input from UDOT staff. Each of these programs was developed for very different reasons so it is understood that their goals are likely different than that of the research team and UDOT. Because of this each program will be analyzed to identify its performance relative to UDOT's preferences, even though those criteria may not have been important for the subject programs use and development. Any negative discussion about programs is only meant to help UDOT identify features of their system, it is not meant to point out shortcomings of other systems.

The specific programs that will be addressed in this chapter include Plan4Safety, UMassSafe, Arizona Department of Transportation (DOT) system, The Bay Citizen Bike Accident Tracker 2.0, CMAT, saferoadmaps.org, Maryland Spatial Analysis of Crashes, Esri Executive Dashboard, CARE, usRAP, and MassTRAC.

3.5.1 Rutgers University: Plan4Safety

Rutgers University Center for Advanced Infrastructure and Transportation (CAIT) has developed Plan4Safety, a crash data analysis tool for the New Jersey Department of Transportation. The system includes statewide crash data, roadway characteristic data, basic statistical analyses, network screening layers and models, and GIS analytical tools. The program was not available to be personally tested as part of this research, but some information was available through the user guide and online information (Rutgers CAIT 2011).

The analysis tools in Plan4Safety are based on sets of filters that query and select data. This data can then be input to some basic statistical tools. The first tool is the Cluster Finder. The Cluster Finder seems to work similar to the strip analysis tool produced by Esri (Section 3.4.2). The tool finds all segments with a crash cluster and counts the number of crashes over that segment. The second tool is Cross Tab. Cross Tab creates tables of data cross-tabulated to help identify patterns and correlation of variables within datasets. The final tool is Frequency Analysis. This tool calculates crash frequency on routes along with the frequency of different types of crashes and their attributes. All of these tools are basic descriptive statistical methods that can help identify high crash locations. There do not appear to be any tools that will normalize data according to AADT, segment length, or population, or tools that will account for roadway and crash variables in the high crash location process.

The GIS user interface for Plan4Safety is similar to Esri web-apps and templates available on Esri's website (Esri 2011b). The interface has a number of filtering operations that can help view the data and prepare it for analysis.

With the limited information about the program it is difficult to make any certain conclusions. The analysis workflow does seem a little rigid. The user has to create filters then load them into the analysis section of the program. It is preferable that the filter could be created and adjusted dynamically during the analysis, rather than having to rebuild the query and load it back into the program.

3.5.2 UMass-Amherst: UMassSafe

UMassSafe at the University of Massachusetts-Amherst developed the Commercial Motor Vehicle (CMV) Crash Data Tool for crash data analysis in Massachusetts. The CMV Tool is designed primarily for law enforcement officers to identify enforcement and officer management or training needs, rather than roadway improvements. The tool was developed in two pieces, the Data Explorer Application and the Crash Mapping Application. Neither application was available for testing but information was available online. The Data Explorer presents a tabular view of the data with quality measures, summary statistics, and the ability to explore the data. The Crash Mapping Application shows the crashes across the state visually with the same data as the Data Explorer (UMassSafe 2006).

The GIS interface for the program is effective and seems easy to use. Figure 3-21 shows how filters and data selectors are available on the left, while the yellow panel shows the symbology, legend, and counts for each type of crash. The map to the right then displays the data. This interface is good in that it allows the analyst to adjust queries, symbol definitions, and map location very quickly throughout the analysis. The mapping tool does not appear to have as much statistical functionality as the Data Explorer tool. This may have been done for practical reasons, but the ideal method is to include all tools in one interface. Figure 3-22 shows another feature of the tool, useful for examining individual crashes. The pop-up window includes tabs for the crash, vehicle, and occupant data, along with a measure of the data quality (UMassSafe 2006).

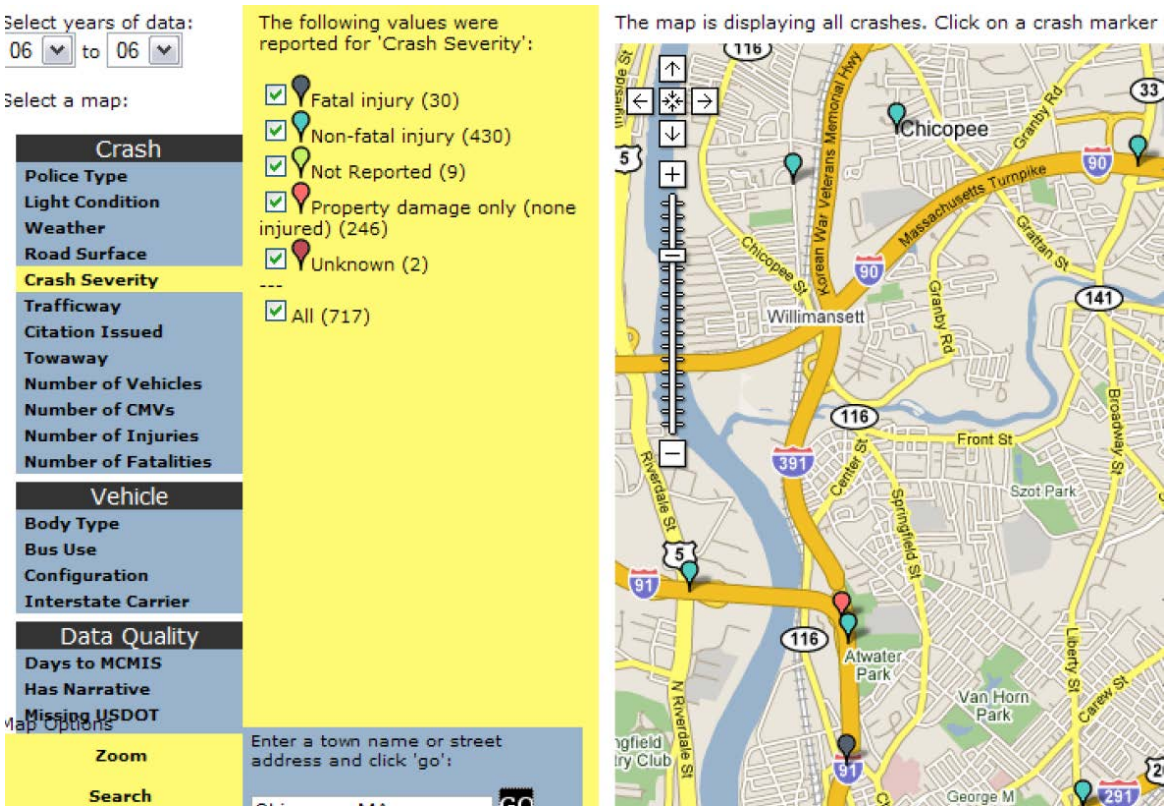


Figure 3-21. CMV Mapping tool user interface (UMassSafe 2006).

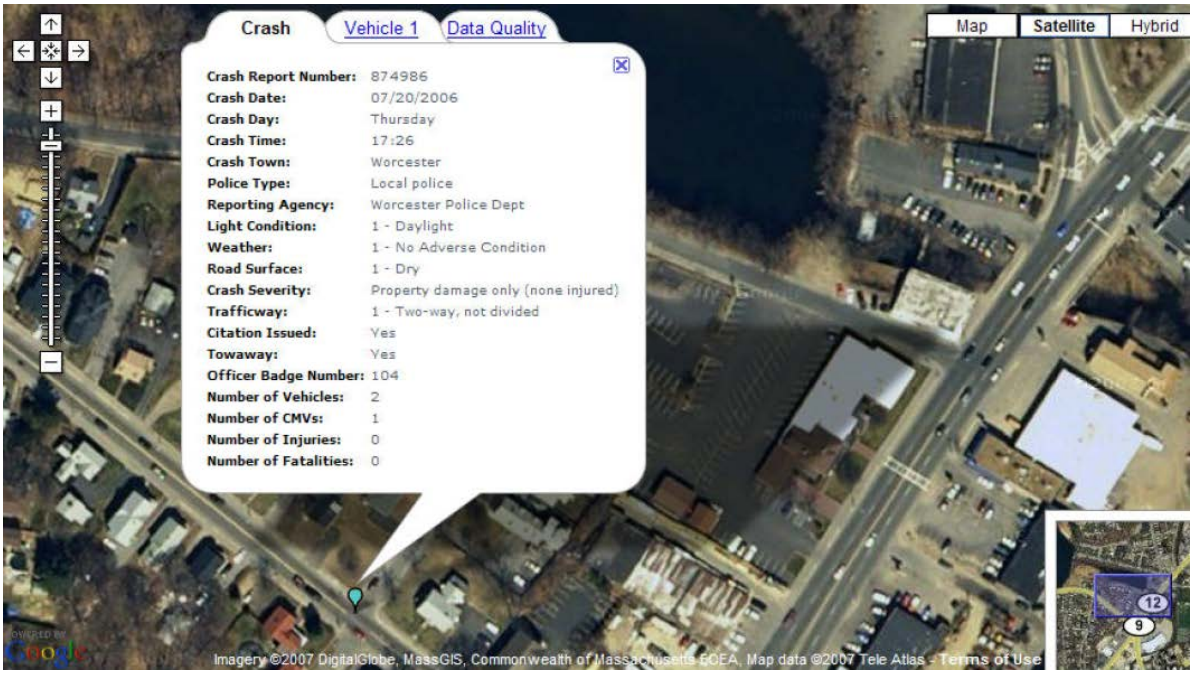


Figure 3-22. CMV tool crash attribute pop-up window (UMassSafe 2006).

Because the tool was created primarily for law enforcement it does not include roadway attributes, but instead has extensive information about the officer and police troop/barracks investigating or reporting the crash. In an engineer-oriented system it is preferred that more roadway data is included. For the mapping system at least 25 percent of crashes had to be hand located because of bad or missing data.

3.5.3 Arizona DOT GIS System

Breyer explains an early attempt of the Arizona DOT (ADOT) to integrate its crash database with GIS software and roadview imagery. The result was a system that could perform macro and micro analysis and generate reports with street-level images for the crash sites. The tool used a basic interface including a legend, map display, data tables, and access to photo log imagery (Breyer 2000).

The macro tools in the ADOT system include the Spatial/Grid view and the Translated Network view. The Grid View uses grid cells projected on the roadway to identify locations with high crashes. There are two limitations to this method. First, using grids for linear data divides the roadway and aggregates data in unusual patterns. Second, the grid cells will often include crashes that occur on multiple non-intersecting routes, skewing the data and making crash cause identification very difficult. The research identified that intersection crashes could be collected this way, but not on routes that do not intersect such as overpasses or frontage roads (Breyer 2000). The Translated view tool was designed to remedy this problem.

The Translated view tool will collect all crashes that intersect a study route within a specified distance and analyze them as if they were on the same route. This method acknowledges that causes at intersections may be associated with both routes and thus they should be analyzed together, separate of the segment crashes (Breyer 2000). This is an operation that now could be improved even further with the spot analysis tool created by Esri as outlined previously in Section 3.4.2.

Micro tools used in the system include attribute filters that analyze only user-specified crashes and situations. They also include the mapping and photo log viewing tools. The tool also used highway centerline files to analyze curve and grade of the roadway (Breyer 2000).

Considering the fact that this program is somewhat dated because of its age (written in 2000) the system developed incorporates many of the best practices mentioned in this literature

review. It includes display tools to show crashes by colors, symbols, and scaled lines. Statistical analyses available include simple rates and more advanced grid-density plots. The system also possesses the ability to query different routes and types of crashes for the analysis. The greatest feature of this system is the ability to link video log images to route locations. The system uses LRS to allow the analyst to pull up any street-level image and see the site as shown in Figure 3-23. Having these directly linked is a huge advantage and time saver. It also can help in creating attractive and informative presentation materials. The system actually takes this a step further by digitizing some of the features shown in the video log. Figure 3-24 shows how features such as signs and slopes seen in the video log have been represented on the roadway along with crashes. This further integrates the data and can allow more in depth querying and crash selection based on roadway features from the video log.

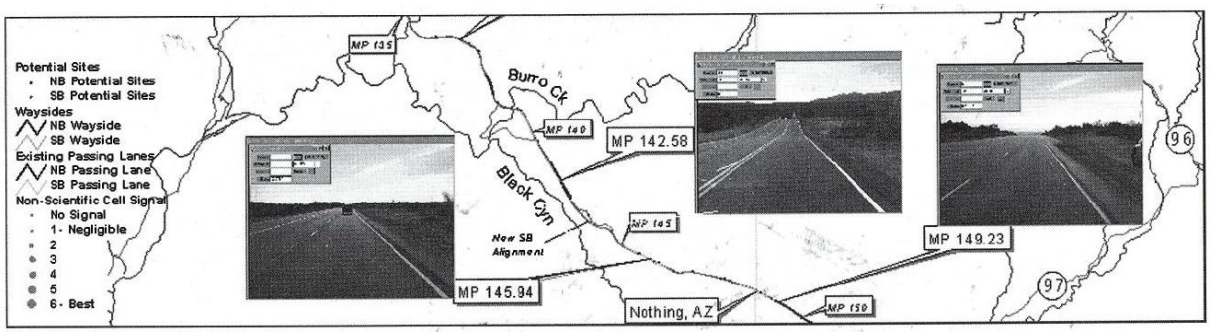


Figure 3-23. Video log images improve analysis of crash locations (Breyer 2000).

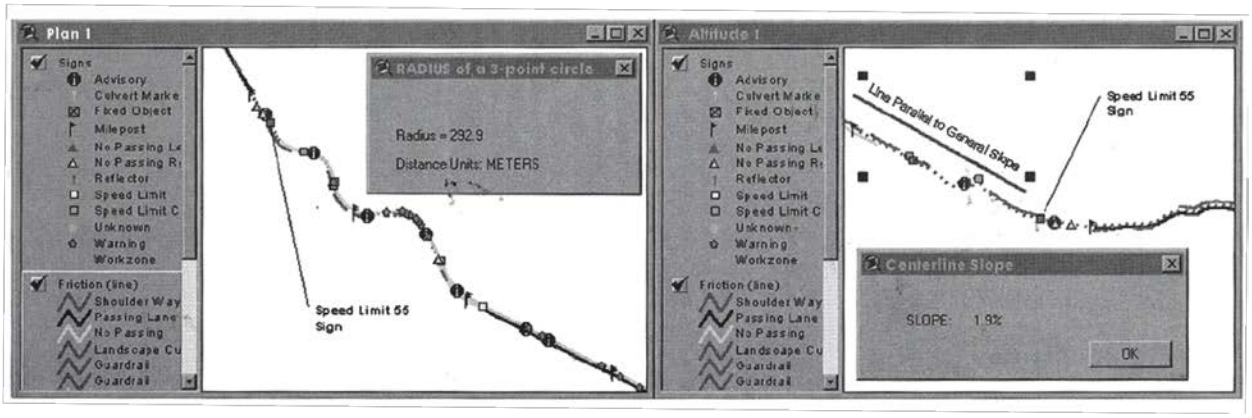


Figure 3-24. Video log data is digitized to points on the map display (Breyer 2000).

3.5.4 The Bay Citizen Bike Accident Tracker 2.0

The Bay Citizen (2011), a nonprofit citizen-run newspaper created the Bike Accident Tracker to increase awareness of bicycle safety issues and informs cyclists of safe and dangerous areas. The application is not designed to be a full-scale safety analysis program, but does incorporate some features worth addressing in this section.

The first feature of the Bike Accident Tracker worth noting is a fairly clean and accessible interface as shown in Figure 3-25. The interface has all of the basic tools and information right in front, without complicated sub-menus to search through. To the left is the filtering menu that offers attribute-based querying, along with some additional filters on the top "Overlay" bar. Tabs are added at the top to view summary charts or raw data (The Bay Citizen 2011).

Another useful feature is that clicking on any crash pulls up a window with additional helpful information as can be seen in Figure 3-26. The window shows most essential information a user would be interested in including the date, time, violation, and case number. Graphics are used to show the manner of collision (car hit bike) along with who was at fault, the injuries, fatalities, and if it was a hit and run. This is an attractive and simple way to access the information and is much easier to understand than reading a crash report or looking through data tables. Combining this idea with the multi-tabbed pop-up window used in the CMV mapping tool (shown previously in Figure 3-22) would create a good method of presenting different levels of information. Pop-up windows like this could even be customized to show data that the user is specifically interested in.

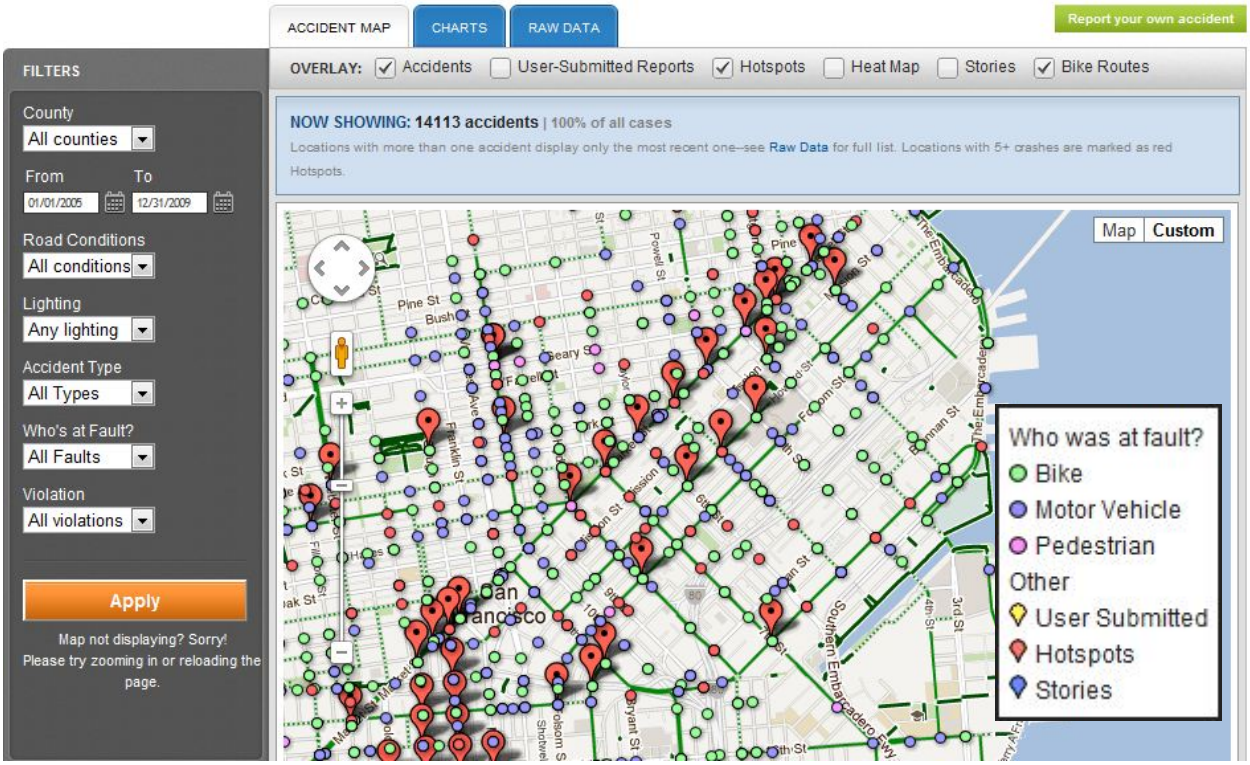


Figure 3-25. BayCitizen Bike Accident Tracker 2.0 interface (The Bay Citizen 2011)

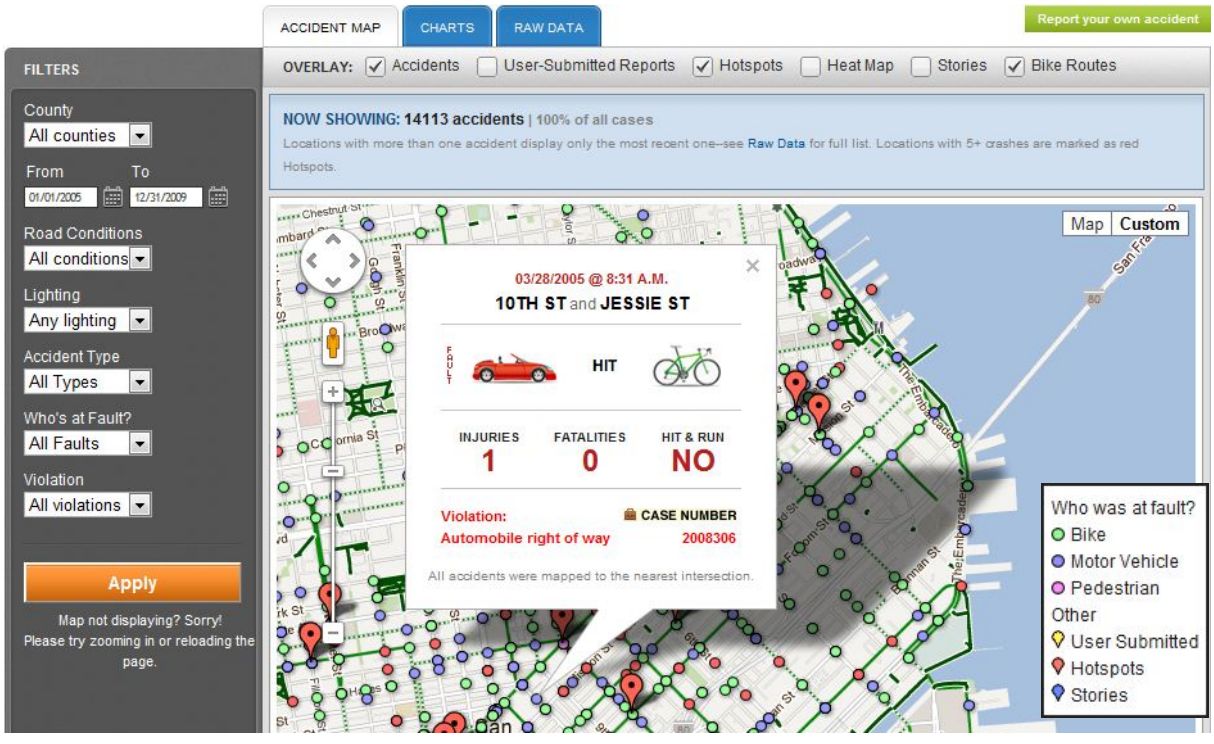


Figure 3-26. Crash detail window provides additional crash information (The Bay Citizen 2011).

3.5.5 CMAT

The Crash Mapping Analysis Tool (CMAT) was developed by the Center for Transportation Research and Education (CTRE) at Iowa State University for the Iowa DOT office of Traffic & Safety (IDOT 2012). The purpose of the program is to facilitate developing safer roads and safety programs by providing public agencies and contracted consultants in Iowa with access to information about crashes. The program now contains 10 years of crash data (2001-2010) on all state roadways in the state and has several features for the user to view and analyze those crashes. While CMAT is one of the most comprehensive programs currently in use that was available to be investigated for this report, it has also been in use longer than most programs and by different states. CMAT Version 3.7 was made available for use in this research. As of Fall 2011 an updated version was being worked on but no completion date was available (Robert Schultz, personal communication, May 2011).

Figure 3-27 shows the CMAT user interface. The program uses a basic window with map display in the center. All of the smaller windows seen in Figure 3-27 are options added by the analyst. Because the primary goal of CMAT is to get data into the hands of various decision makers, the program is built around finding crashes, not analyzing them.

CMAT uses “Filters” and “Finders” to locate and identify crashes. Filters are used to identify crash variables the analyst wants to consider. Finders are used to identify the area or location where crashes will be selected. Finders include city, county, intersection, case number, milepost, node, map coordinate and many others. Figure 3-28 shows the intersection finder dialog box. As parameters are modified in each drop-down box the options for other boxes are limited accordingly, making it easier to locate the desired options (IDOT 2012).

Other options available in CMAT for locating crashes are manual selection methods such as point, polygon, or route segment. These allow the analyst to manually select the area for analysis. Route segment selection methods are limited because the segment lengths and locations are static to the user. Having manual and filter-based options for selecting locations makes the program much more user friendly and flexible (IDOT 2012).

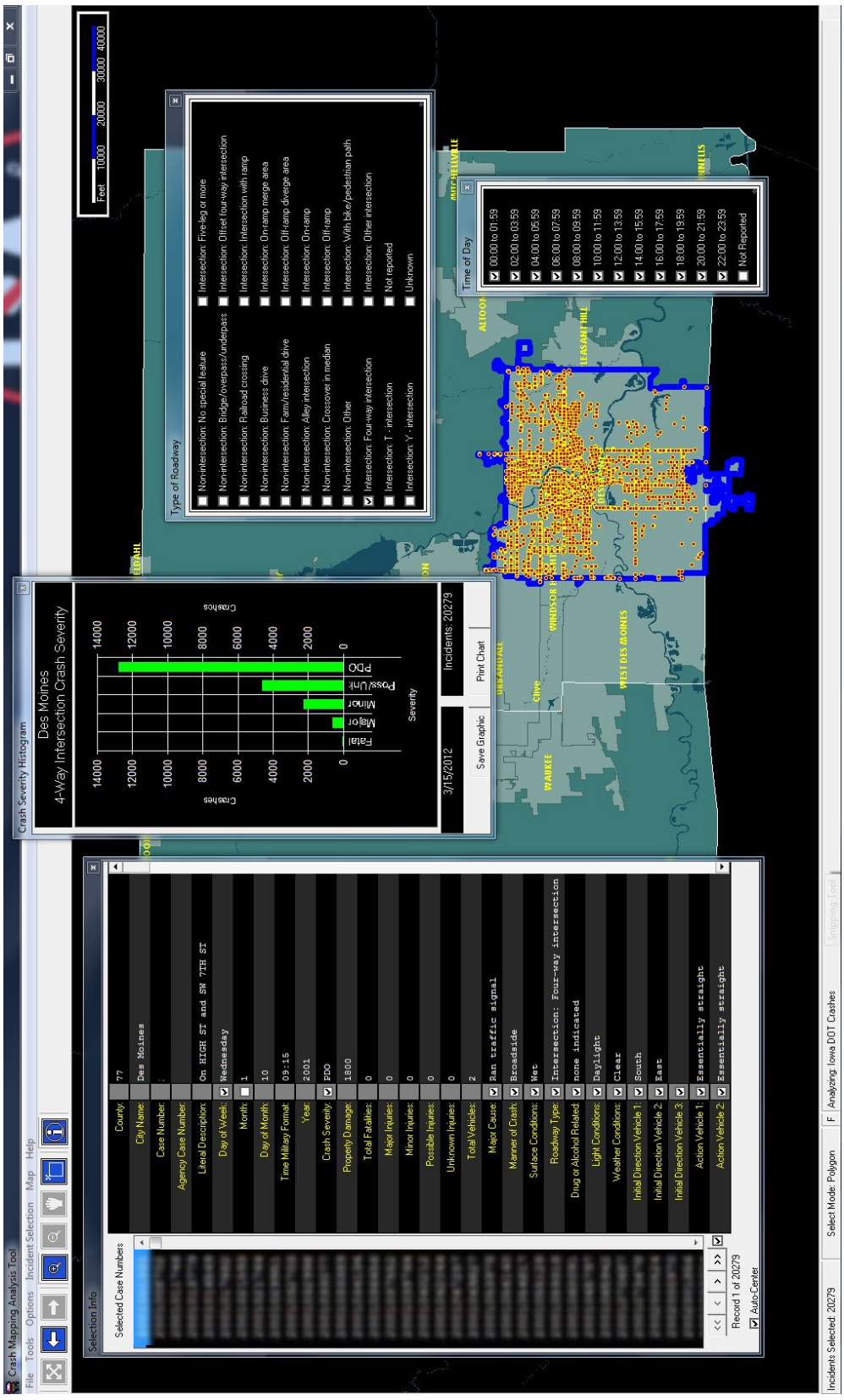


Figure 3-27. CMAT interface with filters, charts, and crash details (IDOT 2012) (Crash ID numbers have been intentionally blurred for privacy).

Once a finder or selection tool is used to identify crashes, filters can be used to pull out crashes of interest. Filters allow the analyst to decide which crashes will be shown based on attributes of the crash. Figure 3-29 shows the filters for crash severity and year. The analyst can select which crashes they are interested in, and the crashes displayed will be adjusted accordingly on the map (IDOT 2012).

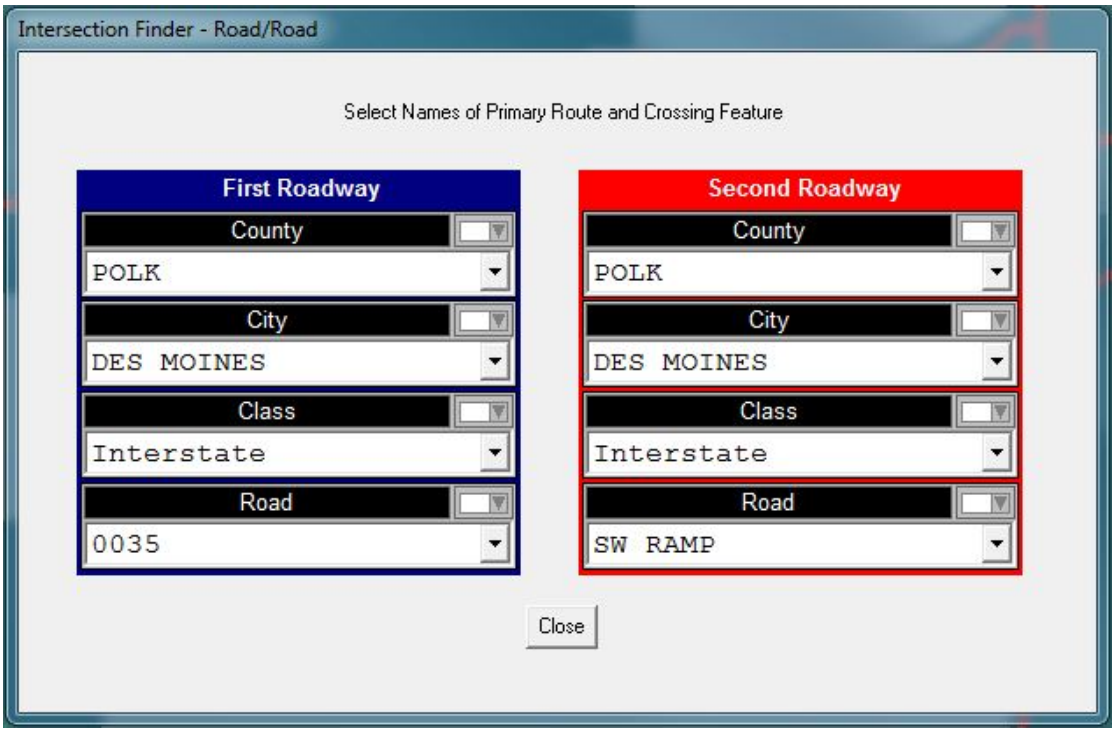


Figure 3-28. CMAT Intersection Finder dialog box (IDOT 2012).

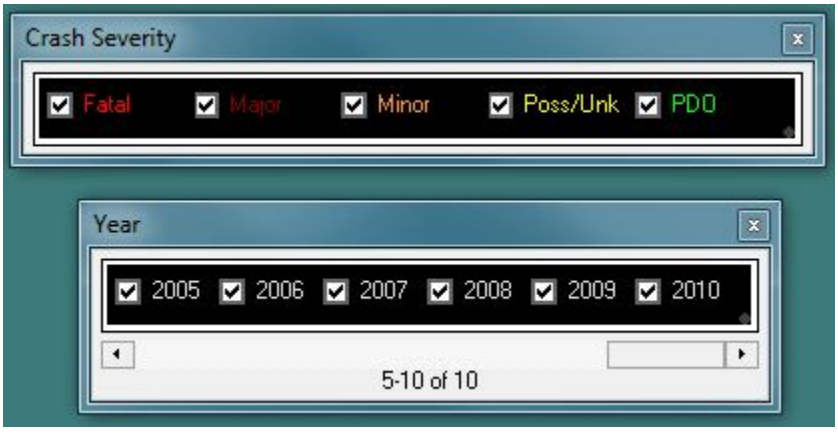


Figure 3-29. CMAT Crash Severity and Year filter dialog boxes (IDOT 2012).

This filter and finder system is effective, but is somewhat cumbersome to operate. The dialog boxes have to be opened, closed, and moved around repeatedly to be able to view the map. Some of the filters include 30 or more options and the analyst must uncheck every single one that he/she is not interested in. Filters can be saved to make this faster. Updating the system to work with query builders and logic statements would have the same ability as the filters and be more user friendly.

After a batch of crashes has been identified, charts are automatically produced summarizing the selected crashes as shown in Figure 3-30. These charts allow the analyst to create his/her own title and then print or save the chart for future use. These charts are very useful for results analysis and presentation but, like the filter boxes they can be cumbersome. Every time the crash selection is changed the charts are updated and appear in the window again. This setting can be changed, but making the charts appear manually after the selection is made can be equally cumbersome.

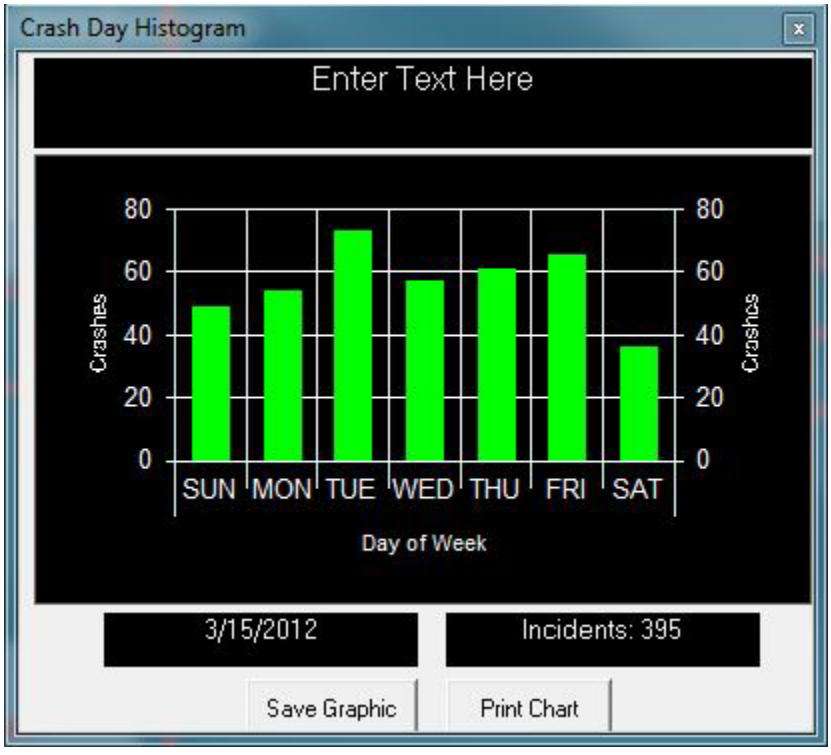


Figure 3-30. CMAT day of week summary chart (IDOT 2012).

After a selection is complete the details of each crash can be seen in an on-screen window (Figure 3-31) or printed in a PDF report. The on screen window is useful for finding a specific crash and viewing attributes about it. The list of selected crashes can be seen on the left (blurred intentionally for privacy) and each crash can be selected and viewed. An important feature here not seen in other programs is the use of text descriptions rather than codes from the crash report. Many other programs reviewed had only the numeric code entered by the officer in the crash data. Having the actual text description instead of the code makes it much easier to interpret and understand the data. By default CMAT shows the text description but has the option to view the code as seen in Figure 3-31. The box next to "Day of Week" is checked by default, so "Sunday" shows instead of the numeral. The box next to "Month" has been unchecked so the numeral "1" shows instead of reading "January". This ability to show text descriptions instead of codes or at least to have a key readily available should be essential in any crash data analysis system.

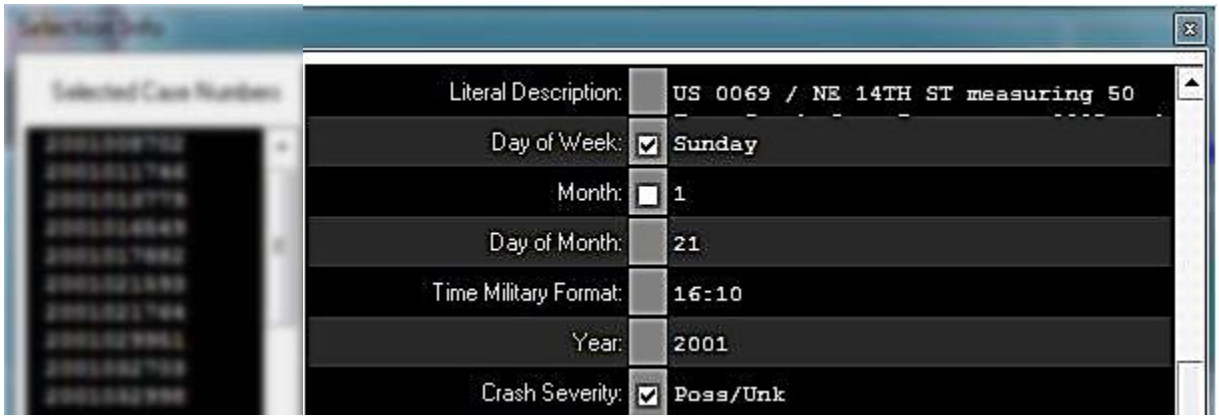


Figure 3-31. CMAT detailed crash information (IDOT 2012)
(Crash record numbers on left intentionally blurred).

CMAT allows the user to save queries and other settings for future use. Once a subset of crashes is identified they can be exported in a PDF report, as a CSV, or as a shapefile for use in other GIS systems. Because CMAT is not built for statistical analysis these export features are essential to be able to analyze crashes more thoroughly. Overall CMAT is an excellent program; it is recommended many of its features and concepts be incorporated into any new crash data analysis software that is developed.

3.5.6 SafeRoadMaps (saferoadmaps.org)

SafeRoadMaps creates maps and tools for the public to view fatal crashes across the entire United States. The system is based on showing crash points with Heat Maps as discussed in Section 3.3.2 and shown previously in Figure 3-13. The analyst can view crashes by selecting a route and buffer distance, a neighborhood, city, state, or a national map. When viewing crashes a pop-up window is used that shows information about the crash along with maps and street view as previously shown in Figure 3-17 and in Figure 3-32 (SafeRoadMaps 2012).

The important features this system uses are the "Heat" maps for hotspot analysis and the informative pop-up windows. Heat maps are limited in functionality but can be very useful for presentation. The pop-up windows provide instant information about the crash including maps and street view images without the analyst having to use a separate program. These features should be utilized in GIS-based crash data analysis.

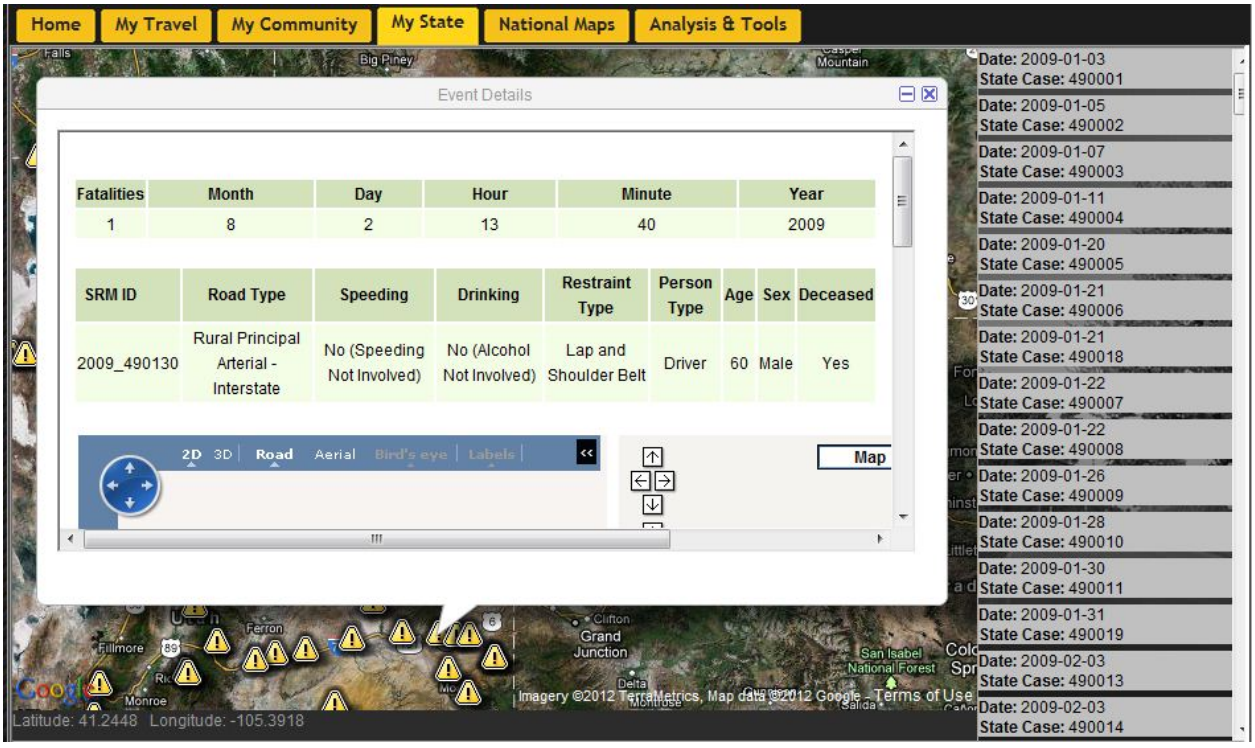


Figure 3-32. SafeRoadMaps interface shows crash information and maps (SafeRoadMaps 2012).

3.5.7 Maryland Spatial Analysis of Crashes

The National Transportation Center at Morgan State University developed the *Maryland Spatial Analysis of Crashes* system for the State of Maryland. The system was developed as a prototype in 2003, no information could be found about its continued development. The program is an online interface that helps get information about crashes to public agencies. It was designed specifically for crashes involving CMVs, similar to the UMassSafe CMV data analysis program (see Section 3.5.2). The system has a few analytical tools, but is mostly for viewing and querying crashes (Bapna et al. 2003).

The interface for the *Maryland Spatial Analysis of Crashes* system is fairly clean and accessible. Figure 3-33 shows the basic window and features. The map display is at the center with a toolbar above it. To the side are toolbars that show current functions and selections, and the current submenu (Accident Time Period) that is being examined. It is helpful to have these menus available all the time for two reasons: 1) the analyst can see what selections they have made and 2) they can quickly adjust those selections and instantly see the display update. At the bottom of the screen is summary information. This summary doesn't provide as much information as CMAT for example, but it is easier to work with because it is docked to the bottom of the screen. This report window can be used to show detailed crash reports (shown in Figure 3-33) or summary statistics of the currently selected crash features.

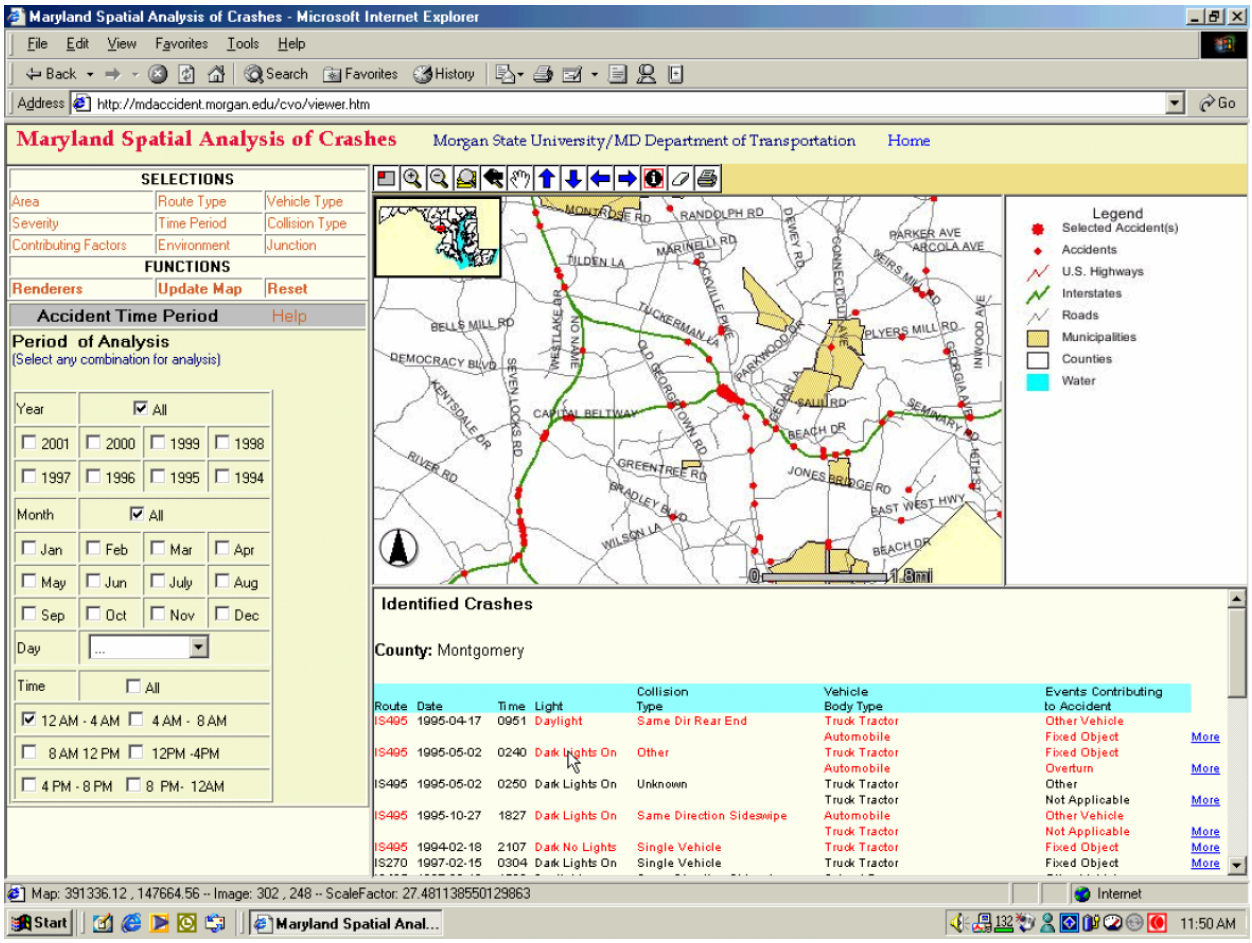


Figure 3-33. Maryland Spatial Analysis of Crashes interface (Bapna et al. 2003).

3.5.8 Esri Executive Dashboard

Along with the tools developed by Esri that have already been discussed in Section 3.4.2, Esri has also developed an executive dashboard for viewing crash data and statistics (Esri 2011a). This interface is used to view and analyze crash data from a high level, such as a city/county administrator, DOT executive, or Governor's office. The dashboard provides statistics by county, shows safety projects planned, completed, in progress, and still allows viewing of individual crash information. This dashboard has the potential to be connected with other display and analysis tools, which would make the system even more powerful. However, in its current design the system is best suited for high-level analysis and presenting broad information, not specific crash countermeasure projects.

Figure 3-34 shows the interface and features of the Esri executive dashboard. The figure shows statistics for each performance measure color-coded by county. The bottom of the screen

shows different statistics that can be viewed, and the right side shows a bar chart for county comparison. The selected county will be highlighted in red on the map and in the chart and table. Viewing crash data at a macro scale can help point out broad trends and problems that might be washed out in all the details of a micro analysis. Crash data and analysis must be presented effectively to decision makers in order for results to be seen, and high-level displays such as these are typically more effective than micro analysis tools in those situations (Esri 2011a).

3.5.9 CARE

The Center for Advanced Public Safety (CAPS) at the University of Alabama has developed a data analysis system known as CARE (Critical Analysis Reporting Environment). The CARE program was built primarily for crash analysis, but is designed to be useful for analyzing any kind of data and is not tied to crash data only. The system has also been developed with a GIS extension to allow spatial analysis. An online version of CARE and a desktop version were available for review as part of this research however; the GIS portion of the system could not be directly analyzed (CAPS 2009a).

The basis of the CARE system is to provide tools to sort, analyze, and compare data using variables in the crash data. Display graphics and some simple statistics are used to show what variables are most correlated with crashes and where safety improvements may have the greatest impact. Figure 3-35 shows the CARE online interface, which provides the best methods of querying data and seeing impact of variables (CAPS 2009a). The left panel consists of filters and variables that are used to identify which crashes will be analyzed. This allows the analyst to select filters such as ‘Crashes in Tuscaloosa County and in the city of Tuscaloosa,’ or ‘Crashes in ADOT region 5 involving commercial motor vehicle.’ These filters can be locational or from the crash/person/vehicle databases.

The desktop version of CARE takes a similar approach as the online version, but with a menu-based structure, rather than the visual approach seen online (CAPS 2009a). The series of menus and sub-menus for filtering in the desktop version work more or less the same, but are not as intuitive or as easy to use as the online version. The online version also has the graphical representations of each variable, giving easy to read summary statistics.

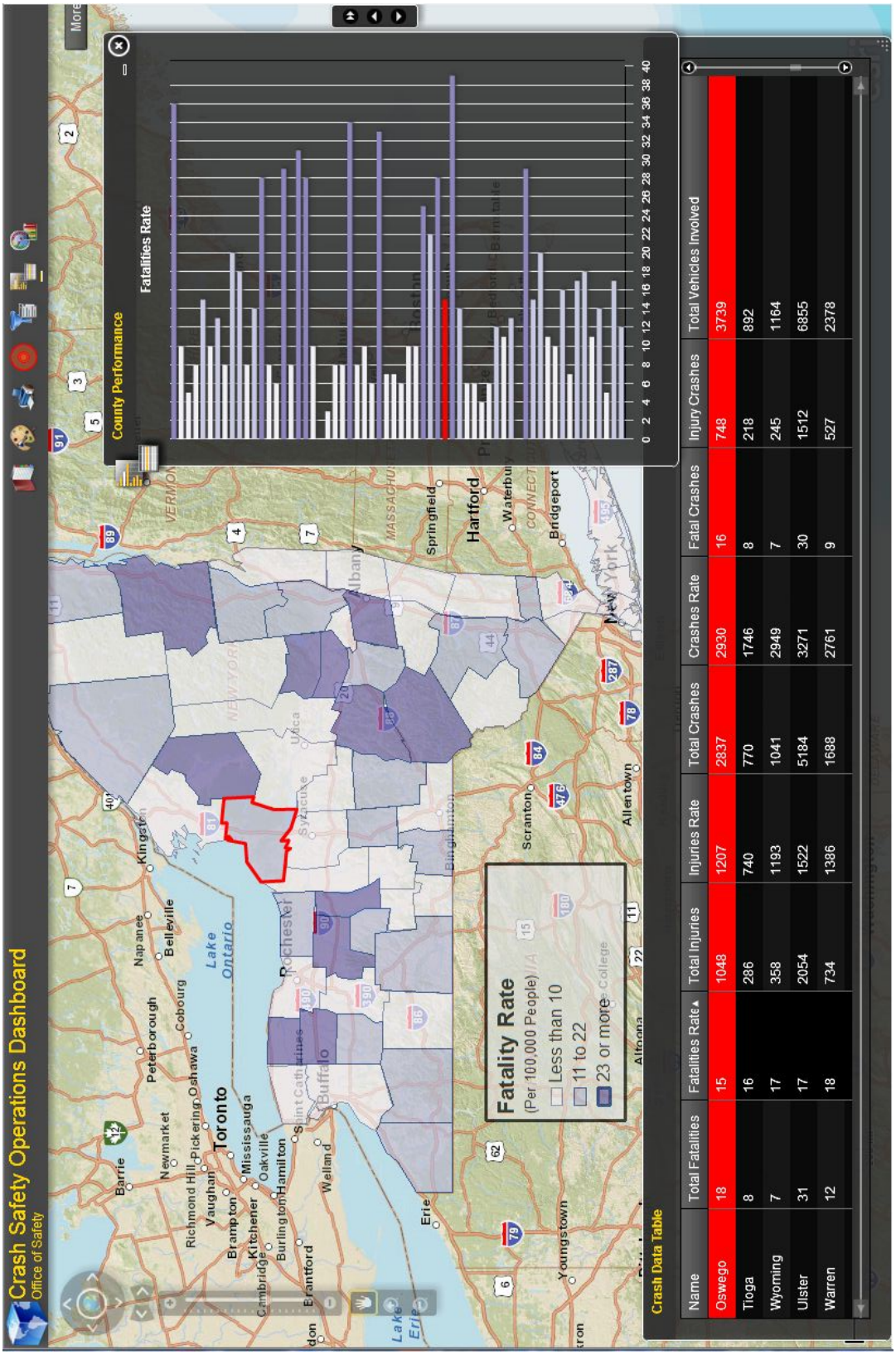


Figure 3-34. Esri Dashboard shows high-level crash statistics (Esri 2011a).

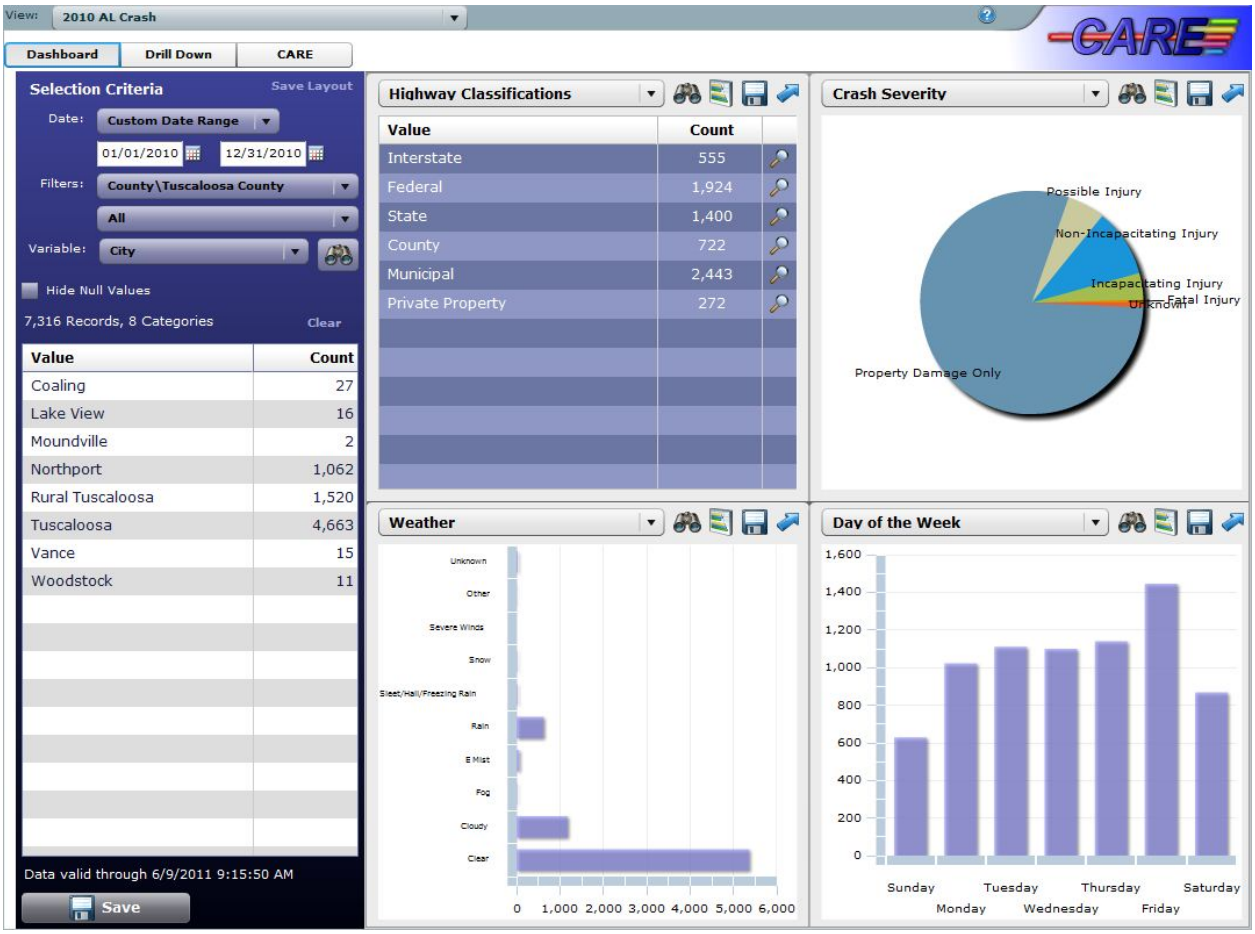


Figure 3-35. CARE online interface (CAPS 2009a).

Where the desktop program far exceeds the online version is in statistical ability. Figure 3-36 shows results of a frequency comparison of alcohol and non-alcohol related crashes by time. The desktop program is able to perform two other important statistical analyses. The first is referred to as Crosstab. Crosstab will take a set of crashes and two variables and then identify which conditions for those variables are correlated. For example, the graph in Figure 3-36 shows that non-alcohol related crashes (blue bars) are more common at low speeds, while alcohol-related crashes are more common at high speeds (red bars). The Crosstab would be able to quantify how over-represented these crashes were to help determine if the difference was statistically significant. This Crosstab analysis supplements the graphical analysis from the bar chart and helps the analyst find an effective countermeasure to address the need. The second statistical analysis tool available in the desktop program is IMPACT. IMPACT also analyzes over-representation, but works with only one dataset rather than two and will calculate how

much crash improvement could potentially be seen by countermeasure implementation. This helps in determining the cost/benefit of safety improvements.

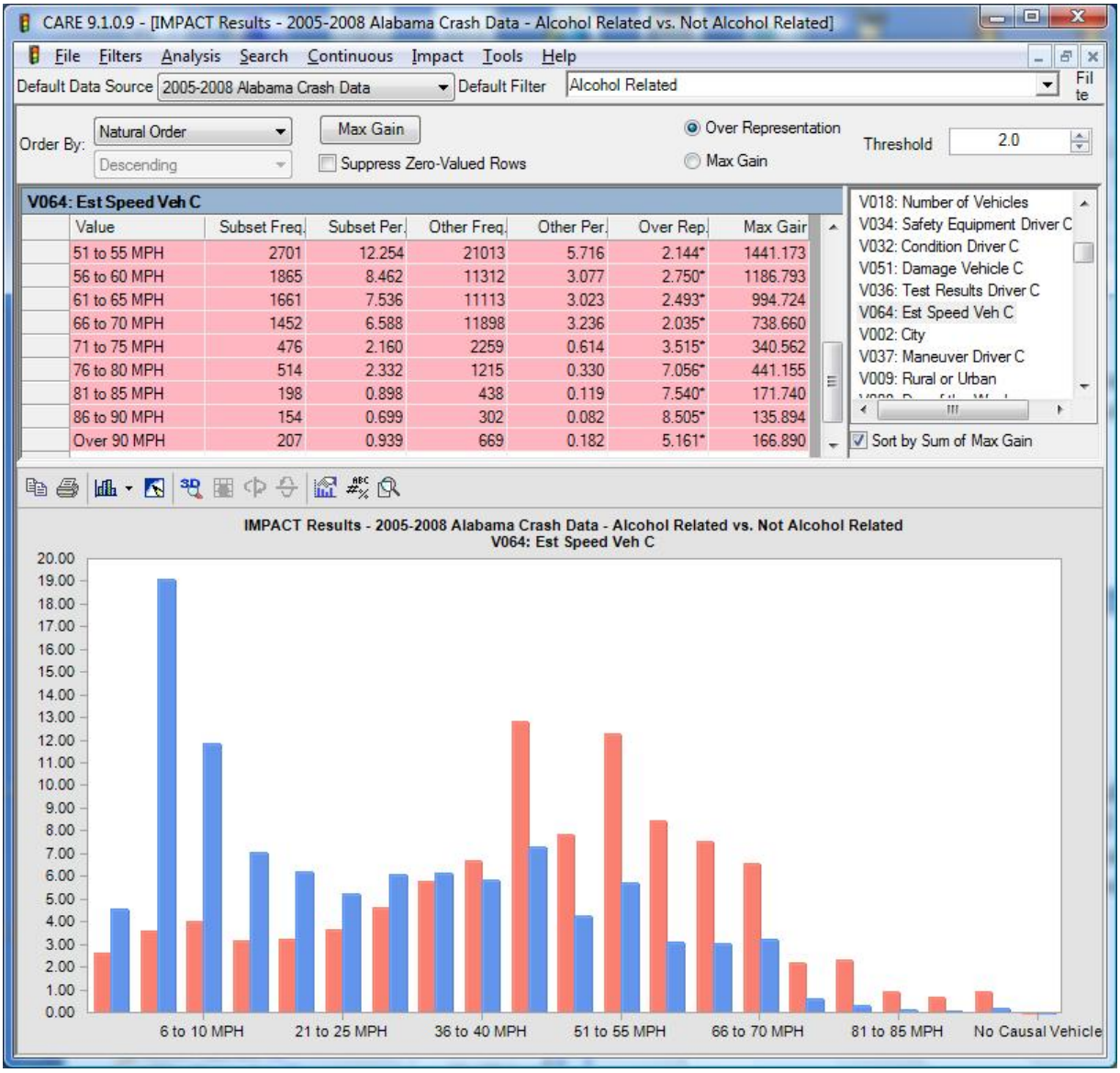


Figure 3-36. CARE desktop interface (CAPS 2009a).

Because the GIS portion of CARE could not be analyzed, limited conclusions can be presented. Figure 3-37 shows the map portion of the program and what it adds to the statistical analysis. The map gives spatial context to the data, while the line diagram provides a good view of where crashes are occurring along the route. This method of stacking in a horizontal view is unique to CARE and seems easier to understand than some other stacking methods. No mention

was made of any spatial statistics or rates calculated by the GIS system. From Figure 3-37 it is apparent that the system does some kind of hotspot analysis. No information could be found about how this analysis is done, but based on the graphic it is most likely a count of crashes within a defined segment length, similar to the strip analysis tool described in Section 3.4.2 (CAPS 2009a).

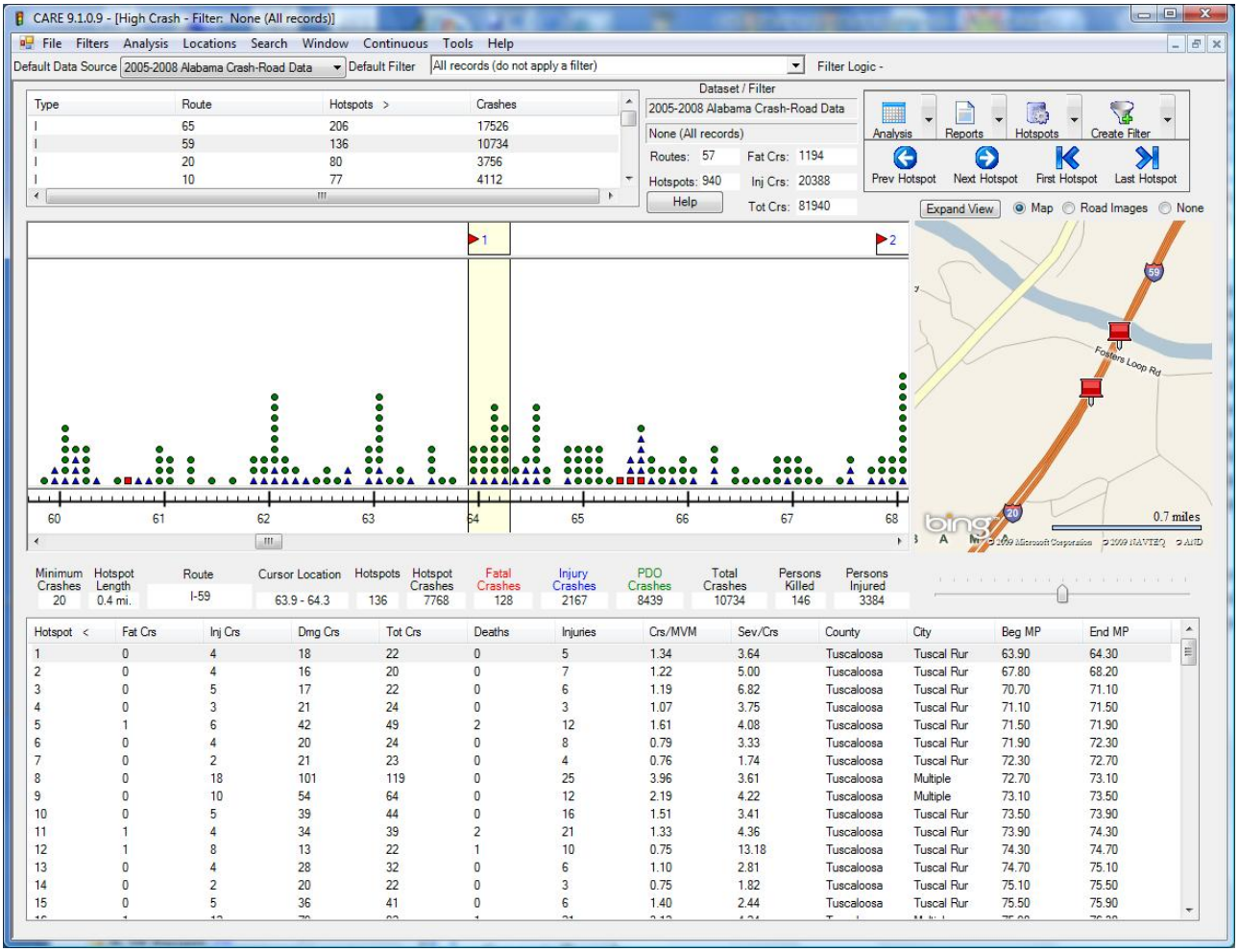


Figure 3-37. CARE mapping interface and tools (CAPS 2009a).

CARE provides one of the best overall analysis systems examined for this research. The program allows the analyst to filter and subset crashes and variables until variables or situations have been found that pose a significant risk to traffic safety, which is the primary goal of crash data analysis. The analytical tools (Crosstab and IMPACT) also provide good measures of effect

from those variables. The GIS system seems adequate, but without a hands-on analysis no definite conclusions can be made.

3.5.10 usRAP

The United States Road Assessment Program (usRAP) was created by the American Automobile Association (AAA) Foundation for Traffic Safety. The program has been piloted in several states and is now moving out of the pilot stage to become fully functional (Harwood et al. 2010). The usRAP system is not a GIS program like many others investigated in this report, but is a GIS-based method and practice of analyzing crash safety (AAA 2010).

An important advancement made in the usRAP system was to analyze all segments of the entire roadway system in a state and classifying safety across all of those routes. Most systems use only crash or demographic data, but usRAP used roadway data also. The statistics used by usRAP were only descriptive, but the method of classifying the whole state system based on a single criterion will be helpful to state DOT's and decision makers. usRAP calculated crash rates (crashes per VMT) and densities (crashes per mile) for different subsets of crashes and produced maps to represent the results. Common subsets included fatal or major injury crashes, alcohol involved crashes, and aggressive driving crashes. Crash rates or densities were also compared with similar routes in the state to obtain a crash rate ratio (Harwood et al. 2010).

Figure 3-38 shows a sample output of the usRAP analysis for the State of Utah. The image gives valuable information for the entire state but not enough detail to determine crash countermeasures. The results are shown at a very high level, which is difficult to tie to specific locations. Figure 3-38 shows where high-risk corridors are in the state, but it does little to help identify specific risk locations, why crashes are occurring there, or how to improve safety. Also, providing only a static map makes the use of the information very limited. Having the data and results in a GIS would be beneficial.

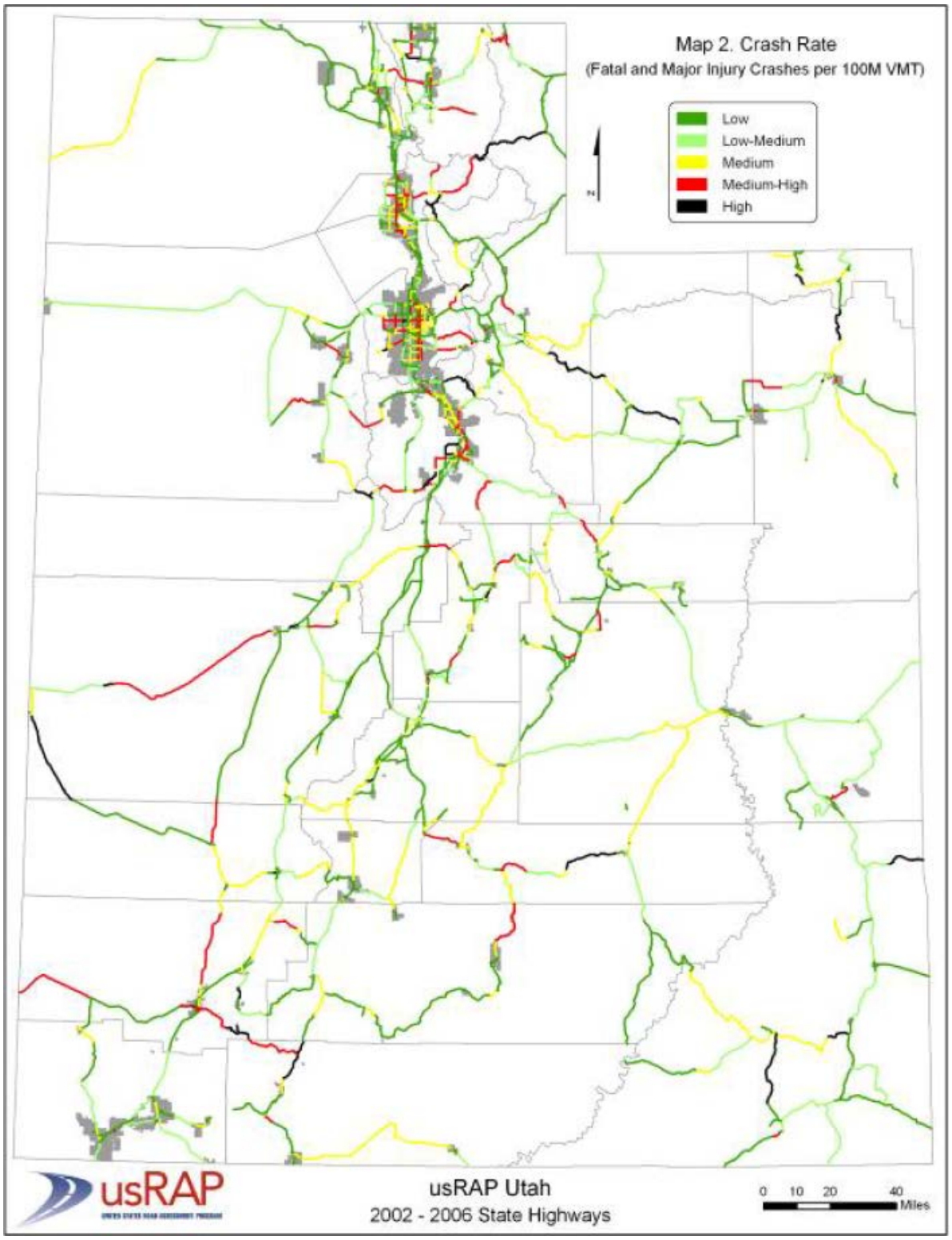


Figure 3-38. usRAP output for Utah (Harwood et al. 2010).

A second element of the usRAP program is known as the Star Rating system. This program uses roadway data collected by driving every roadway in the state to create statistical models that can predict the crash frequency of a section of roadway. The model inputs the roadway characteristics to create a star rating that signifies how safe the road is. After completing the study it was determined that the star ratings were correlated with crash frequencies, but not in all types of crashes or roadways studied. The star rating does hold significant potential though in being proactive at determining what safety hazards exist and fixing them before an obvious problem develops. This is commonly referred to as the systemic approach. The systemic analysis approach is greatly enhanced by GIS because it provides more information than is available in a crash database (Harwood et al. 2010).

3.5.11 MassTRAC

The Massachusetts Traffic Records Analysis Center (MassTRAC) provides access to traffic crash records and information for public safety officials in the state and authorized professionals. The developers have established an online portal to facilitate analysis and acquisition of traffic safety data. The program has been made available to UDOT for internal review (Jan. 2012, email communication with Barbara Rizzuti).

MassTRAC maps crashes in a GIS environment that includes many different layers of information such as schools, fire and police stations, roads, census data, highway exits, and various base maps including imagery. Having these layers available is a huge advantage when examining high crash risk areas and locations because it improves the analyst's ability to understand the location and determine what can be done to improve safety. The GIS portion of the program is completely functional online and uses Esri's ArcGIS API for Flex application (Esri 2011b). Figure 3-39 shows the MassTRAC interface.

MassTRAC allows the analyst to filter crashes, locations, and attributes a number of different ways. First, the analyst selects an analysis year(s), locations, and filters using the buttons seen on the left of the interface. Locations can be selected by county, city, census zones, user (manual) selection, and many other ways. The filters allow the analyst to select which type of crashes they would like to view. This includes data about the crash, people involved, vehicles involved, and citations/violations. This three step selection process (year, location, and filter) is

an excellent way to begin an analysis and provides a good framework for using the MassTRAC application (MassTRAC 2011).

After the initial three filters have been applied in MassTRAC, the analyst will see the map interface shown in Figure 3-39. This view provides access to different data layers and allows the analyst to zoom in and out to view each crash. The variable used to symbolize crashes is shown at the left (manner of collision is used in Figure 3-39) and can easily be changed to represent other crash variables. Tabs on top of the screen provide additional information about the selected crashes. The "Records" tab shows complete crash records for all crashes included in the analysis and can be exported for use in other programs. These records can also be sorted and then found on the map to help the analyst find a specific crash. The "Tabulations" tab provides summary statistics. It allows the analyst to choose what variables the crashes will be summarized by, what type of charts will be used, and to export the results. The "Reports" tab will generate a detailed report with standard performance measures used in Massachusetts, a map, and any trends seen in the selected data.

One of the most important features of MassTRAC is that all of the tabs, filters, and maps are dynamically linked. Changing the filter selection, map location, or summary statistics will be automatically accounted for and updated in other portions of the program. Having the tabular and geographic data linked like this in one analysis framework is the most efficient way to perform analysis and is a key feature that has been missing in other programs. MassTRAC's clean, accessible interface and easily adaptable analysis methods make it very useful in crash data analysis.

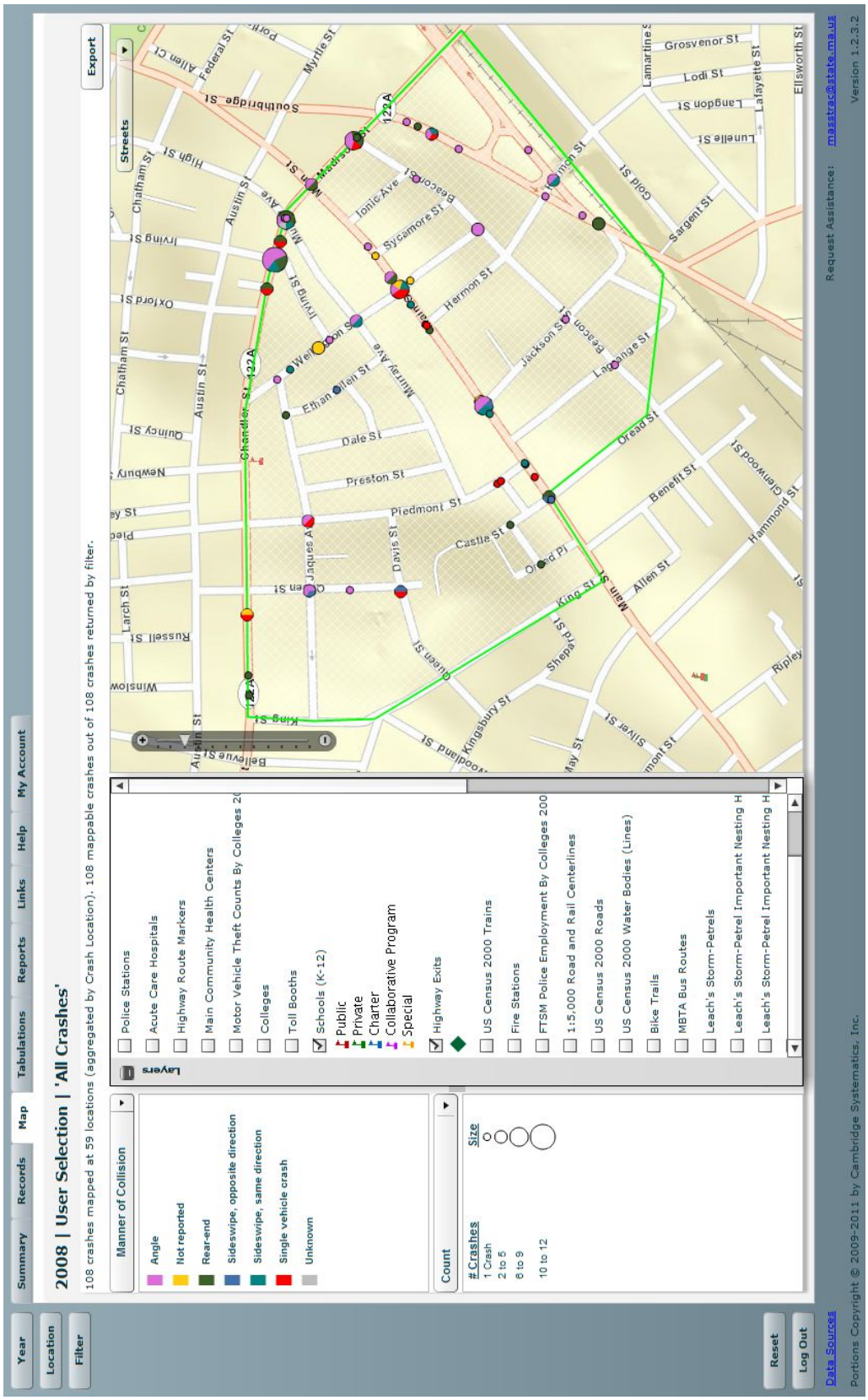


Figure 3-39. MassTRAC interface shows the GIS and filter tools available (MassTRAC 2011).

3.5.12 Section Summary

Each program discussed in this report has numerous features and capabilities that are useful in crash data analysis. This section will provide a cumulative summary of those features in order to establish a basis for how to develop a future GIS crash data analysis programs, tools, or workflows that incorporates the best features possible.

Filtering and querying data is part of nearly every analysis technique and program, and deserves detailed discussion. There are two ends of the spectrum when it comes to filtering. At the high end is ArcMap, which allows creation of logic statements and several geographic filters for selecting data. At the low end are several programs that have a simple drop-down menu that allows the user to select one variable to filter with. Some programs have tried to find a balance between these. CMAT provides a separate window for every attribute in the crash data, which allows the user to be as detailed as they want. This level of detail is preferred, but it needs to be packaged in a more user-friendly way. ArcMap allows the same detail but uses logic queries instead of selection menus. This is usually easier, but having the entire menu available like in CMAT is good too because the analyst can see all of the options. CARE also has multi-layer filtering ability, but it is accomplished with drop-down menus so it is limited to the number of menus available. A feature that wasn't available on any of the programs examined was the ability to exclude certain crashes. Excluding certain crashes can be accomplished simply by selecting all the other crashes, but having the option to exclude is often easier.

Another consideration regarding filters is that there are different types of filters. These include data filters, geographic filters, road/route filters, year filters, and others. MassTRAC has the best system for organizing these types of filters with the three stage (year, location, filter) setup. But it is still static; the user cannot use two location filters and no data filter if they want to. Other systems don't differentiate; it is just left to the user to determine how to create the filter.

An absolutely essential filter capability is that different filters must interact and update as each is changed. In CMAT adding one filter or location query would result in all results being automatically updated accordingly. The MassTRAC system also did this. If the map view is changed or a new tabulation or summary statistic is changed, the other maps and tables will be updated. CARE online reached a sort of middle ground in this regard. The initial location and data filters were layered, but the summary charts produced were not. They showed statistics but had no ability to extend the query based on those statistics. Filter selections should also be

dynamic. Some programs require the user to build a filter then load it into the program. This might add capability to the filtering but is not conducive to data mining or searching visually for high crash locations. The filters should be easily adjustable and the results should automatically be updated on the map.

A separate issue that assists with the filtering process is having summary statistics easily accessible. Several programs currently incorporate this feature. As filters are created a small table or panel should show a summary of the currently selected data such as number of crashes, type of crashes, routes selected, etc. These summary statistics should be customizable so the user can see what they are interested in. Having these available without having to calculate them or use an additional filter makes analysis easier and faster because more information is available.

Statistical tools should be linked to the mapping element of analysis. CARE provides a good example of this by presenting statistical results side-by-side with mapped results. The statistical analysis should be represented in the map. Statistical tools are often limited by computing power, but as often as possible they should be integrated into the map interface for simplicity. Statistical analyses should also be adjustable. There is no one-size fits all statistical analysis, so analysis programs should allow different methods and processes for statistical analysis.

Many programs discussed incorporate the ability to see detailed data corresponding to specific crashes or roadway segments. CMV Crash Mapping Tool, BayCitizen, and saferoadmaps.org all used very simple and informative pop-up windows to show detailed information. The pop-up used in CMV contains multiple tabs, the saferoadmaps.org pop-up had a street view window, and the BayCitizen pop-up had graphical summaries of the crash. Combining these three concepts would create a very informative way of viewing crash data. The pop-up could include the graphical summary on the first tab, street view image and even photos of the crash on another tab, and tabular crash data on other tabs.

Several other minor features were observed that are important to include. The ability to show “Heat” maps is useful for presenting results. Macro scale analysis tools like those in the Esri dashboard are important for high-level analysis. Actual data values should be shown rather than codes that were entered in the crash report as done in CMAT. Some kind of crash stacker like those available in CMAT and CARE is necessary to be able to see crashes that occur at the same location. The user should have the option to normalize data, especially crash counts, with

respect to population, length, AADT, or other variables to facilitate comparison between different geographical areas and routes. Finally, the ability to determine the potential impact of crash countermeasures as is done in CARE is valuable for early analysis of potential crash countermeasures.

Overall, the programs studied offer a wide range of methods and styles, and incorporate most of the desired features of a GIS system. The only glaring hole in all of them is the ability to analyze crashes from the perspective of roadway attributes, rather than only crash attributes. This is sometimes referred to "bottom-up" analysis where analyzing with crash data is referred to as "top-down" analysis. Bottom-up analysis is using attributes of the roadway system to identify 'hotspots' and then viewing crash data to determine potential causes and countermeasures. Top-down analysis is using crash data to find 'hotspots,' and then looking at the roadway for countermeasures. Top-down analysis is more common because it can be done with only crash data. Bottom-up analysis requires crash data and roadway attribute data, which is often very hard to acquire. In many ways bottom-up analysis makes more sense because it is using the attribute that engineers are able to control (the roadway design) to identify hotspots and then determining countermeasures by considering the crash data, which reflect human factors. Using only crash data incorporates the human factors but does not tie them to roadway characteristics that can be changed. This makes it difficult to get results out of the analysis that will actually make a difference. usRAP attempted this with their star rating system. But that system uses statistical models to predict crashes without any consideration for the actual crash data. Improved bottom-up analysis capability is the most necessary and important development for improved crash data analysis capability.

3.6 Chapter Summary

Multiple GIS programs, tools, and methods have been developed for GIS-based crash data analysis. Many of these are based off of Esri's ArcGIS software. Linear referencing is a tool that facilitates the collection, storage and analysis of crash data. The benefit of linear referencing is that it allows different datasets to communicate with each other, enabling better analysis. Visualization of crash data is the primary advantage of using GIS systems because the data can be represented graphically using colors, lines, 3-D images, gradient maps, aerial imagery, and

other graphical effects. In addition to visualization, GIS analysis tools are available to help interpret the data. Query and filtering tools enabled sorting and summarizing of data and are the first steps in beginning more in-depth statistical analysis. Esri has developed the strip, spot, and sliding scale analysis tools to help with data interpretation. Statistical methods in GIS provide numerical evidence of analysis results and are becoming more and more available in GIS.

Several states and agencies have developed programs that utilize different features of GIS for crash data analysis. All of these programs have made important steps towards developing a GIS crash data analysis framework that could be adapted to a statewide scale and used in micro analysis for countermeasure selection. In GIS systems filtering capabilities should be dynamic, easily accessible, update relative to other filters, and allow very simple or very detailed selection. Summary statistics should be shown to aid the filtering process. Statistical tools should be adaptable instead of taking a one size fits all approach. Detailed data needs to be presented in the GIS using pop-up windows or tabular presentation. The major element missing from crash data analysis systems is the ability to analyze data relative to roadway attributes rather than only crash data. This advancement will improve the ability of engineers to determine effective countermeasures.

4 DATA

The availability and quality of data is often the limiting factor when determining what crash analysis can be done and how valid the results are. This chapter will explain some general considerations regarding data use, datasets that were acquired for this research and how the data were prepared for use in analysis, recommendations for data uniformity, and an explanation of how data were used in analysis. Finally, a discussion on the value and potential of electronic crash data collection will be presented, followed by a chapter summary.

4.1 General Data Considerations

There are four general data considerations that are important before using any dataset in crash analysis. These include accuracy, availability, coverage, and usability. Each is described in the following sections.

4.1.1 Accuracy

Accuracy is the first important issue when considering data for use in crash analysis. Accuracy is important in order for the analysis to be valid and lead to real safety improvements. This is especially important in GIS analysis. In GIS analysis data layers are often merged, overlaid, buffered, or joined. Between those operations there are typically several steps of selecting and querying data also. Because the GIS analysis is multi-layered a small data error can be propagated a number of times. Quality control checks should be in place for assuring that data are accurate. Some quality control checks should be standard, automated operations with all data sets. Other quality control checks depend on the analyst being thorough and observant while working with data.

4.1.2 Availability

Availability is the second data consideration. Datasets should be widely available to encourage analysis and sharing of results. If data are protected and not shared then they are of little to no use. Tools such as UPLAN (UDOT 2011b) and the Utah Automated Geographic Reference Center (AGRC) (2012) are invaluable when it comes to data availability because they provide one location from which data can be shared. Acquiring data now is not the only consideration, but how available data will be in the future is also important. Undergoing a massive data collection effort and building an analysis around these data is of little long-term value if there is no plan in place to continue to collect that data. If the data are not re-collected or updated, accuracy will become an issue in the future. In the course of this research several programs and research studies were reviewed in which a unique one-time use dataset was developed. This was necessary for the project to be completed, but offers little value beyond that for further analysis.

4.1.3 Coverage

Coverage is the third data consideration. It is preferred that data include the entire state. All datasets need to be checked for the extent of coverage and any future plans to extend that coverage. Missing large sections of data will affect results because the missing section will be lost in the analysis. If the missing sections are not totally lost in analysis then at minimum some variables may be lost. Coverage is also lacking sometimes after new construction or route realignment. When either of these occurs it can take some time to update master centerline files.

4.1.4 Usability

Usability is the final data consideration. It is important to understand the types of data and how they can be used in analysis. Some datasets represent only geographic features with very few attributes. These are important for identifying location but offer very little information that can be used in analysis. Other datasets may not match geometry very well or may be in a format this is difficult to work with. A benefit of using ArcGIS is that there are many tools that convert datasets. When these tools are used properly they can help turn most data into something useful for analysis.

4.2 Dataset Preparation

This section will review datasets that were used in this project. Discussion will include how each was acquired, steps needed to prepare them for analysis, as well as any important issues regarding the data that need to be considered. Table 4-1 provides a summary of the datasets and their source, format, and future availability. Appendix A provides a step-by-step recounting of what was modified for each dataset in order to prepare it for the GIS model.

Table 4-1. Dataset Source Summary

Dataset	Source	Format	Future Availability
State Routes	Utah AGRC	LRS Feature Class	Updated Regularly
Crash Data	Scott Jones	CSV Tables (Excel)	Updated at least Annually
AADT	Frank Pisani, Lee Theobald	Excel Spreadsheet	Updated Annually
Truck AADT	Frank Pisani, Lee Theobald	Excel Spreadsheet	Updated Annually
Speed Limit	Larry Montoya	Shapefile	TBD
Functional Class	Interplan-Charles Allen	Excel Spreadsheet	TBD
Through Lanes	UGATE	KML file	TBD
Urban Code	UGATE	Shapefile	TBD
Road Geometry	Gary Kuhl	Access Database	New inventory underway
Skid Index	Gary Kuhl	Excel Spreadsheet	Half of state updated each year

4.2.1 State Route Inventory

The Utah State Route inventory file is made available through the Utah AGRC and is maintained by UDOT staff. The file can be accessed by any online user and imported to ArcGIS, or connected through the AGRC database server. The file is updated regularly as changes are made to mileposts, routes are realigned, or new routes are constructed. The file is linear-referenced and includes route references, directional indicators, and milepost locations. Each route in the file is created as one segment, even if there are breaks in the route. This route inventory file is absolutely essential for GIS analysis and linear referencing. Linear referencing requires a base route file with mileposts in order to use other data. No processing or modification was done to this file in order to use it for this research.

An important element of the route inventory file is the use of labels to indicate unique routes. Each route is referred to by a four digit route code that includes the route number with leading zeroes to equal a total of 4 numerals. For example, Route 6 is indicated by “0006” and route 154 is indicated by “0154”. This four digit code is referred to as the route ID. In addition to the four digit route ID a letter is used to indicate the direction of measure on the route. The route ID plus the letter indicating direction form a unique five digit code reference for each route and is known as the label.

The “P” direction code indicates that route milepoint measures increase in the positive direction. In the State of Utah the positive direction is north and east, meaning that mileposts increase from west to east and from south to north. The “N” direction code indicates that mileposts increase in the negative direction. This is used for the southbound or westbound portion of divided highways. Finally, the “X” direction is used as a surrogate measure for the “N” direction. The “X” direction follows the same geometry as the “N” direction but has milepoints that match the “P” direction. Figure 4-1 shows an illustration of the different route directions and their associated milepoints. In a real GIS the “X” and “N” route would overlap but they are kept separate in the figure for clarity.

Because the “N” direction routes accumulate mileage backwards there are several issues that can arise with representing data on “N” routes. Milepoints in datasets are all recorded in the “P” direction. Data cannot be directly transferred from the “P” route to the “N” route because the milepoints are backwards. The “X” route allows easy transfer from the “P” route because the milepoints are the same so the direction indicator is the only change that needs to be made. For this research the “P” direction will be used for all routes and the “X” direction will also be used where it is available. The “N” direction will not be used.

All interstate routes and some expressways are represented by three separate lines as shown in Figure 4-1, each indicating one of the three directions. Other routes have only one line which is represented by the “P” direction only. The exception to this is non-interstate routes that are fully or partially divided. These routes have the “P” direction for their entire length and an additional “N” direction line for only the divided portion, with no “X” direction line. Because there are only a few of these sections in the state and they are not continuous across the entire route there is very little data available for them. This is one aspect of UDOT’s current data management practices that should be addressed.

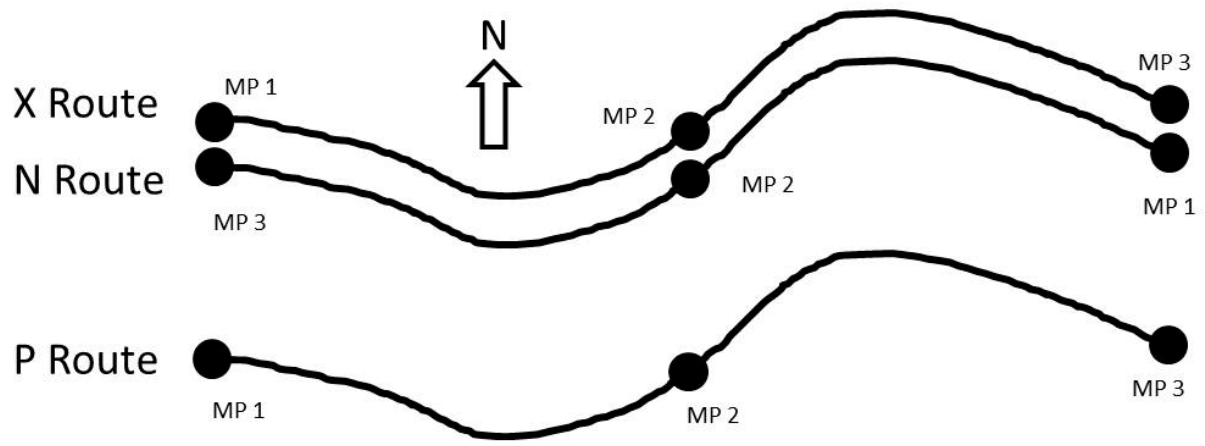


Figure 4-1. Illustration of route directions and milepoints.

4.2.2 Crash Data

Crash data are maintained by UDOT Traffic & Safety Division. Data are collected through paper forms filled out by police officers. The data are then manually transferred to digital format after which UDOT staff organizes the data. The most recent dataset at the time of this report was completed for the years 2006-2010. It was organized into seven different files described below.

1. crash: General information regarding the crash, events, outcomes, and roadway
2. crash_comments: Reporting officers' written comments about the crash
3. crash_location: Location information including route, milepoint, latitude and longitude, and jurisdictional information
4. crash_rollups: Contains summary information about the crash including injuries, fatalities, vehicles involved, people involved, and others
5. motorcarrier: Information about motorcarriers involved in crashes
6. people: Information about each person involved in a crash
7. vehicle: Information about each vehicle involved in a crash

Each file contains reference ID numbers that can be used to relate one dataset to the others. The crash_rollups file is a new data file that is still in development. Most of the fields that will be included in the rollups file will be quantity indications (i.e., number of vehicles), or a binary indication (yes or no) to indicate if a certain thing did or did not happen (such as alcohol

involved or not). This rollups file can be used for basic crash information and then the more detailed files can be used for in-depth investigation.

Three of the above files were used to create the crash dataset used in the GIS model for this research. These include the crash file, the crash_location file, and the vehicle file. The crash_location file contains most of the information needed to map crashes on the State LRS including the route ID and milepoint. It also includes lat/long coordinates that can be used to map crashes in the GIS environment. Having lat/long is important because it preserves the absolute location in case the LRS changes. The only missing information in the crash_location file is the direction indicator for the route reference. The vehicle data file is needed to get this information. The vehicle file contains a "TRAVEL_DIRECTION_ID" field indicating which direction the vehicle was traveling at the time of the crash. This can be used to determine which route direction the crash should be assigned to. If the vehicle direction indicates north or east or is a non-interstate route then the crash is assigned to the "P" direction, if it is a crash that occurred on the interstate and indicates a south or west direction then it will be assigned to the "X" direction. This method is effective but has a few sources for potential error that need to be understood.

The first source for error in interpreting the travel direction is the fact that many crashes have multiple vehicles involved, often with different directions of travel. Because the crash involved vehicles traveling different directions it is very possible that different aspects of each roadway were involved in causing the crash; however, it is more likely that attributes associated with the route direction of the at-fault vehicle were more involved. Because of this, the direction of the at-fault vehicle is used to determine the crash location. This leads to another potential error because the at-fault vehicle cannot be directly determined from the crash data. In order to account for this the general assumption was made that officers list the at-fault vehicle first when reporting a crash. This is not likely to always be true, but is consistent with what officers are instructed to do.

The second source for error in interpreting the travel direction comes in the officers' indication of direction in the crash record. The officer indicates which direction the vehicle was traveling, not the major direction of the route. A north-south route could have meandering sections that travel east-west, (and vice versa) and many routes are not exactly aligned with their nominal direction. For example, Interstate 215 is technically a north-south route, but has large

sections that are entirely east-west. Advanced programming logic could be used to check each crash location and find the exact direction, but that was beyond the scope of this research.

Even with these two sources for error in interpreting the travel direction, it is believed that the number of crashes incorrectly located on the LRS due to these errors is very small. Because all non-interstate routes use only the "P" direction there is no opportunity for error on those routes, and most Interstates in Utah are fairly straight and follow their nominal direction so the risk of error on these routes is fairly small also.

Once the route direction indicator is acquired from the vehicle file the crash file is added to the crash location file so that some basic crash information will be contained in the dataset. In the future the crash_rollups file will be used instead. After this file was created it was imported into ArcGIS and mapped on the State LRS.

The method chosen for locating crashes in GIS needs to be carefully considered as it will affect data quality. Most systems use absolute geographic referencing such as lat/long to locate crashes. Other options include geocoding and linear referencing. There are several pros and cons with all of these methods that will be discussed further in the following sections.

4.2.2.1 Latitude/Longitude Crash Location

Lat/long references are preferred for locating crashes in GIS because it is the most precise, is easy to store the information, and the location won't change with time. It does require that the crash be located geographically, either by the officer filling out the report or during processing. A significant drawback is the increased difficulty in having the crash data "communicate" with other data during analysis. The lat/long reference does not guarantee that the crash will match the roadway exactly, meaning the crash point will not exactly overlay or intersect other datasets. This can pose accuracy problems when transforming the data or connecting different roadway attributes to crashes. Another problem is that route realignments or interchange reconfigurations can move roadways enough that a crash will no longer appear to be on the roadway, meaning it would have to be moved manually or end up as lost data.

4.2.2.2 Geocoded Crash Location

Geocoding uses address systems to locate where crashes occur. Researchers have shown that geocoding can potentially be very labor intensive and possibly inaccurate (Pulugarth et al. 2007), but is often the only option due to lack of useable data. Geocoding requires a road

network with names and addresses associated with them. The crash can then be matched to the address. Unfortunately information is often missing, making it impossible to locate crashes. Algorithms can be developed that will locate crashes within an acceptable margin of error. Some studies that use geocoding have shown location error as high as 25 percent (UMassSafe 2006). Geocoding is most common in urban areas because address systems are fairly complete and roads are too small and discontinuous for linear referencing to be feasible.

4.2.2.3 Linear Referencing Crash Location

Linear referencing uses a route and milepoints along that route to locate where crashes occur. An advantage of LRS methods is that data can easily interact with multiple datasets as discussed in Section 3.2. Different data sets in the LRS can merge attributes and create new segments as necessary to match roadway attributes. Linear referenced data can also be used the same way lat/long data could in most analyses. The disadvantage to using LRS data comes when routes change. If a route is realigned or re-measured, crash points may no longer be accurate because the crash point is tied to the milepoint that moved with the realignment. The typical answer to this issue is to increase every crash milepoint by the amount of roadway that was added in the realignment. However, this is not always accurate or correct. Different sections may be realigned or re-measured but, especially in re-measure situations, the changed value may not affect all milepoints proportionally. While this is a disadvantage it is a better situation than if routes change and only lat/long data are available. With LRS the crash will always be located on the roadway, even if the milepoint is slightly off. It is likely that a hybrid database that stores both lat/long and LRS data should be used. Data could be located with whichever method is preferred for the analysis and adjusted in whatever method is easiest if routes change.

4.2.3 AADT and Truck AADT

AADT data are recorded in permanent and temporary count stations located around the state. These data are used to estimate AADT for all other routes. Every year a master file is created with AADT for all state routes. For this research AADT data were acquired in Excel spreadsheet format. Spreadsheets are very easy to import into ArcGIS and map to the state LRS. The master data file created includes AADT from 2008-2010 along with truck counts and percent trucks for 2010. Historic AADT data were also acquired and prepared. A master historic

AADT file was created with AADT data for all state routes from 1981 to 2009. Historic truck AADT data were available as separate files for each year from 2003 to 2009. No major processing of the AADT data was necessary to prepare it for use in ArcGIS.

4.2.4 Speed Limit

UDOT Traffic & Safety recently created a shapefile of speed limits across the state highway system for UDOT. The file was not originally an LRS file but was easily modified to enable use in the LRS. All that was required was creating the 4-digit route ID from a simple numeric identifier and adding the direction as milepoints and speed limit values were already included. It was noted that several routes were missing from the file. Speed limits were acquired for these routes using other datasets and added to the master file. According to UDOT staff the master speed limit file is now up to date with the missing routes. Another file has recently been created with points indicating locations of speed limit signs. This could also prove useful for analysis in the future, but was not incorporated into this research.

4.2.5 Functional Classification

Functional classification data were acquired from InterPlan, a consultant firm for UDOT Traffic & Safety. The file was a spreadsheet containing LRS information for each route and the associated functional classification segments. After adding the "X" direction to this file it was mapped in ArcGIS on the State LRS. This file was not created as part of any standard UDOT project, but rather was a one-time effort. Because of this there are no definite plans for updating it in the future. Because functional classification data are essential for the statistical model plans need to be put in place to keep this information up-to-date.

4.2.6 Through Lanes

The through lanes file was downloaded as a KML file from UGATE, UDOT's online data portal (UDOT 2012). The original source of this data is not known, and the future use of UGATE is to be determined at this point, so work needs to be done to ensure that these data will be available in the future.

The KML file included route ID, milepoints, and the number of through lanes. Through lanes does not include any auxiliary lanes such as acceleration/deceleration lanes, median lanes, turning lanes, or truck lanes. KML files are built for Google Earth but can be directly imported to ArcGIS; however, the import is not a LRS-friendly file. To use the file with LRS the route, milepoints, and number of lanes information was extracted from the imported dataset table using Excel. From this a LRS through lanes file was created.

4.2.7 Urban Code

Urban code data are derived from census information for the state. Urban areas in Utah include rural, small urban, Salt Lake City, St. George, Provo-Orem, Ogden-Layton, and Logan. These data are not directly applicable to crash analysis. Demographic data like these are more applicable for high-level analysis and generation of general crash statistics. The best possibility for use of these data currently exists in the potential to analyze the rural to small urban transition areas. Drivers in rural areas often operate in isolation at high speeds, which can lead to safety problems at transitions to small urban areas that are surrounded by rural routes. It is also good to use when creating analysis segments because it will separate rural and urban areas on highways that pass through both.

The urban code data were also downloaded from UGATE as a shapefile (UDOT 2012). Some slight modifications were made to the shapefile data to enable mapping on the State LRS. As mentioned the future use of UGATE is unknown and the original source of this data is also unknown. If the urban code data is desired for continued use in the future work must be done to ensure that it is updated and available.

4.2.8 Road Geometry

Road geometry data were collected from a 2009 project to inventory the entire State Highway system. A data collection van recorded several different roadway design and condition variables and took street view images. The images and their associated data are now available online (UDOT 2011a). All the data from this project was packaged in an Access database and made available for this report.

Data from the road geometry data report was opened in Excel and prepared for mapping on the LRS. Curvature (radius), grade, and cross slope were the only values used from the

geometry data. Radius data contains the radius in feet of curved sections of highway. A positive value indicates a right curve and a negative value indicates a left curve. Grade records the longitudinal slope of the roadway, with positive values indicating upward slope and negative values indicating downward slope. Cross slope records the cross section slope of the roadway. Positive values indicate a downward slope to the right, with negative values indicating a downward slope to the left. Each of these data values are direction-dependent, so it is important to know which direction the vehicle was traveling when collecting the data. To accommodate this, a field titled "Direction_Collected" is present in the dataset to indicate which direction the vehicle was driving when recording the values. In order to understand each value correctly this field should be checked. In the future it is recommended that the data be normalized to a single direction to avoid possible confusion.

After preparing the geometry dataset and mapping it in ArcGIS, accuracy issues were observed that need to be considered before utilizing the curve radius data in any future analysis. The first problem is that the radius values are influenced by how the data collection vehicle operates, so they do not always reflect conditions that the driver experiences. Figure 4-2 shows how one large and fairly significant curve appears in the data to be a sequence of shorter curves. In addition, there are several tangent sections (radius of zero) between each of these curves. The result of this is that crashes occurring on this curve will be located on different geometric segments, even though they all occur on the same curve. In addition, many of the crashes would show a "0" value for radius, even though the crash occurred on a curved section. Figure 4-3 shows another example of this where two curves occur very close with a short tangent between them. Geometrically the curves have to be connected by a tangent, but in reality the driver does not experience this. The roadway appears to begin the second curve as soon as the first is finished. A final example of data inaccuracy is shown in Figure 4-4. Here two very sharp radius segments appear where the road is obviously straight. It is likely that these were caused by a lane change or similar maneuver made by the data collection vehicle. All three of these situations are common throughout the state.

Figure 4-2. Example of inconsistencies in existing geometry data (radius in feet is shown in white).

Figure 4-3. Example of short tangent between two curves (radius in feet is shown in black).

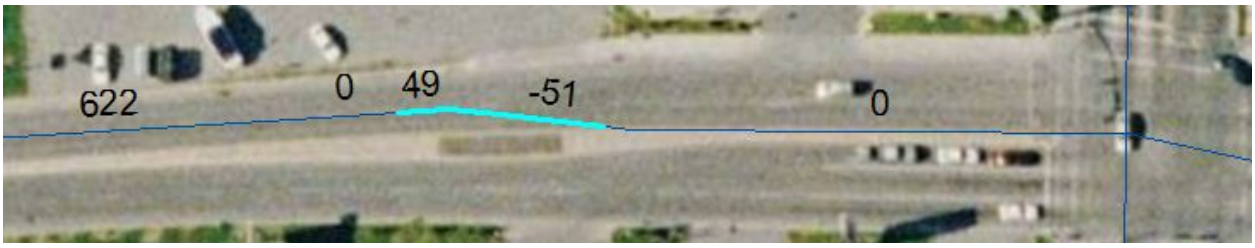


Figure 4-4. Example of radius data caused by van maneuver (radius in feet shown in black).

A second inaccuracy caused by geometry data is that crashes can be incorrectly located to a segment that does not reflect where the crash occurred. Officers typically record the milepoint location of a crash to a tenth of a mile. This location is determined purely by their judgment and by any milepoint markers within sight. Because of this, it is very easy for crashes to have a

recorded milepoint that is off by at least a tenth of a mile and up to a half of a mile. In the geometry file over half of the 25,633 segments in the state are less than a tenth of a mile in length. Considering the inherent error in milepoints recorded by officers and the abundance of short segments, at least half of crashes could very easily be located on an incorrect segment in the geometry file, which would result in inaccurate results when used in the statistical model.

The errors outlined with respect to the road geometry files will limit the ability of an engineer or analyst to accurately identify high-risk crash locations based on road geometry. Careful examination needs to be done on the geometry data to determine a way to account for possible inaccuracies. One option is to develop an algorithm that will aggregate radius data based on adjacency of curved sections. Segments could be aggregated to remove unnecessary tangents, or they could be rated on a scale based on the general level of curvature of the roadway in a given analysis area. This will be more useful as input into the model and also help account for the effect of curves that occur in succession. The detailed geometric data, even with its inaccuracies, is still valid for micro-scale analysis, but is currently not considered suitable for input into the statistical models. Because of these inaccuracies the geometry was not used directly in this research; however, the data were prepared so that it could be used and examined in future research.

4.2.9 Skid Index

Skid index data were collected in 2009 and 2010. Between the two years every route in the state was recorded at least once. The data were provided in an Excel spreadsheet with point locations for the milepoints and all other route information already present.

Two separate files were created in ArcGIS for using Skid Index data. A point file was created to show the actual point where the roadway was tested along with the corresponding value of the skid index results. To help with using the data in the model a segment file was also created. This file was created by turning each point into a segment that covered half the distance between adjacent points on either side. This provides a generalization of pavement condition, and makes the data more useful in the statistical model because it can be overlaid with the other LRS data. However, because the segments are very small they were not used in the current research, but were prepared so that they could be used and examined in future research.

4.3 Recommendations for Data Uniformity

Data uniformity is critical to ensure accuracy, simplify processing, and to allow easy addition of new data to the model. Because data are collected by different departments within UDOT and there are different standards for each, it is difficult to develop an agency wide standard. However, because of how data are used in GIS the only element of the data that has to be consistent is the column headings. These headings are used to identify which data to use in processing. Having uniform data headings will allow any dataset to be inserted into the model without having to modify automated processes or manually prepare the data. The following list contains five data fields that are recommended for all datasets. These fields correspond with what is required for use with the State LRS and for use in models developed for this research. All other data can be recorded as desired. Many of the datasets provided by UDOT for this research already follow these naming recommendations.

1. "ROUTE_ID": Contains 4 numeric digits with the route number and leading zeros
2. "DIRECTION": Contains P, N, or X corresponding to the route direction
3. "LABEL": Five digit code with the ROUTE_ID and DIRECTION fields joined
4. "BEG_MILEPOINT": Beginning milepoint of the segment
5. "END_MILEPOINT": Ending milepoint of the segment

4.4 Dataset Use in Crash Analysis

This section will discuss the use of data in crash analysis. This includes three concepts for how data are used, selection of how to store roadway data, and using data for roadway segmentation. More detail for how data are used in analysis is presented in Section 6.2 of this report.

4.4.1 Three Concepts for Data Use

There are three basic concepts for how data can be used in crash analyses. The first is using data as a variable in a statistical model. This is discussed in more detail in Chapter 1 of this report. Using data as a variable input in a model allows the model to better determine if a

location is a hotspot. A similar use of data is calculating normalized crash rates or densities based on a variable.

The second concept for data use is identifying homogeneous segments of the roadway. When analyzing crashes from the perspective of roadway sections it is necessary to break the roadway into segments that have constant attributes throughout. This helps to eliminate potential confounding variables. LRS offers tools that are uniquely suited for accomplishing this as will be described in Section 3.2.2. Several issues regarding segmentation exist that will be discussed in depth in Section 4.4.2.

The third concept for data use is examining data with good engineering judgment to help identify crash causes and potential countermeasures. This is the most inexact use of data, but is always necessary in order to translate statistical outputs into projects that will improve roadway safety and reduce crashes.

4.4.2 Roadway Segmentation

An essential step in analysis is to divide the roadway into segments for analysis. This process is fairly common in crash data analysis, especially when performing an analysis from the perspective of roadway design characteristics (Harwood et al. 2010). The roadway is divided into homogeneous segments with regard to the study variables. Creating homogeneous segments is important because it allows the analysis to be connected to the actual roadway, which helps in identifying potential crash countermeasures, and it eliminates potential confounding variables. Another method is to use a defined increment to segment the entire roadway. This is used because it eliminates issues that can arise with varying segment lengths.

A common problem with roadway segmentation is that it produces very short segments. These are impractical because a few crash occurrences on a small segment would result in a very high crash rate. It also can result in data accuracy problems due to issues described in the road geometry data section (Section 4.2.8). Bayesian statistics can account for short segment lengths because of the sampling method used (see Chapter 1), but it is still preferable to have longer segments.

usRAP used a method similar to that proposed in this research for roadway segmentation (Harwood et al. 2010). Using roadway inventory data from each state piloted the researchers broke the roadway into segments based on the characteristics in the data. This resulted in many

very small segments, which were determined to be unrealistic for analysis. To account for this, the researchers developed an algorithm that could be used to join short segments to adjacent segments. These rules included factors such as segments on the same route in the same county, speed limits within 5 mph, AADT within 20 percent, or segments in towns with population less than 5,000. In some cases this still did not remove small segments, so more rules were relaxed to join segments within larger ranges of AADT and speed limit (Harwood et al. 2010).

In research conducted for this report problems were also encountered with very short segments, especially when geometry and skid index data were used in segmentation. Table 4-2 presents a comparison of how segments lengths changed as each successive dataset was incorporated into the segmentation process. The "AADT" column represents segments as they exist only in the AADT file. The "Speed Limit" column represents segments based on homogeneous AADT and speed limit, and the final column shows segments based on all of the datasets. Because this is just an illustration, some datasets were not included in the table. Creating segments with AADT, speed limit, and through lanes results in an average length of 1.64 miles and approximately 57 percent of segments longer than a half mile. The upper quartile is actually less than the average for through lanes and geometry. This is probably caused by some very large segments that skew the average. Once geometry and skid index were added to the segmentation, segments became extremely short with over 90 percent less than a half mile in length. This illustrates the dilemma that exists when trying to incorporate geometry data as was described in Section 4.2.8. The geometry data is considered important for analysis, but is not practical to use in its current condition.

There is no universally accepted optimal segment length for analysis (Harwood et al. 2010). usRAP suggests that 2-3 miles is adequate, but many other corridor analyses, including some done in Utah, use quarter and half mile segments. For this research it is assumed that segments created using any data besides geometry and skid index will produce segments of sufficient length. Additional work needs to be done to develop ways to incorporate geometry and skid index data, and to determine what an optimal segment length is.

Table 4-2. Comparison of Segment Lengths with Progressive Segmentation

Segment Length (mile)	Datasets Included: Progressing Left to Right									
	AADT		Speed Limit		Through Lanes		Geometry		Skid Index	
Segment Count	2127		3405		3935		28,057		33,308	
Average	3.21		1.98		1.64		0.23		0.19	
Upper Quartile	3.71		2.01		1.57		0.22		0.23	
Lower Quartile	0.64		0.34		0.27		0.04		0.04	
Longest	50.06		49.38		49.38		21.28		4.42	
Shortest	0.0320		0.0020		0.0010		0.0010		0.0000	
>5 miles	387	18.2%	350	10.3%	311	7.9%	51	0.2%	0	0.0%
>2 miles	837	39.4%	857	25.2%	818	20.8%	344	1.2%	7	0.0%
>1 mile	1341	63.0%	1427	41.9%	1425	36.2%	1017	3.6%	61	0.2%
>=0.5 mile	1787	84.0%	2170	63.7%	2235	56.8%	2781	9.9%	3278	9.8%
<0.5 mile	340	16.0%	1235	36.3%	1700	43.2%	25276	90.1%	30030	90.2%

4.5 Electronic Crash Data Collection

Current practice for crash data collection in Utah is for officers to fill out a paper report with crash information. The form is then submitted and recorded digitally by hand. This system allows for potential human errors when the police officer is recording the information, and again when it is being recorded digitally. In addition, it takes time to complete this process and for data to become available for analysis. Collecting crash data digitally in real-time provides significant potential for improving data accuracy and availability when compared with existing methods (Carmody 2011). In 2010 The National Highway Transportation Safety Administration (NHTSA) released a report titled *E-Crash: The Model Electronic Crash Data Collection System* explaining current practices, advances, and recommendations for electronic crash data collection. This section of the report will explain some basic issues and considerations regarding electronic data collection. At least 15 states have already implemented an electronic reporting system (DeLucia and Scopatz 2010).

Developing an electronic crash data entry system for use by police officers would result in more accurate geographic data. Electronic entry would allow officers to locate exact coordinates where the crash occurred, guaranteeing that the crash could be accurately mapped. Using the same mapping technology officers could draw an electronic crash diagram on an aerial

image. This would be easier than hand drawn diagrams, and it would help the officers to be more detailed because they can connect events with locations on the map. This would also be incredibly valuable for analyzing crashes and determining causes. For example, if a specific location was being studied all of the diagrams for crashes occurring there could be overlaid, making it very easy to see similarities in location and contributing factors between crashes.

Electronic reporting tools also simplify the data reporting process. Alabama has developed a system called *eCrash* that officers use to record crash information (CAPS 2009b). The system allows officers to scan a driver license and/or license plate to automatically populate much of the information in the report. A lot of the information on crash reports is also redundant, being recorded two or three times on the same form. With an electronic report the information could be entered once and then populate to all relevant entry locations, thus minimizing the potential for human error in data entry, while also saving time in the data recording process. The *eCrash* system also uses data validation checks for all entries to make sure that they meet standards and are appropriately placed. This helps to catch errors before the report leaves the officers hands (CAPS 2009b). DeLucia and Scopatz (2010) discussed the potential to have even more advanced quality checks that will compare diagrams with written descriptions and entered data to ensure that the narrative of the crash is complete. In the event that data was incomplete the report could be returned to the officer within a day for clarification.

Utilizing and electronic reporting tool would also allow roadway data to be auto-populated. Once the officer indicates the crash location the crash report could be populated using a GIS system that would input number of lanes, speed limit, geometry information, functional classification, shoulder presence, traffic control devices, and other variables. This would make the officer's job easier and provide information that historically has not been available for safety analyst's use.

Another benefit to electronic crash data collection is that data could be submitted electronically from the officer's vehicle, and become available for viewing by safety analysts almost immediately. In the event that new crash hotspots were developing due to increased traffic or construction the data would be available immediately for analysis and countermeasure design. Electronically submitted data could also be sorted automatically into different databases, eliminating another step that has historically been done manually and provides additional opportunity for error.

The primary opposition to electronic data collection is the additional equipment it would require, and the fact that officers would need a connection to a data network. Connections to a data network will continue to become less of an issue over time as satellite and wireless technology has advanced enough data connections are available in most areas of the country. It is possible that in very rural areas officers would have to fill out the form without internet access, but these situations would be rare and the form could still be completed once returning to the station. Regarding equipment, much of the capability needed is available through officer's on-board computers, and relatively low-cost programs could be used for other features. A barcode reader for driver licenses would be needed. Many state agencies already use these, and for those that don't low-cost USB scanners and wireless devices are available at reasonable cost. As far as GPS systems are concerned it has been shown in research that very inexpensive handheld tools or computer programs are sufficient for determining crash location (Graettinger et al. 2001). The system could also take advantage of GIS software on officers' laptops. This would allow the officer to simply click a point on the map and the coordinates will be output automatically. Overall these concerns are becoming very minor issues due to advances in technology.

The possibility of moving to real-time digital crash data collection is becoming more and more real with advances in technology and with documentation of lessons learned by other states. Many systems have already been implemented and proven effective (DeLucia and Scopatz 2010). Digital crash data collection would greatly improve data accuracy, provide information that was not previously available, and make data more accessible for analysis.

4.6 Chapter Summary

Accuracy, availability, coverage, and usability are four important issues to consider when using and preparing data for crash analysis. Accuracy affects validity of results, availability determines if the data can be used, coverage affects how comprehensive the analysis will be, and usability affects how easy it is to incorporate the data into an analysis processes. Data were acquired and prepared for use in this research with each of these considerations in mind. Crash data, AADT, speed limit, functional classification, through lanes, urban code, geometry, and skid index data were all prepared for use in safety analysis. Methods for using data in safety analysis include input as variables, using it to segment roadways, and using it to interpret statistical

results. Using roadway segments based on existing data holds the most promise for producing accurate results that can be easily interpreted. Segment lengths need to be considered when performing analysis; however, segments based on geometry or skid index are not considered practical at this time. Electronic collection of crash data has become a very realistic possibility in recent years due to advances in technology. Electronic data collection improves data accuracy and provides information that is not currently available.

5 STATISTICAL MODEL

A hierarchical Bayes model was developed to analyze crashes on all state roads in Utah. This chapter discusses the theoretical basis for the hierarchical Bayes model, a summary of the components used to develop the model, the resulting output of the model, and a chapter summary.

5.1 Theoretical Basis – Hierarchical Bayesian Zero-Inflated Poisson Regression

When using the Bayesian framework it is important to choose a distribution that fits the data as well as prior distributions to summarize what is believed about the parameters used to model the distribution of the data. Since the response variable is number of crashes, it would be intuitive to model the data using the Poisson distribution, a distribution commonly used to model count data. When using the Poisson distribution, it is necessary to assume that the mean and variance of the data are equal. Since a large number of road segments being analyzed in this study have zero crashes, this assumption is not met. This high number of zero crash segments causes the variance to exceed the mean resulting in overdispersion of the data.

Because of the nature of the data for this analysis and the assumption that the mean and variance are not equal, a different distribution that preserves the ability to model count data while also allowing for excess segments with zero crashes is recommended. This distribution is called the zero-inflated Poisson (ZIP). The ZIP is a mixture of zero with probability p and the original Poisson distribution with probability 1-p as illustrated in Equation 5-1, where λ_{ijk} and p_{ijk} are the parameters of the ZIP distribution (Lambert 1992). In this equation the number of crashes for the i^{th} segment on the j^{th} route with the k^{th} functional classification (denoted Y_{ijk}) is assumed to follow a ZIP distribution.

$$Y_{ijk} \sim ZIP(\lambda_{ijk}, p_{ijk}) \quad (5\text{-}1)$$

where:
- Y = number of crashes,
- λ_{ijk} = the mean and variance of the crash count for segment *i*, route *j*, and functional class *k*,
- p_{ijk} = the probability that the crash count is zero,
- i = segment,
- j = route, and
- k = functional classification.

The probability density function for Y_{ijk} is given in Equations 5-2a through 5-2c.

$$P(Y_{ijk}) = \begin{cases} p_{ijk} + (1 - p_{ijk})e^{-\lambda_{ijk}} & if\ Y_{ijk} = 0 \\ (1 - p_{ijk})\frac{e^{-\lambda_{ijk}}\lambda_i^{Y_{ijk}}}{Y_{ijk}!} & otherwise, \end{cases} \quad (5\text{-}2a)$$

where:

$$\log(\lambda_{ijk}) = \beta_{0jk} + VMT_{ijk}\beta_{1jk} + SpeedLim_{ijk}\beta_{2jk} \quad (5\text{-}2b)$$

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \gamma_{0jk} + VMT_{ijk}\gamma_{1jk} + SpeedLim_{ijk}\gamma_{2jk}. \quad (5\text{-}2c)$$

It is important to note that this model is not estimating λ_{ijk} directly, but rather λ_{ijk} is modeled using VMT and speed limit (*SpeedLim*) in Equation 5-2b. In order to assess the effects of these two variables on λ_{ijk}, the variables β_{0jk}, β_{1jk}, and β_{2jk} are introduced. Similarly, to model p_{ijk} in Equation 5-2c, VMT and speed limit (*SpeedLim*) are used and the variables γ_{0jk}, γ_{1jk}, and γ_{2jk} are introduced to measure the corresponding effects. Notice that in both Equation 5-2b and Equation 5-2c models are created for functions of λ_{ijk} and p_{ijk}. These are called link functions and are quantities that are expected to be more linear than λ_{ijk} and p_{ijk} alone (Lambert 1992).

Non-informative multivariate normal (MVN) prior distributions are utilized in the model as outlined in Equations 5-3 through 5-6. In these equations the matrix **I** represents an identity

matrix of appropriate dimension. The identity matrix is multiplied by 100 in order to ensure that the priors are non-informative, indicating that the variance of each parameter is 100.

$$\vec{\beta}_{jk} = \begin{pmatrix} \beta_{0jk} \\ \beta_{1jk} \\ \beta_{2jk} \end{pmatrix} \sim MVN(\vec{\mu}_k, 100\mathbf{I}) \tag{5-3}$$

$$\vec{\gamma}_{jk} = \begin{pmatrix} \gamma_{0jk} \\ \gamma_{1jk} \\ \gamma_{2jk} \end{pmatrix} \sim MVN(\vec{\Gamma}_k, 100\mathbf{I}) \tag{5-4}$$

$$\vec{\mu}_k = \begin{pmatrix} \mu_{0k} \\ \mu_{1k} \\ \mu_{2k} \end{pmatrix} \sim MVN(\vec{0}, 100\mathbf{I}) \tag{5-5}$$

$$\vec{\Gamma}_k = \begin{pmatrix} \Gamma_{0k} \\ \Gamma_{1k} \\ \Gamma_{2k} \end{pmatrix} \sim MVN(\vec{0}, 100\mathbf{I}) \tag{5-6}$$

The parameters $\vec{\beta}_{jk}$ and $\vec{\gamma}_{jk}$ have prior distributions depending on other parameters $\vec{\mu}_k$ and $\vec{\Gamma}_k$, called hyperparameters. These can be interpreted as parameters in the linear model for the k^{th} functional classification, or average parameters for the routes in the k^{th} functional classification. For example, the average effect of VMT on $\log(\lambda_{ijk})$ is given by β_{1jk}, which is specific to the j^{th} route of the k^{th} functional classification. However, Γ_{1k} gives the average effect of VMT on the entire k^{th} functional classification.

Hierarchical Bayesian methods were utilized to obtain posterior distributions for each parameter in the model and for every combination of route and functional classification. In the statewide data, there were six parameters in the linear models, six hyperparameters, and 304 routes nested within seven functional classifications, yielding a total of 1,866 parameters. The joint posterior distribution of the parameters is proportional to the product of the ZIP distribution for each crash count multiplied by each of the priors. Samples from each conditional posterior are obtained using Markov Chain Monte Carlo (MCMC) and Gibbs sampling methods (Qin et al. 2005). This results in posterior distributions of $\vec{\beta}_{jk}$ and $\vec{\gamma}_{jk}$ for each route and posterior

distributions of $\vec{\mu}_k$ and $\vec{\Gamma}_k$ for each functional classification. This process is called hierarchical Bayesian ZIP regression.

5.2 Model Development

The model was developed using the R programming language because of its versatility and abundance of statistical functions and packages. R is also available as a free download and runs on a variety of computer platforms (RPSC 2012). Hierarchical Bayesian modeling using MCMC methods, especially with the number of parameters used in this analysis, requires heavy computation. Running the desired number of iterations could take hours or even days depending on the amount of data being analyzed and the capabilities of the computer hardware running the computations.

As part of the computation, a candidate generating distribution is used from which MCMC draws are determined to be probable and accepted as samples from the posterior distribution (Gelfand and Smith 1990). Determining the variance of the candidate generating distribution can be challenging. The process of trying a candidate generating distribution variance, analyzing the results, and changing the variance accordingly is called tuning. Though most tuning in the model is done automatically, it can take up to a full day. Further, the automatic tuning is not a guarantee that the choice of candidate variance is good. Before using the results of an MCMC run, the trace plots output by the R function should be analyzed to ensure that they are acceptable. Detailed instructions on how to use the model can be found in Model User's Guide.

5.3 Model Output

Using the posterior distributions obtained for all of the parameters described above, predictive distributions are constructed for each segment. Posterior predictive distributions give a distribution of the number of crashes that would be expected on a segment given its VMT and other variables (speed limit in the analysis presented in Chapter 6). One can then determine where the actual number of crashes falls in the posterior predictive distribution by observing the area to the left of the actual number of crashes in the posterior predictive distribution, or the percentile of the actual number of crashes (between zero and one). A high percentile (near one)

would indicate that the actual number of crashes is larger than predicted on that segment, while a percentile near zero would indicate that the segment had less crashes than predicted.

An example posterior predictive distribution produced by the model is shown in Figure 5-1. The bars represent the distribution of the number of crashes that would be expected on this segment based on analysis of all segments in the same functional classification and route, having the same VMT and speed limit. The solid vertical line represents the actual number of crashes for this segment. The proportion of the area of the distribution to the left of the solid vertical line is the percentile.

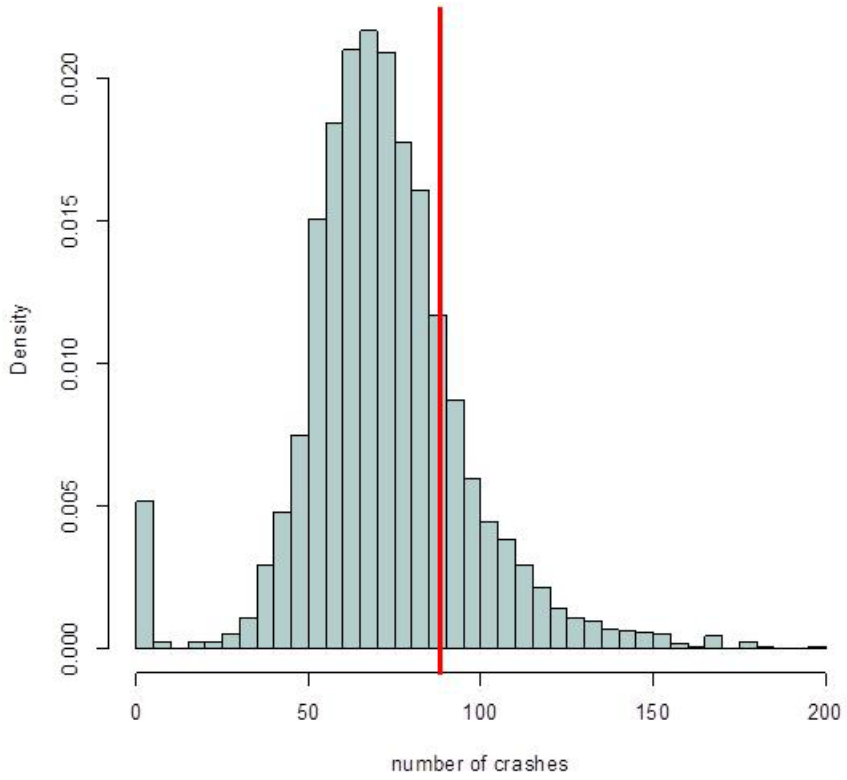


Figure 5-1. For a specific segment the percentile is the area of the predictive distribution less than or equal to the actual number of crashes (solid vertical line).

In some cases, the number of crashes predicted is low but the actual number of crashes is only slightly larger (for example: if the mean of the posterior predictive distribution is one and the actual number of crashes is two). The percentile for this segment would likely be very high but the difference between the predicted and actual values is very low. If only the percentile

were considered when identifying a hotspot this segment would be identified since the number of crashes is statistically significant, but it may not necessarily be practically significant. Thus the mean of the posterior predictive distribution is included in the model output as well. The mean of the posterior predictive distribution can then be compared to the actual crash value and the difference can also be analyzed. The combination of the percentile and the difference between the predicted mean and actual number of crashes will indicate how dangerous a segment may be expected to be. This process will be illustrated in the examples presented in Chapter 6.

5.4 Chapter Summary

To analyze crashes on Utah roadways a hierarchical Bayesian statistical model using ZIP regression was developed using the R programming language. The ZIP is necessary because there are a high number of segments in the data with zero crashes causing the data to be overdispersed.

Posterior predictive distributions for each roadway segment are developed using MCMC and Gibbs sampling methods. By comparing the posterior predictive distribution with the actual number of crashes for a given segment it can be determined if more crashes have occurred on that segment than would normally be expected.

6 FRAMEWORK FOR GIS-BASED CRASH DATA ANALYSIS

The chapters leading up to this point have provided a literature review and outlined the background on traffic safety analysis and GIS applications. In addition, the theoretical basis of a statistical model has been presented to analyze traffic safety and establish the basis for hotspot analyses. To be able to best utilize the results of the statistical model and create a framework that will help engineers and analysts to identify hotspot locations and perform a safety audit on these locations, the results of the safety analysis and GIS application must be combined. This will allow the user to most effectively address the safety concerns in the state.

This chapter establishes the framework for GIS-based crash data analysis. The four steps of the framework will be presented followed by an example in which the framework is applied using ArcMap to a real analysis situation.

6.1 Analysis Framework

The basic concept of crash data analysis in GIS is to use the data to identify locations that are candidates for safety improvement. It is important that the framework for analysis focuses on this final output, and is adaptable to any system and any method of analysis. This allows the analyst to select the platforms, procedures, and outputs while still following the framework guidelines. Figure 6-1 shows the analysis framework graphically. After a brief discussion of the preliminary steps necessary for the framework, each major heading and steps associated with it will be discussed in this section.

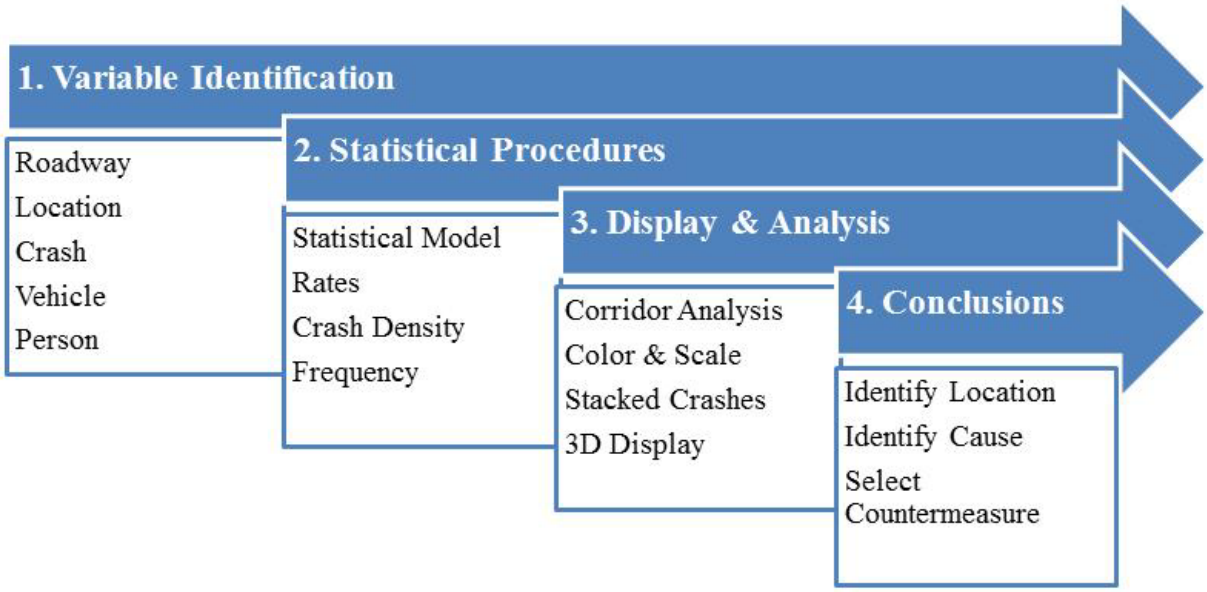


Figure 6-1. Framework for GIS-based crash data analysis.

6.1.1 Preliminary Steps

Before applying the analysis framework it is important to identify the purpose of the analysis. In other words, the analyst must determine the question that is being answered. This step is not a direct part of the GIS analysis process, but serves as a precursor to analysis. Examples of analysis questions include the following:

1. Which interchanges have the greatest crash risk?
2. Which rural highway sections are most susceptible to single-vehicle crashes?
3. Do enforcement activities in different jurisdictions impact the number of crashes caused by traffic violations?
4. Does AADT have a correlation with crash occurrences on this route?
5. Which roadway design features can be changed to reduce crashes on this route?
6. Which half mile segments in the state have the greatest crash risk?

These and many more questions may be the focus of an analysis. In addition to the examples listed, each should have the question “Why” at the end. The why question is how the analysis results become a feasible countermeasure that can improve safety.

6.1.2 Step One: Variable Identification

The first step is to identify the variables that will be used in analysis. This will depend on the question being answered or purpose of the analysis. Some variables that should be considered include the crash location, roadway attributes, and information from the crash, person, and vehicle databases. Location data include both geographic and demographic variables such as county, city, or land use along with route and milepoint location. Roadway characteristics include AADT, speed limit, functional classification, number of lanes, and any others that are needed in the analysis. Identifying locational variables is identifying the environment in which the crash occurred. Information from the crash, person, and vehicle databases is collected by the officer and provides variables related to how the crash occurred, what non-locational factors were involved, potential human-related causes, and what the outcomes of the crash were. Examples include information such as alcohol involvement, traffic violations, number of passengers, injuries or fatalities, vehicle malfunctions, time of day, and sequence of events. This step will also involve some data acquisition and preparation. For more information about using variables in analysis and data considerations refer to Chapter 4 of this report.

6.1.3 Step Two: Statistical Procedures

Statistical procedures provide information that can be used to analyze the data and make decisions regarding hotspots, crash causes, and countermeasures. In general the two classes of statistics used include descriptive statistics and inferential statistics.

Descriptive statistics provide information about the data. They show patterns, distributions, relative values, and help summarize data. Descriptive statistics are useful for understanding and interpreting data, but are not robust enough to draw generalized conclusions or to make predictions. Inferential statistics use more advanced mathematical procedures and sampling methods, and thus are able to draw generalized conclusions about the population and aid in predictive analysis (Laerd Statistics 2012). More details on the type of statistics used are provided in Section 2.3 and Chapter 1.

The type of statistics used will depend on the analysis being done, the question being answered, and the software tools available. The purpose of the statistical procedures performed in this step of analysis is to create outputs that can be used to identify crash hotspots. Outputs

should be such that they can be interpreted in a GIS environment. This means that the outputs need some sort of discrete value that can be represented graphically with a spatial connection.

6.1.4 Step Three: Display & Analyze

Step three consists of taking results of the statistical procedures and displaying them in GIS for analysis. Section 3.3 of this report shows several examples of how this is done. As discussed in Step Two, the statistical procedures need to include output that can be displayed in a meaningful way. This means that each data record (crash point or roadway segment) needs to have some value associated with it. In the examples shown previously in Figure 3-7 and Figure 3-8 each segment has a rate (descriptive statistic) associated with it that is displayed by a color scale. The example shown previously in Figure 3-12 used a similar display method, but in 3-D and shows crash risk (inferential statistic). Figure 3-13 and Figure 3-14 showed density plots that can be used to find concentrations (descriptive statistic). Another example not shown gives an output that shows likelihood (inferential statistic) for each crash occurrence, which could be displayed by colored points as was shown previously in Figure 3-6.

The purpose of using these display methods is to allow the analyst to better understand and analyze the data and the outputs of the statistical model. Displaying results graphically improves the quality and simplicity of analysis. Once the results have been displayed and analyzed, conclusions can be developed regarding the answer to the question posed and potential solutions.

6.1.5 Step Four: Form Conclusions

The final and most important step of crash data analysis is to form conclusions. In this step the question posed at the beginning of the analysis should be answered and a decision for what actions to take regarding the issue should be determined. Conclusions should be checked by the analyst and others to assure that they are reasonable, can bring tangible improvement to the system, and are backed by statistical results.

Once conclusions are made the process of selecting future steps to be taken depends entirely on the results. If a specific countermeasure and location were determined in the analysis then the solution is ready for implementation, subject to time and budget constraints. In other analyses the results may lead to further studies for more detailed results. This is often where

iteration occurs in the process as it may be determined that different variables need to be used or a modified statistical approach taken. Theoretical examples are provided here of the sample questions posed in Section 6.1 with possible answers to the questions and steps to be taken. It should be noted that the steps outlined are examples only and are not provided as solutions to any specific problems.

1. Which interchanges have the greatest crash risk?
    a. Answer: Interchanges at Exit 453 and 527 have the greatest risk
    b. Steps: Study those intersections to determine how to make them safer
2. Which rural highway sections are most susceptible to single-vehicle crashes?
    a. Answer: Rural highway sections 10 miles or more from a city or roadway junction are most susceptible to single-vehicle crashes
    b. Steps: Install drowsy driver warning signs and/or rumble strips at these sites
3. Do enforcement activities in different jurisdictions impact the number of crashes due to traffic violations?
    a. Answer: Jurisdictions with strategic enforcement goals have fewer violation-related crashes
    b. Steps: Require traffic law compliance enforcement plans from all jurisdictions
4. Does AADT have a correlation with crash occurrences on this route?
    a. Answer: Crash occurrences are correlated more with speed limit than AADT
    b. Steps: Determine candidate sites for speed limit changes
5. What roadway design features can be changed to reduce crashes on this route?
    a. Answer: Curves that occur within 0.1 miles of each other are correlated with crash occurrences on this route
    b. Steps: Examine signage, speed limit, and realignment options for curves that meet this criteria
6. Which half mile segments in the state have the greatest crash risk?
    a. Answer: Milepost 37.5 to 38.0 and milepost 12.0 to 12.5 on these routes
    b. Steps: Study these sites to select countermeasures for implementation

6.1.6 Order and Iteration of the Framework

This framework is meant to adaptable and flexible. It is designed as a four-step process, but the four steps do not have to be sequential. In some cases the steps will be performed at the same time or in differing order. Display methods will often be used to select variables of interest, rates may be calculated to help identify locations to study, and conclusions will often lead to a modification of one or more of the steps and then starting over. The graphic used to represent this process in Figure 6-1 is designed to show this. Every step is necessary, but they overlap in different ways and can be done in whatever order is needed. The purpose always is to use this process to identify safety improvements that can be made to the transportation system.

6.2 Framework Example Using ArcMap

This section will provide a step-by-step example of how the framework can be applied in a real GIS-based safety analysis process. This example will incorporate information described in the literature review (Chapter 3) and will utilize data that was prepared in accordance with the procedures described in Chapter 1. It will also utilize outputs from the statistical model. Because crash data analysis should always be different based on the situation, this example should not be considered a prescribed approach, it is merely an illustration to bring together concepts from this report in a tangible way. The section headings will generally match the steps recommended in the framework in Section 6.1 and shown in Figure 6-1.

6.2.1 Custom GIS Models

Three models were developed in ArcMap to perform different functions necessary in analysis. These models were developed using ArcMap's ModelBuilder and were assigned parameters so that they could be run like any tool in ArcMap. Figure 6-2 shows these models as they appear in a toolbox within ArcMap. Each of the seven "Overlay" models shown in the figure are essentially the same model, they each just perform the process for a different number of input files. These tools will be referred to and explained in this analysis example. Each is explained in detail in Appendix B.

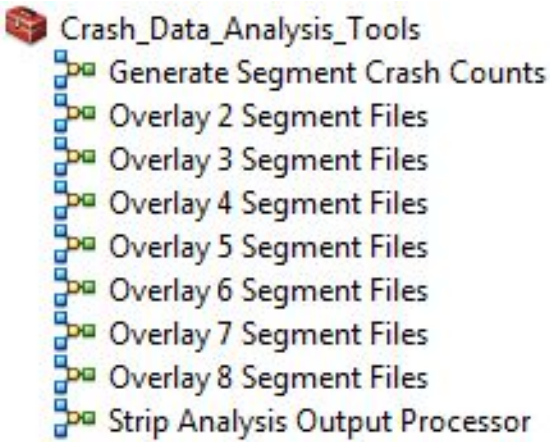


Figure 6-2. Models created for analysis procedures.

6.2.2 Preliminary Steps

Before the analysis begins, the preliminary step is to identify the purpose of the analysis, and create a question that will be answered. For this example, the purpose will be to use the newly developed statistical model results to analyze all state highway segments and identify a few potentially high-risk locations (see Chapter 5). Usually a more specific purpose and question would be developed, but for an example this is sufficient. The question to be answered is, "What roadway segments across the state pose high crash risk, and why?"

6.2.3 Step One: Variable Identification

The first step in the overall framework is to identify what variables will be used in the analysis. Because the model looks at all state highways it is good to incorporate as much data as possible in order to help distinguish segments. In a more specific analysis one or two data variables may be sufficient. This analysis will use all data that is prepared and considered accurate. This includes functional classification, AADT, percent trucks, speed limit, through lanes, and urban code. These datasets will be used to create homogeneous segments for the entire state. AADT (or more specifically, VMT to account for segment length) and speed limit will then be used as variables in the model. No specific locations will be identified at this stage because this analysis will examine the entire state. Crash data from 2006-2010 will be used because it is the newest dataset available. For this analysis all crashes will be included. In future analyses it may be necessary to include only a subset of crashes such as only crashes with severity 3 through 5.

6.2.4 Step Two: Statistical Procedures

The statistical model developed for this research will be used to determine crash risk on the study segments (see Chapter 5). The model utilizes a hierarchical Bayesian model that allows input of variables. Functional classification will be used as the hierarchy for the segments and AADT and speed limit will be used as input variables. The following sections explain steps to prepare data for the model, run the model, and prepare results for use in ArcMap.

6.2.4.1 Dataset Preparation (Pre-Processing)

A dataset will be prepared that can be input into the statistical model. Linear referencing and dynamic segmentation will be used to overlay each of the input roadway variables to create a file of homogeneous segments (see Section 3.2 for a review of linear referencing and dynamic segmentation). The input dataset must include segment lengths, attribute values, crash counts, and the necessary identifier fields for mapping.

The custom tool “Overlay 5 Segment Files” will be used to overlay the roadway datasets. This model uses the linear referencing tool “Overlay Route Events” (which performs dynamic segmentation) to overlay five roadway segment datasets and create one dataset with homogeneous segments for the entire state. Figure 6-3 shows the inputs for this model. The output is a segment file that has been split to create homogeneous segments and includes all data values from each dataset. The output table from running this tool is manually cleaned up to remove unnecessary fields and prepare it for the statistical model.

Now that a segment file has been produced the crash counts must be generated for each segment. For this the “Generate Segment Crash Counts” model will be used. This model takes input segments and crashes and outputs the segment file with a new field that shows the number of crashes on each segment. Figure 6-4 shows the input box for this window. After performing this step the dataset is ready for input into the statistical model.

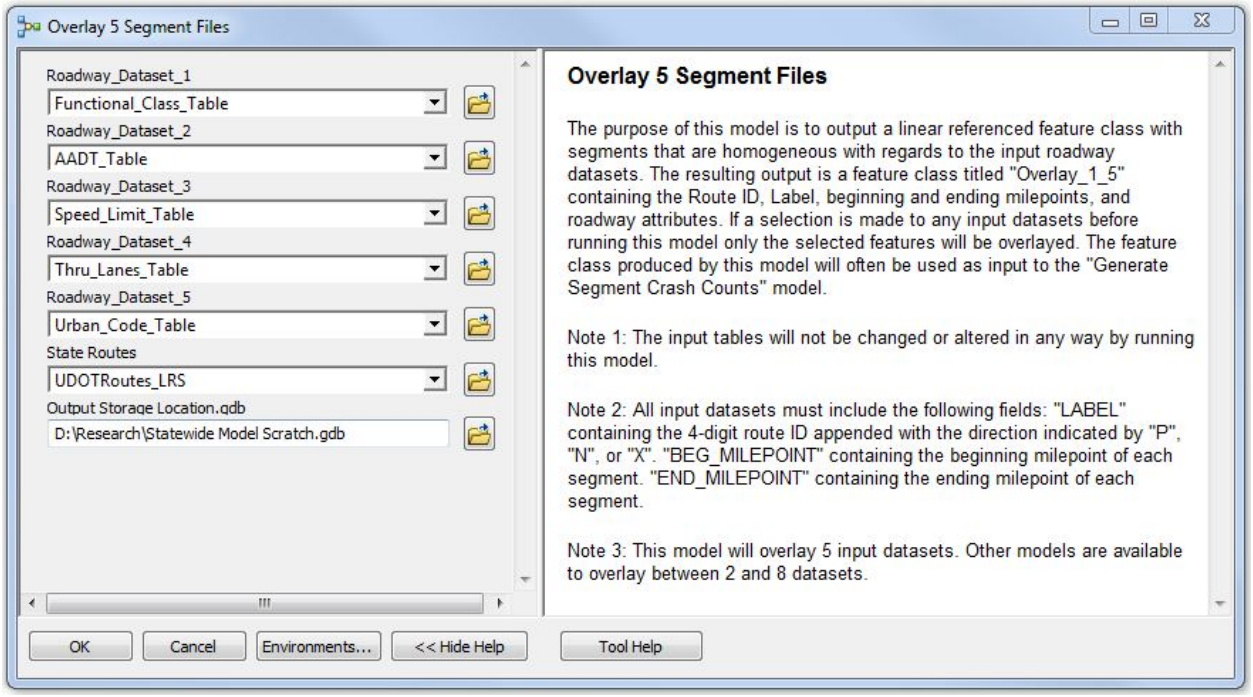


Figure 6-3. Input box for overlay model.

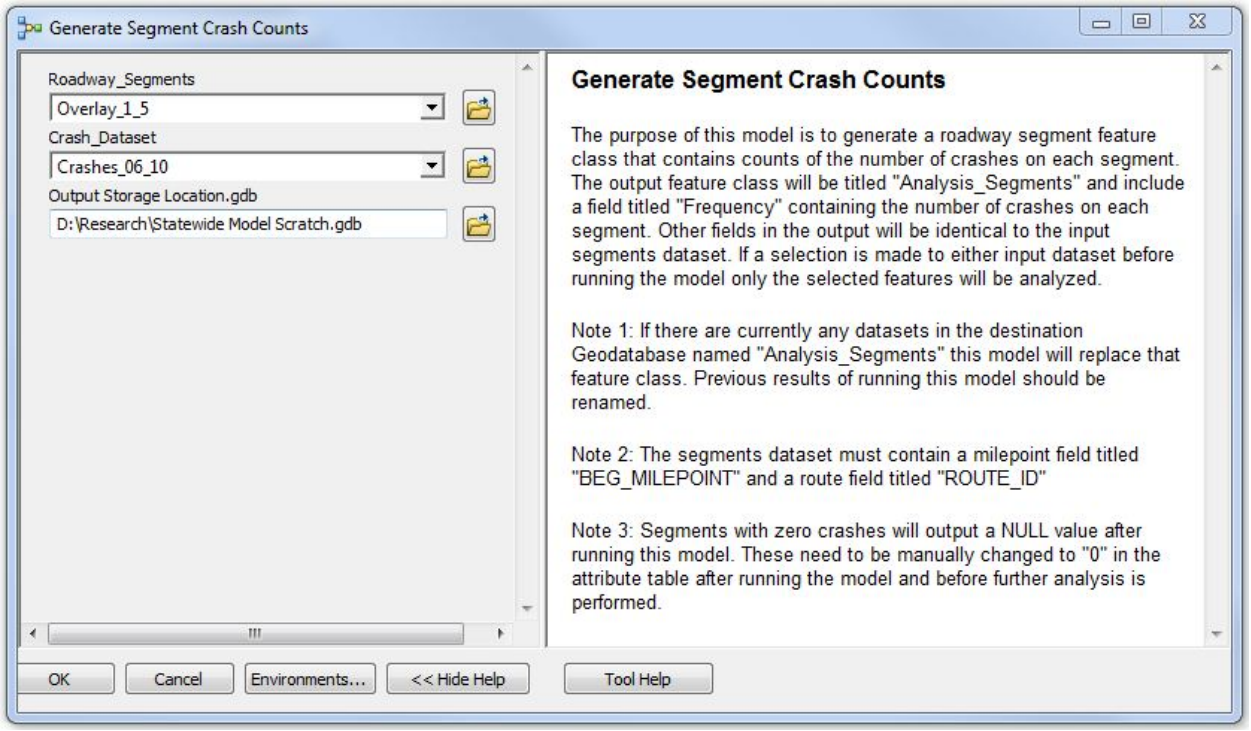


Figure 6-4. Input box for crash count model.

6.2.4.2 Running the Model (Processing)

The dataset table is output to a CSV file, which is then used as input into the statistical model. After processing, the model outputs the same table with two fields added. The new fields are “Posterior_Mean” and “Percentile.” The posterior mean field can be interpreted as the number of crashes expected to occur on that segment given the input variables. The percentile represents the likelihood of the observed number of crashes actually occurring. More details on the model process are found in Chapter 1 of this report.

6.2.4.3 Calculate Rate and Difference (Post-Processing)

To assist with analysis of results the crash rate will be calculated. For this example the crash rate per MVMT will be calculated using the crash count, length, and AADT. Another statistic that will be computed is the difference between the posterior mean (expected number of crashes) and the actual number of crashes. This will help determine practical significance of a high risk segment. Now that all of the necessary statistics have been calculated the results can be imported to ArcMap for display and analysis.

6.2.5 Step Three: Display and Analysis

After importing results to ArcMap the segments were mapped using the State LRS. The results were displayed using a color scale and the percentile to represent crash risk. Figure 6-5 shows the results for the entire state.

The results of Figure 6-5 show a logical representation of crashes, indicating that the hierarchical Bayes model appears to be effective. Evidence of this is that both red segments (hotspots) and blue segments (coldspots) appear throughout the state and in both urban and rural areas. This suggests that the model is doing a good job of determining expected crashes and identifying high risk segments according to the hierarchy and input variables.

The percentile ranks for the color-coding have been selected to represent how an actual percentile would be calculated. Section 5.3 provides additional detail about these outputs and how the percentile can be interpreted. The ranges were selected so that 5 percent of segments would fall in the highest and lowest groups (red and blue respectively), 15 percent of segments would fall in the next highest and lowest groups (orange and green respectively), and 60 percent of segments would fall in the middle range (light green).

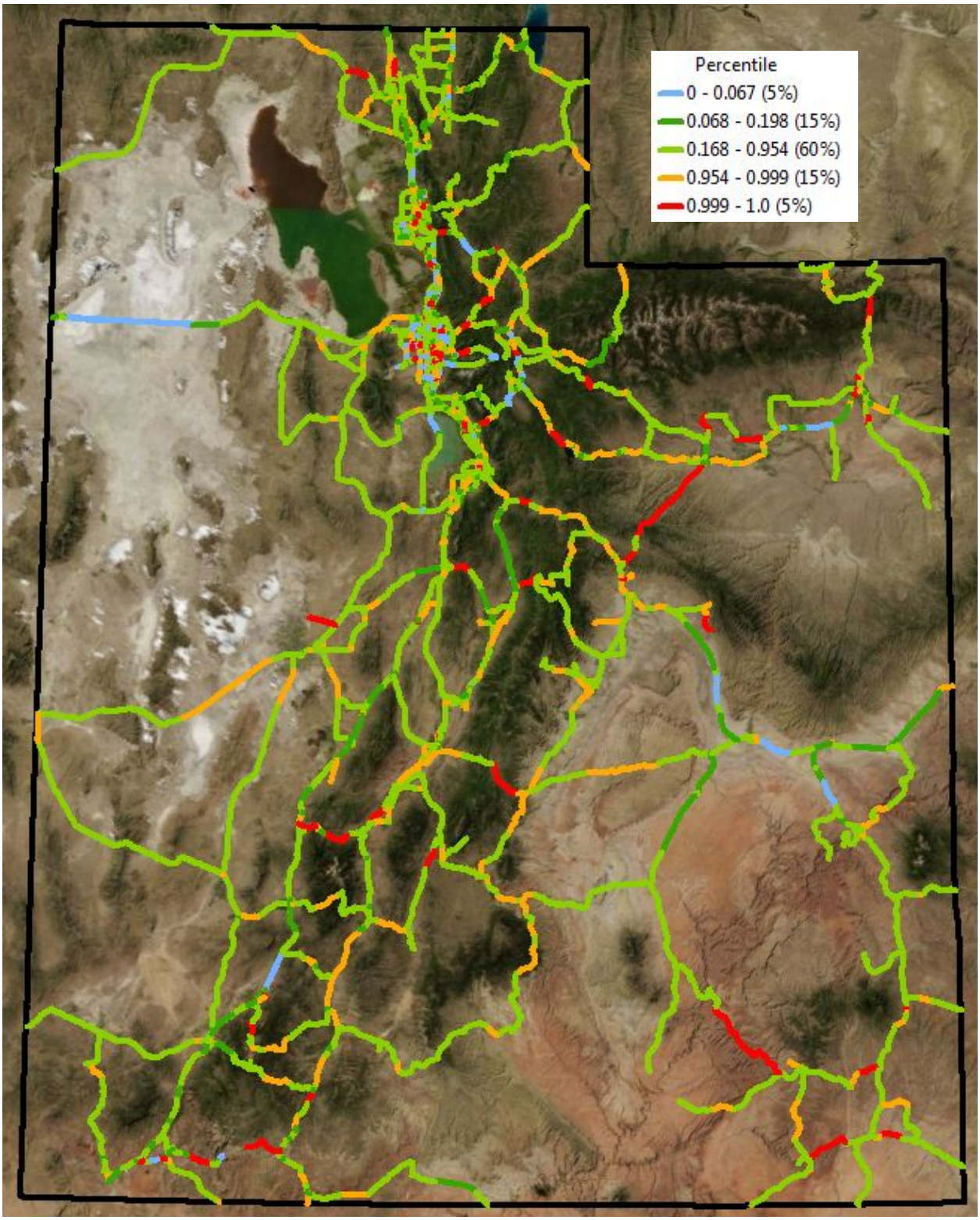


Figure 6-5. Statewide model results displayed by percentile.

Figure 6-6 shows results based on the difference between the posterior mean and the actual number of crashes. This helps to show the practical significance of results by indicating if any crash countermeasures could actually have a significant impact. For example, using only percentile as a segment performance measure a segment may be a high percentile, even though there were only 10 crashes there, because the expected number of crashes was one or two. This segment would be considered statistically significant but not practically significant. A segment that has a large number of expected crashes and a much greater observed number of crashes would be practically significant. Symbolizing segments by the difference helps to represent this practical significance.

A hybrid method of displaying results is shown in Figure 6-7. This map uses the percentile color coding along with the line weight to show crash difference. By including both factors the map shows high risk locations and potential high impact locations graphically, making it easier to identify sites that warrant further study.

The remainder of this example will focus on two different locations to illustrate some micro scale analysis methods. The first location will be US-89 through Logan Canyon from milepoint 470 to milepoint 478. This area received significant publicity in the winter of 2012 due to a few high-profile crashes that occurred at milepoint 473.7. The corridor will be examined to determine if the location could be considered a hotspot, and what variables may be contributing to it. The second location is US-191 between US-6 and US-40, or milepoint 253.1 to 294.1. The results of the analysis indicate that this section contains three separate segments with crashes in the top 5 percent of all crashes, as well as crashes in which the difference between the posterior mean and the actual is relatively high. The area will be examined in detail to determine what may be causing the high crash risk and how it can be mitigated.

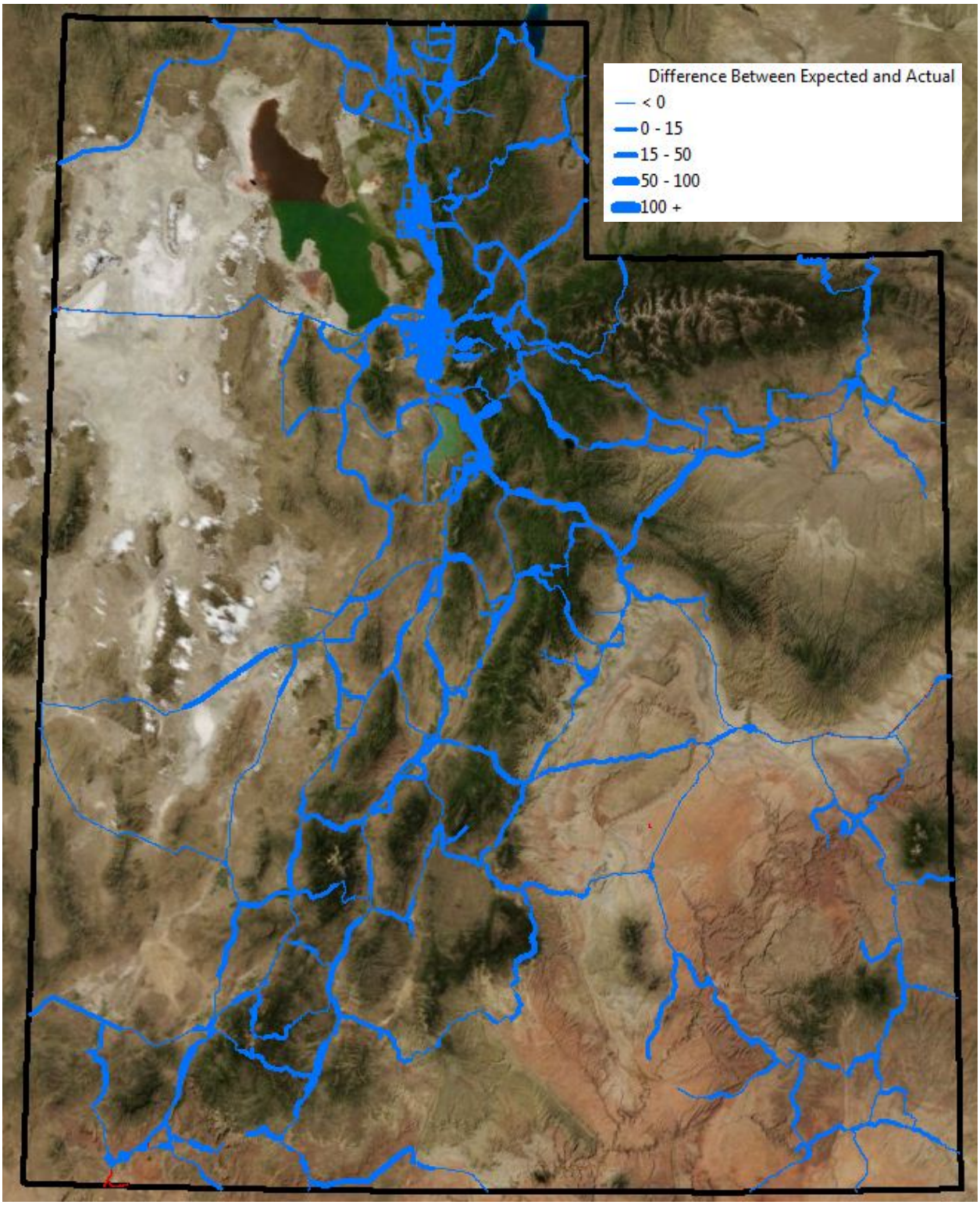


Figure 6-6. Statewide model results displayed by crash difference.

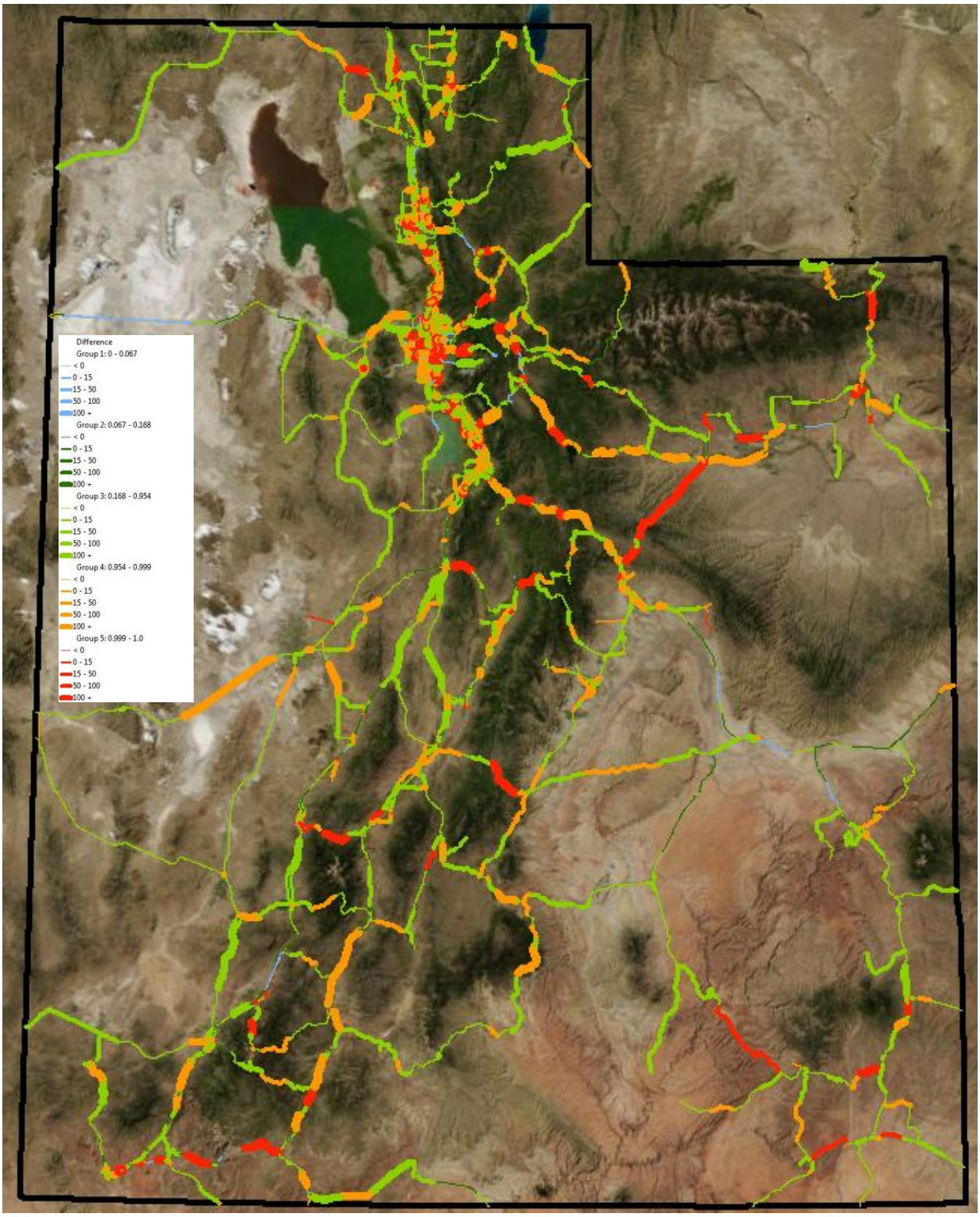


Figure 6-7. Statewide model results displayed by percentile and crash difference.

6.2.5.1 US-89 Milepoint 470 to 478 (Logan Canyon)

US-89 in Logan Canyon between milepoint 470 and milepoint 478 was identified in the winter of 2012 by UDOT staff as a corridor with potential negative safety impacts. Previous highway safety studies have not identified it as a high crash location, but there is circumstantial evidence that it does pose a significant risk. Milepoint 473.7 has been the primary subject of concern due to several crashes at that location where vehicles have ended up sliding off of the road into the river, particularly during storm events. To date there have been no fatalities, but this type of crash poses significant risk for future fatalities. There have been three serious injury crashes, all of which involved motorcycles (Personal Communication with Scott Jones, Jan. 2012).

Based on the statistical performance measures this does not appear to be a high risk area. All of the segments are in the third (0.168 to 0.954 60%) or fourth (0.067 to 0.168 15%) percentile groups, and none of them have significant differences in expected and actual crashes as seen in Figure 6-8.

Because the segments collectively do not pose significant crash risk, the next analysis step will be to examine this section in ¼ mile segments to determine if one specific location poses significant risk. This will provide better micro-scale information than is given in the model results. The strip analysis tool will be used to break the highway into ¼ mile segments, the crashes will then be analyzed for each ¼ mile segment.

A custom tool is used to work with the strip analysis output. The strip analysis tool does not use linear referencing, and the output does not have any of the information necessary to convert the file to an LRS dataset. The “Strip Analysis Output Processor” was developed to make it possible to use strip analysis outputs with the LRS. The strip analysis was run, then the output processor tool was run and the outputs are shown in Figure 6-9.

Results of the strip analysis show that four locations could potentially pose high crash risk, based on the number of crashes in each ¼ mile segment. Figure 6-10 through Figure 6-13 show each of these four locations with crash points and the curve radii labeled. Each site has a curve radius of approximately 100 feet, which is relatively small, and crashes that occur at the curve. This suggests that curvature could be closely related to crashes at each location. With this discovery, it is best to now go through the crash database and determine how many of the crashes include roadway departure and look for other variables that may indicate that the curves were a

factor. For milepoint 473.7, 95 percent of crashes included roadway departure, which supports the initial conclusion that there are concerns with slide offs at this location. Next, a detailed analysis will be performed for the site at milepoint 473.7 to develop conclusions and garner ideas for appropriate countermeasures.

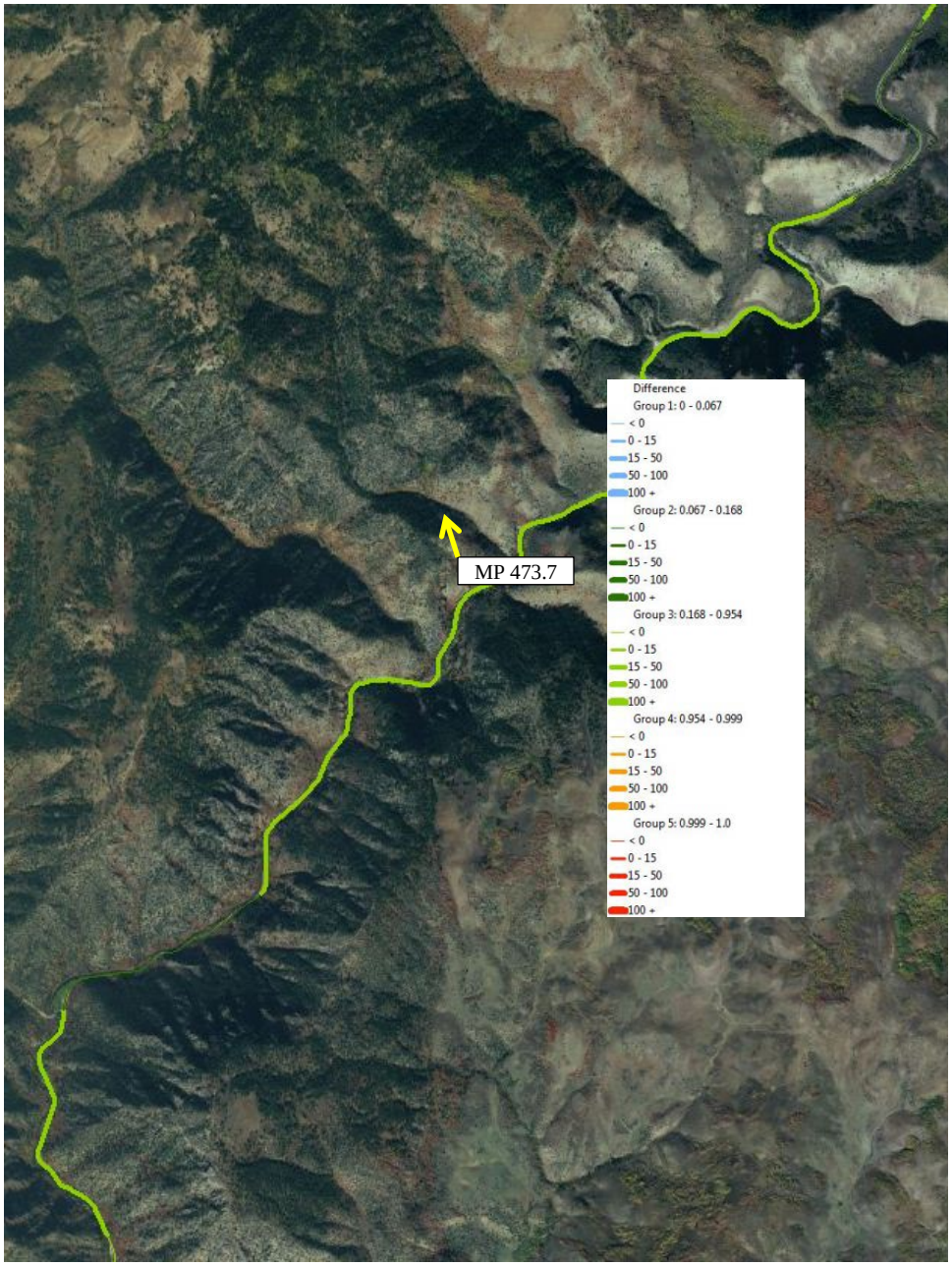


Figure 6-8. US-89 milepoint 470-478 model output.

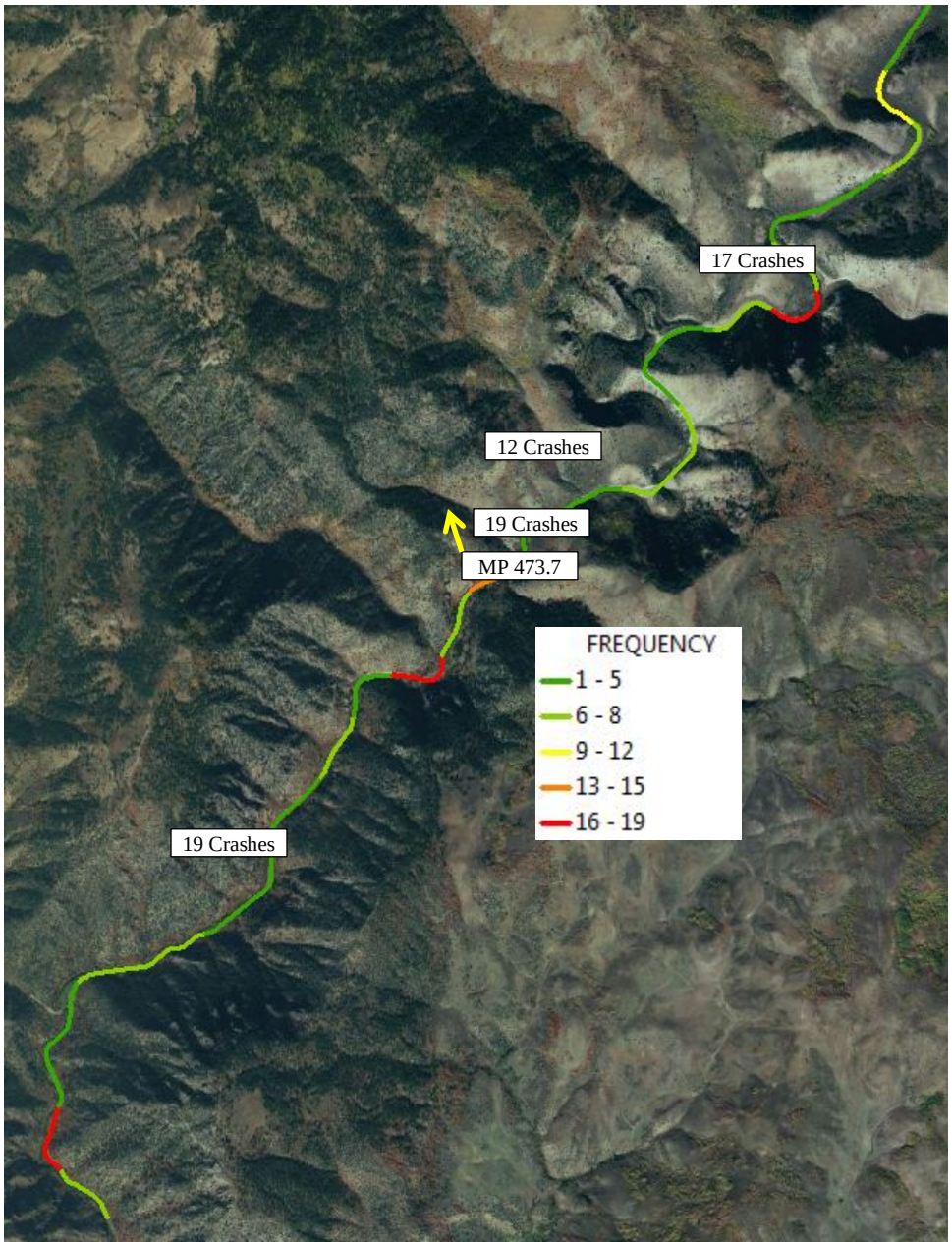


Figure 6-9. US-89 milepoint 470-478 ¼ mile analysis.

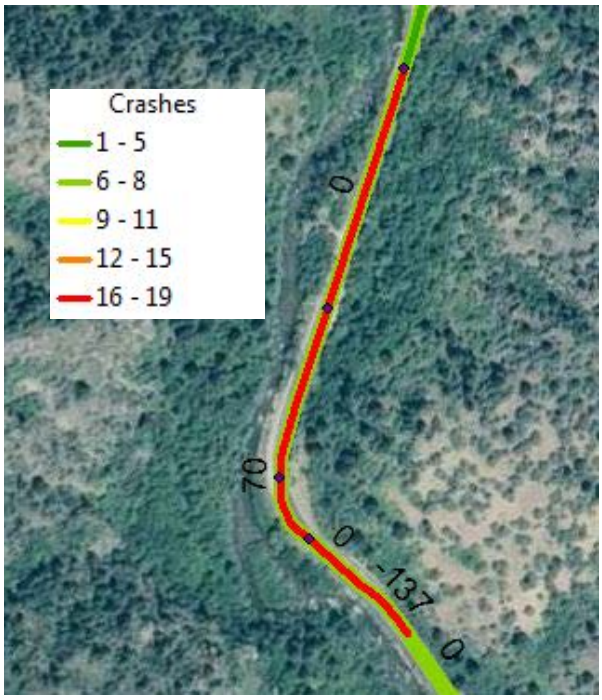


Figure 6-10. Crash points and geometry at US-89 milepoint 471 (19 crashes).

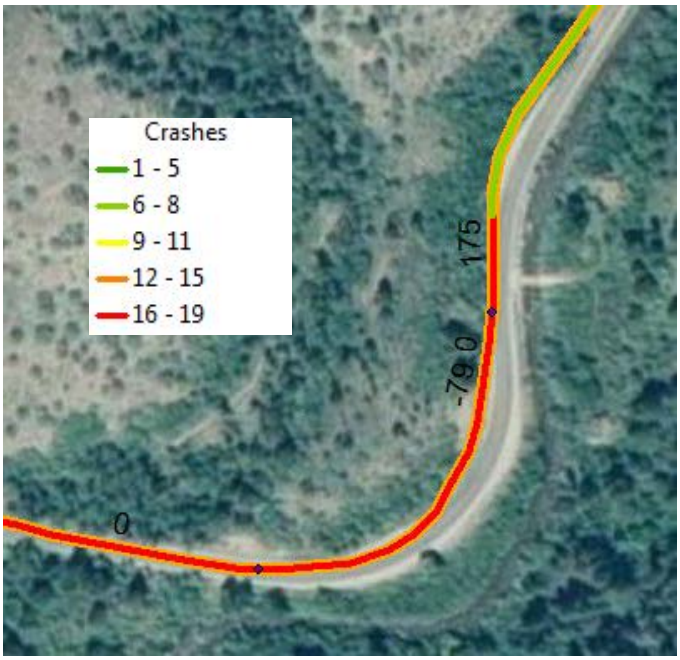


Figure 6-11. Crash points and geometry at US-89 milepoint 473.7 (19 crashes).

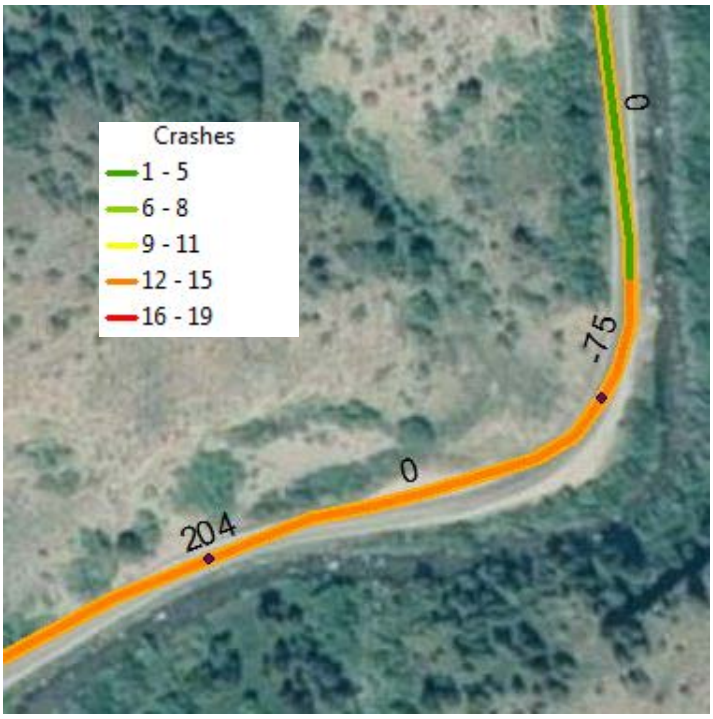


Figure 6-12. Crash points and geometry at US-89 milepoint 474.2 (12 crashes).

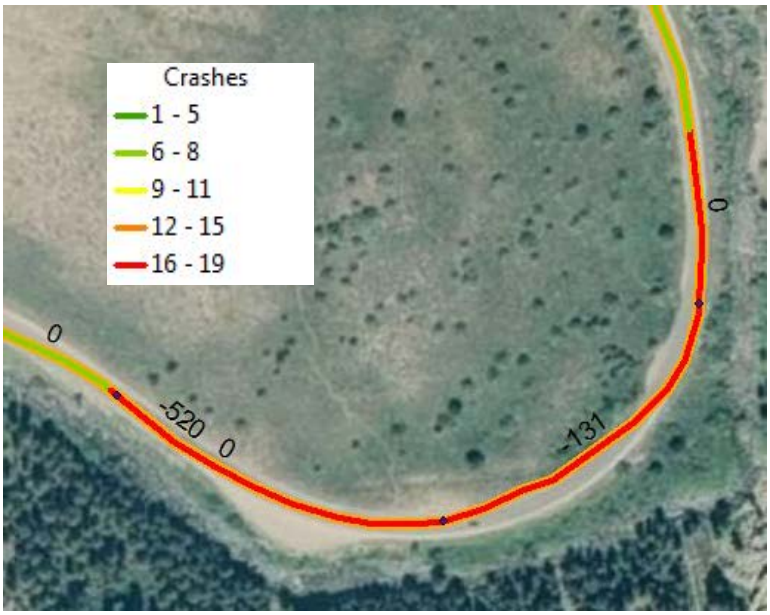


Figure 6-13. Crash points and geometry at US-89 milepoint 476.1 (17 crashes).

A detailed analysis of this location was performed using ArcMap and Google Streetview. The results were prepared in a PowerPoint presentation with comments and presented to the Technical Advisory Committee. Appendix C shows these slides with comments below each. The analysis indicates that poor signing, a lack of chevron curve indicator signs, or a lack of proper skid resistance on the curve could be contributing to crashes at this location. Preferred countermeasures would be to install an approach sign that better indicates the actual road geometry, installation of chevron curve indicators, and possible treatment of the surface to improve skid resistance. A possible speed reduction could also be investigated. Other curves identified in Figure 6-9 could be analyzed in a similar way.

6.2.5.2 US-191 Milepoint 253.1 to 294.1

US-191 from milepoint 253.1 to 294.1 was found to be a high risk crash location from the statistical analysis. Four different segments make up this section of highway, three of which appear in the highest percentile group and have a crash difference of 15-50 crashes. Table 6-1 presents a summary of these segments. Figure 6-14 shows an aerial of the study section with model outputs.

Table 6-1. Study Segments on US-191

Segment Number	Beginning Milepoint	Ending Milepoint	Crashes	Expected Crashes	Percentile
1	253.1	259.1	34	6	1.00
2	259.1	262.8	10	5	0.92
3	262.8	269.7	29	6	1.00
4	269.7	294.1	78	12	1.00

The same basic method will be used to analyze this section of highway as was used in the example in Logan Canyon. The strip analysis tool and strip analysis output processor custom tools were used to analyze the section in ¼ mile segments. The results of this analysis are shown in Figure 6-15. Surprisingly, not a single ¼ mile segment has more than eight crashes out of 151 total crashes for the study section. Two ¼ mile segments have eight crashes, and of the rest, 88 percent, have only one or two crashes.

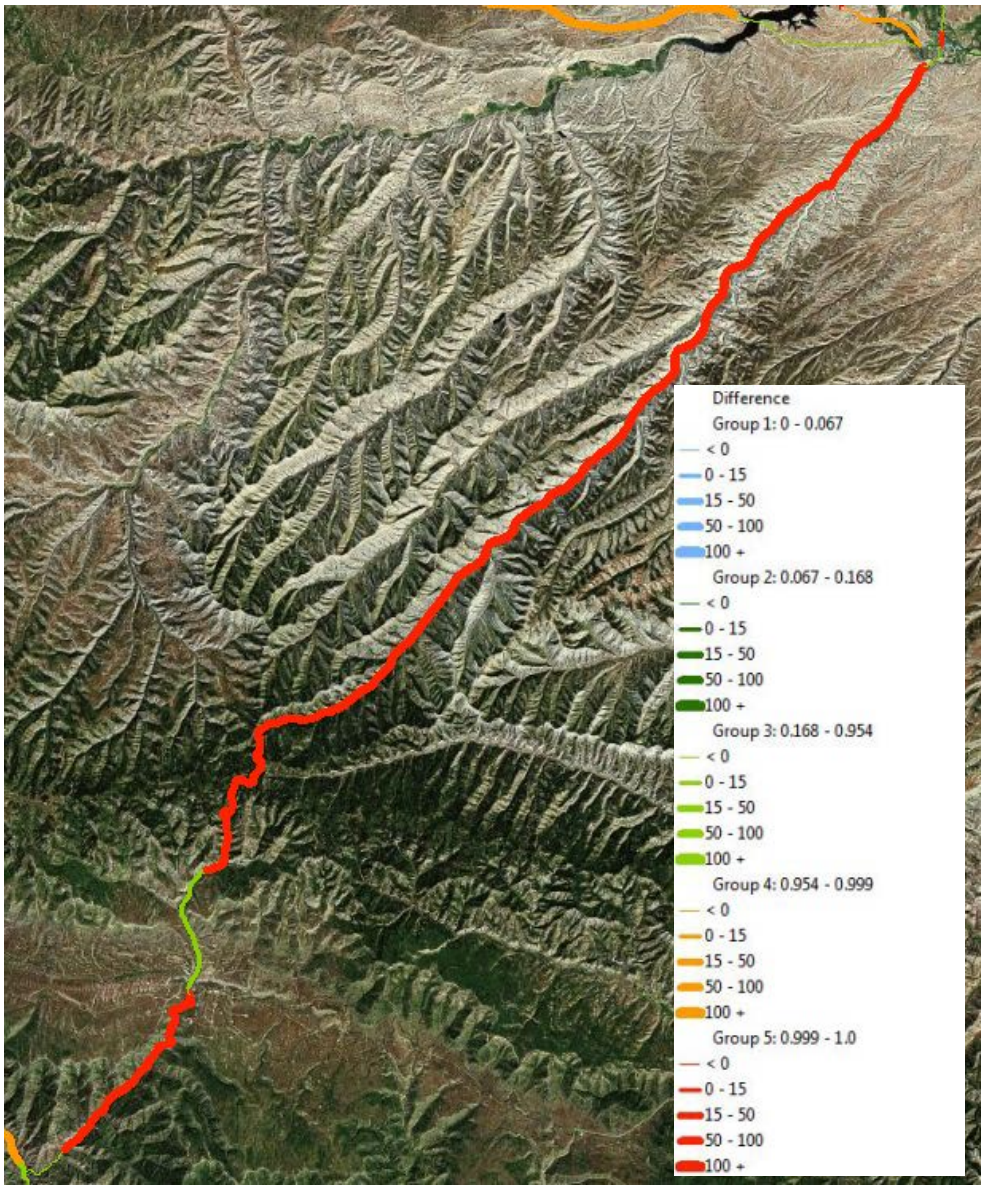


Figure 6-14. US-191 milepoint 253.1-294.1 model output.

Figure 6-15. US-191 milepoint 253.1-294.1 ¼ mile analysis

The fact that three different segments are percentile 1.0 in the model, and no single location appears to be high risk, suggests that the safety problem here is very general. It is possible that because this is a long, isolated mountain pass drivers are more likely to make mistakes. The road does appear to have some irregular curvature, so in general the alignments may be difficult to handle. It could also be drowsy driving causing crashes. A detailed examination of the crashes occurring on this route may reveal additional information. At this point the best recommendation that could be made is to consider placing warning signs

encouraging drivers to use caution, and examining the crash data more carefully to determine if there are patterns in the crashes.

6.2.6 Step Four: Conclusions

Step four has already been demonstrated for each example. For the US-89 case study it was determined that chevron curve indicator signs need to be placed on the curve at milepoint 473.7 and that a more accurate curve approach warning sign needs to be installed. For the US-191 example it was determined that the crash data needs to be examined more carefully, and that signs along the corridor warning drivers to use caution could be effective.

6.3 Chapter Summary

The framework for GIS analysis of crash data is based on the principle of using data to identify crash hotspots and countermeasures that could improve safety at hotspot locations. A four step flexible process is recommended. The steps include variable identification, statistical procedures, display and analysis, and forming conclusions. The framework is designed to work in any analysis and to be iterative within a single analysis. GIS models were developed to assist with crash analysis within the framework.

An example of how the framework can be applied in a real analysis was provided using US-89 milepoint 470 to 478 (Logan Canyon) and US-191 from milepoint 253.1 to 294.1. US-89 through Logan Canyon at milepoint 473.7 does not appear to be a hotspot statistically, but geometry could be impacting safety at that location and the current model does not account for that. A detailed analysis shows that improved signing could improve safety at this location. From the results of the analysis on US-191 from milepoint 253.1 to 294.1 there do not appear to be any specific locations on this segment that are dangerous, instead it appears to be a general problem. Further investigation could indicate specific problems, but at this point warning signs appear to be the best option for improving safety on this segment.

7 CONCLUSION

The purpose of this study was to advance the level of safety research in the state of Utah by developing a statistical model for use in identifying crash hotspots on state roadways using a hierarchical Bayes model. In order for engineers and analysts to successfully interpret the results for each individual segment analyzed by the statistical model, a GIS prototype framework was developed to display the results of the model graphically allowing for a simplified comparative analysis and to more accurately identify crash hotspots. This report presents the procedure used to develop the statistical model and GIS framework.

This chapter briefly summarizes the results of the research and provides recommendations for future research that should be considered to continue to advance safety research in Utah.

7.1 Results

Traditional before and after methods commonly used to analyze automobile crashes are limited in that they do not account for the mean and variance of the data not being equal (as in a Poisson regression) or they do not account for RTM bias. The EB method accounts for both the difference between the mean and variance and RTM, but can be complicated to employ and has limitations of its own. The analysis made possible by the hierarchical Bayes method developed in this study will be an improvement over analyses carried out using before and after or EB methods.

The hierarchical Bayes model developed as part of this research analyzes each roadway segment in the state and determines the severity of the safety needs for each segment. The GIS framework was developed to prepare the data for analysis by creating segments based on three characteristics: 1) functional classification; 2) AADT (converted to VMT to account for segment length); and 3) speed limit. After the segmented data are analyzed by the model, the GIS

framework then provides a method to display the results for each segment on a color scaled map allowing for easy identification of hotspots using contrasting colors. A sample analysis was presented to demonstrate how the method could be applied for a safety study on hotspots. This will allow staff at UDOT to accurately evaluate the safety needs of roadways in the state.

7.2 Future Research

In the development of the statistical model and GIS framework for this study three areas for further research to enhance the deliverables of this project were identified: 1) the development of a graphical user interface (GUI) for the statistical model; 2) application of the safety model and GIS prototype to identify hotspots for analysis; and 3) determining acceptable methods for including road geometry in the model analysis, while maintaining acceptable segment lengths.

7.2.1 Graphical User Interface

Currently it is necessary to manipulate the model code in order to run the model. Since most transportation industry practitioners are unfamiliar with the R programming language this can prove difficult even when step by step instructions are provided. The development of a GUI would provide a more simple and intuitive way for practitioners to input data and run the model without manipulating the program code.

7.2.2 Model Application and Hotspot Analysis

The next step in the development of safety research in the state of Utah is the application of the model developed in this research to apply the safety model and GIS prototype to identify hotspots for analysis. The results of the hotspot analysis can be used to diagnose crash causes, while proven countermeasures can then be implemented to address the crash causes diagnosed. Following the application of the model and hotspot analysis, a systemic approach to safety can be applied with the proven countermeasure by identifying and prioritizing projects within the state where safety can be improved.

7.2.3 Methods for Including Road Geometry Data

Including road geometry data in the analysis creates several data accuracy issues and creates a large number of very short segments. This complicates the statistical analysis because the short segments result in a large number of segments where zero crashes are reported and a higher number of segments requires more computing time to analyze. Determining an acceptable method of aggregating road geometry values would result in increased accuracy in the statistical analysis because more data could be included without producing an excessive number of segments.

REFERENCES

American Automobile Association (AAA). (2010). "usRAP: United States Road Assessment Program." < http://www.usrap.us/home/>. (22 Aug, 2011).

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

Anselin, L. (1995). "Local Indicators of Spatial Association-LISA." *Geographical Analysis,* 27 (2), 93-115.

Bapna, S., Grebele, F., and McCall, M. (2003). "Spatially Referenced Crash Data System for Application to Commercial Motor Vehicles." *National Transportation Center, Morgan State University.* <www.morgan.edu/Documents/ACADEMICS/SOE/ntc/Spatially_ Bapna_0203.pdf >, (May 4, 2012).

Breyer, J. P. (2000). "Tools to Identify Safety Issues for a Corridor Safety-Improvement Program." *Transportation Research Record: Journal of the Transportation Research Board,* No. 1717, 19-27.

Carmody, M. T. (2011). "Efficiencies in Web-Based Crash Reporting." *ITE Journal,* March 2011, 16-19.

Center for Advanced Public Safety (CAPS). (2009a). "CARE – Critical Analysis Reporting Environment." < http://care.cs.ua.edu/care.aspx>, (Sep 13, 2011).

Center for Advanced Public Safety (CAPS). (2009b). "eCrash – Electronic Traffic Crash Reporting." < http://care.cs.ua.edu/eCrash.aspx>, (Apr. 19, 2012).

Davis, A. D., Xiong, H., and Tao, H. (2009). "Application of Bayesian Statistics to Identify Highway Sections with Atypically High Rates of Median-Crossing Crashes." *Transportation Research Record: Journal of the Transportation Research Board*, No. 2136, 77-81.

DeLucia, B. H., and Scopatz, R. A. (2010). "E-Crash: The Model Electronic Crash Data Collection System." Report DOT HS 811 326, National Highway Traffic Safety Administration (NHTSA), Washington, DC.

Esri. (2011a). “Crash Safety Dashboard Template for ArcGIS 10 – Flex 2.1.” <http://www.arcgis.com/home/item.html?id=9b3c2b82dadb46208e3f5e386b43454a>, (Feb. 28, 2012).

Esri. (2011b). “ArcGIS API for Flex.” < http://help.arcgis.com/en/webapi/flex/index.html>, (Mar. 14, 2012).

Esri. (2011c). “Linear Referencing with ArcGIS Desktop” < http://training.Esri.com/gateway/index.cfm?fa=catalog.webCourseDetail&courseid=1296>, (Jul. 5, 2011).

Esri. (2012a). “ArcGIS for Desktop.” <http://www.Esri.com/software/arcgis/arcgis-for-desktop/features.html> (Mar. 5, 2012).

Esri. (2012b). “Density Analysis.” < http://www.Esri.com/software/arcgis/extensions/spatialanalyst/density.html> (May. 2, 2012).

Esri. (2012c). “Crash Safety Analysis Tools Template for ArcGIS 10.” < http://www.arcgis.com/home/item.html?id=5537f21dc2d54d369fd14ea951648077> (Jan. 17, 2012).

Federal Highway Administration (FHWA) (2011). “’5 Percent Report’ Requirement.” Highway Safety Improvement Program (HSIP), <http://www.fhwa.dot.gov> (April 19, 2012).

Gelfand A. E. and Smith A. F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities". *Journal of American Statistical Association*, 85, 398-409.

Getis, A., and Ord, J. K. (1992). “The Analysis of Spatial Association by Use of Distance Statistics.” *Geographical Analysis,* 24 (3), 189-206.

Graettinger, A., Lindly, J. K., and Mistry, G. J. (2005). “Display and Analysis of Crash Data.” Report 03102, University Transportation Center for Alabama (UTCA), Tuscaloosa, AL.

Graettinger, A. J., McFadden, J., and Rushing, T. W. (2001). “Evaluation of Inexpensive Global Positioning System Units to Improve Crash Location Data.” *Transportation Research Record: Journal of the Transportation Research Board,* No. 1746, 94-101.

Hadi, M. A., Aruldhas, J., Chow, L., and Wattleworth, J. A. (1995). “Estimating Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression.” *Transportation Research Record*: *Journal of the Transportation Research Board,* No. 1500, 169-177.

Harkey, D. L. (1999). “Evaluation of Truck Crashes Using a GIS-Based Crash Referencing and Analysis System.” *Transportation Research Record: Journal of the Transportation Research Board,* No. 1686, 13-21.

Harwood, D. W., Gilmore, D. K., Bauer, K. M., Souleyrette, R., and Hans, Z. N. (2010). "usRAP Pilot Program-Phase III Final Report." Project No. 310537.1.021, AAA Foundation for Traffic Safety, Midwest Research Institute, Ames, IA.

Hauer E. (1997). *Observational Before-After Studies in Road Safety,* 1st Ed., Tarrytown, New York.

Herbel, S., Laing, L., and McGovern, C. (2010). Highway Safety Improvement Program (HSIP) Manual, Report DTFH61-05-D-00026. Federal Highway Administration, Washington, DC.

Iowa Department of Transportation (IDOT). (2012). "Crash Mapping Analysis Tool (CMAT)." <http://www.iowadot.gov/crashanalysis/cmatmain.htm> (May 5, 2011).

Ivan, J. N., Wang, C., Bernardo, N. R. (2000). "Explaining Two-Lane Highway Crash Rates using Land Use and Hourly Exposure." *Accident Analysis & Prevention*, 32(6), 787-795.

Jovanis, P. P., and Chang, H. L. (1986). "Modeling the Relationship of Accidents to Miles Traveled." *Transportation Research Record: Journal of the Transportation Research Board*, No. 1068, 42-51.

Khan, G., Santiago-Chaparro, K. R., Qin, X., and Noyce, D. A. (2009). "Application and Integration of Lattice Data Analysis, Network K-Functions, and Geographic Information System Software to Study Ice-Related Crashes." *Transportation Research Record: Journal of the Transportation Research Board,* No. 2136, 67-76.

Khan, G., Qin, X., and Noyce, D. A. (2008). "Spatial Analysis of Weather Crash Patterns." *Journal of Transportation Engineering,* 134(5), 191-202.

Laerd Statistics. (2012). "Descriptive and Inferential Statistics." <https://statistics.laerd.com/ statistical-guides/descriptive-inferential-statistics.php>. (Mar 23, 2012).

Lambert, D. (1992). "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics,* 34(1), 1-14.

Li, L., and Zhang, Y. (2007). "Bayesian Approach Based on Geographic Information System to Identify Hazardous Roadway Segments for Traffic Crashes." *Transportation Research Record: Journal of the Transportation Research Board,* No. 2024, 63-72.

MacNab, Y. C. (2004). "Bayesian Spatial and Ecological Models for Small-Area Accident and Injury Analysis." *Accident Analysis and Prevention,* 36, 1019-1028.

Massachusetts Traffic Records Analysis Center (MassTRAC). (2011). "MassTRAC." <http://masstrac.chs.state.ma.us/> (Apr. 9, 2012).

Miaou, S., and Song, J. J. (2005). "Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability Concept, Statistical Criterion, and Spatial Dependence." *Accident Analysis and Prevention*, 37, 699-720.

Miller, J. S. (2000). "Geographical Information Systems Unique Analytic Capabilities for the Traffic Safety Community." *Transportation Research Record: Journal of the Transportation Research Board*, No. 1734, 21-28.

Miller, J. S. (1999). "What Value May Geographic Information Systems Add to the Art of Identifying Crash Countermeasures?" VTRC 99-R12 Charlottesville, VA.

Mitra, S. (2009). "Spatial Autocorrelation and Bayesian Spatial Statistical Method for Analyzing Intersections Prone to Injury Crashes." *Transportation Research Record: Journal of the Transportation Research Board*, No. 2136, 92-100.

Ord, J. K., and Getis, A. (1995). "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." *Geographical Analysis*, 27 (4), 286-306.

Pulugarth, S. S., Krishnakumar, C. K., and Nambisan, S. S. (2007). "New Methods to Identify and Rank High Pedestrian Crash Zones: An Illustration." *Accident Analysis and Prevention*, 39, 800-811.

Qin, X., Ivan, J. N., Ravishanker, N., and Liu, J. (2005). "Hierarchical Bayesian Estimation of Safety Performance Functions for Two-Lane Highways Using Markov Chain Monte Carlo Modeling." *Journal of Transportation Engineering*, 131(345), 345-351.

Qin, X., and Wellner, A. (2011). "Development of Safety Screening Tool for High Risk Rural Roads in South Dakota." Report MPC-11-231, Civil Laboratory for Operations and Safety Engineering in Transportation, Department of Civil and Environmental Engineering, South Dakota State University, Brookings, SD.

R Project for Statistical Computing (RPSC). (2012). "What is R," <http://www.r-project.org> (May 2, 2012).

Roche, J. (2000). "Geographic Information Systems-Based Crash Data Analysis and the Benefits to Traffic Safety." *MTC Transportation Scholars Conference*, Ames, IA, 110-122.

Roess, R. P., Prassas, E. S., and McShane, W. W. (2004). *Traffic Engineering*, 3rd Ed., Pearson Education, Upper Saddle River, NJ.

Rutgers Center for Advanced Infrastructure and Transportation (CAIT). (2011). “Plan4Safety.” <http://cait.rutgers.edu/tsrc/plan4safety> (May 19, 2011).

SafeRoadMaps. (2012). “Safe Road Maps: Envisioning Safer Roads.” < http://saferoadmaps.org/>, (Mar. 5, 2012).

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.

Schultz, G. G., Thurgood, D. J., Olsen, A. N., and Reese, C.S. (2010). “Transportation Safety Data and Analysis, Volume 1: Analyzing the Effectiveness of Safety Measures using Bayesian Methods.” Report UT-10.12a, Utah Department of Transportation Traffic & Safety, Research Divisions, Salt Lake City, UT.

Schultz, G. G., Dudley, S. C., and Saito, M. (2011). “Transportation Safety Data and Analysis, Volume 3: Framework for Highway Safety Mitigation and Workforce Development.” Report UT-10.12c, Utah Department of Transportation Traffic & Safety, Research Divisions, Salt Lake City, UT.

Sloboda, B. (2009). *Transportation Statistics,* J. Ross Publishing, Ft. Lauderdale, FL.

Strathman, J. G., Dueker, K. J., Zhang, J., and Williams, T. (2001). “Analysis of Design Attributes and Crashes on the Oregon Highway System.” Report FHWA-OR-RD-09-01. Oregon Department of Transportation (ODOT), Salem OR. and Federal Highway Administration, Washington, DC.

State of Alabama-CARE (2009). “CARE 9 Individualized Training.” http://caps.ua.edu/docs/L.E.%20CARE9%20User%20Manual.pdf, (May 4, 2012).

The Bay Citizen. (2011). “The Bay Citizen Bike Accident Tracker.” <http://www.baycitizen.org/data/bike-accidents/>, (Sep 29, 2011).

UMassSafe. (2006). “UMassSafe Traffic Safety Research Program.” <http://www.ecs.umass.edu/umasssafe/>, (Sep. 6, 2011).

U.S. Department of Transportation (USDOT). (1999). “GIS-Based Crash Referencing and Analysis System.” Report FHWA-RD-99-081, Federal Highway Administration, Washington, DC,

Utah Automated Geographic Resource Center (AGRC). (2012). “Utah GIS Portal.” <http://www.gis.utah.gov/ > (May 4, 2012).

Utah Department of Transportation (UDOT). (2011a). “Roadware Visiweb.” <http://roadview.udot.utah.gov/visiweb/index.aspx> (May. 12, 2011).

Utah Department of Transportation (UDOT). (2011b). “UPLAN.” <http://utahplanning.org/login/default.aspx> (May 2, 2011).

Utah Department of Transportation (UDOT). (2012). “UDOT Data Portal - uGate.” <http://www.udot.utah.gov/ugate> (Feb 13, 2012).

LIST OF ACRONYMS

AAA	American Automobile Association
AADT	Annual Average Daily Traffic
AASHTO	American Association of State Highway and Transportation Officials
ADOT	Arizona Department of Transportation
AGRC	Automated Geographic Resource Center
CAIT	Center for Advanced Infrastructure and Transportation
CAPS	Center for Advanced Public Safety
CARE	Critical Analysis Reporting Environment
CMAT	Crash Mapping Analysis Tool
CMF	Crash Modification Factor
CMV	Commercial Motor Vehicle
CTRE	Center for Transportation Research and Education
DOT	Department of Transportation
EB	Empirical Bayes
FHWA	Federal Highway Administration
GIS	Geographic Information System
GUI	Graphical User Interface
HSIP	Highway Safety Improvement Program
HSM	Highway Safety Manual
IDOT	Iowa Department of Transportation
lat/long	Latitude/Longitude
LRS	Linear Referencing System
MassTRAC	Massachusetts Traffic Analysis Center
MEV	Million Entering Vehicles
MPO	Metropolitan Planning Organization

MCMC	Markov Chain Monte Carlo
MVMT	Million Vehicle Miles Traveled
NHTSA	National Highway Transportation Safety Administration
OLS	Ordinary Least Squares Regression
RTM	Regression to The Mean
SPF	Safety Performance Function
TAC	Technical Advisory Committee
TRB	Transportation Research Board
UDOT	Utah Department of Transportation
usRAP	United States Road Assessment Program
VMT	Vehicle Miles Traveled
ZIP	Zero-inflated Poisson

APPENDIX A. DATA PREPARATION

Various changes were to be made to each dataset to prepare it for use in the GIS. This appendix provides a brief step-by-step explanation of what was done. Table 4-1 is reproduced here as Table A-1 for a review of where each dataset came from.

Table A-1. Data Source Summary

Dataset	Source	Format	Future Availability
State Routes	Utah AGRC	LRS Feature Class	Updated Regularly
Crash Data	Scott Jones	CSV Tables (Excel)	Updated at least Annually
AADT	Frank Pisani, Lee Theobald	Excel Spreadsheet	Updated Annually
Truck AADT	Frank Pisani, Lee Theobald	Excel Spreadsheet	Updated Annually
Speed Limit	Larry Montoya	Shapefile	TBD
Functional Class	Interplan-Charles Allen	Excel Spreadsheet	TBD
Through Lanes	UGATE	KML file	TBD
Urban Code	UGATE	Shapefile	TBD
Road Geometry	Gary Kuhl	Access Database	New inventory project underway
Skid Index	Gary Kuhl	Excel Spreadsheet	Half of state updated each year

AADT, Truck AADT, and Historic AADT

1. Rename column headings to match data uniformity recommendations
2. Copy all Interstate segments and replace “P” with “X” for direction
3. Fix ROUTE_ID for route 89A

Historic Truck AADT File

1. Rename headings to match data uniformity recommendations
2. Copy all Interstate segments and replace “P” with “X” for direction
3. Fix ROUTE_ID for route 89A
4. Delete empty rows
5. Add columns and compute combined truck count and percentage
6. Add single/combo truck count
7. For 2004 File:
    a. Move columns and rename as needed
    b. Delete unnecessary columns
    c. Combine EB/WB counts for uniformity with other historic files
    d. Add four-digit ROUTE_ID

Functional Classification

1. Rename headings to match data uniformity recommendations
2. Copy all Interstate segments and replace “P” with “X” for direction

Speed Limit

1. Export shapefile .dbf to excel for editing
2. Rename headings to match data uniformity recommendations
3. Add DIRECTION and LABEL fields
4. Copy all Interstate segments and replace “P” with “X” for direction

Skid Index

1. Rename headings to match data uniformity recommendations
2. Add DIRECTION and LABEL fields
3. Copy all Interstate segments and replace “P” with “X” for direction
4. Create BEG_MILEPOINT and END_MILEPOINT using excel equation (the existing milepoint value was kept to preserve the actual test location)

Road Geometry

1. Export “HorizontalCurve” table from Access to Excel
2. Rename headings to match data uniformity recommendations
3. Add ROUTE_ID and LABEL fields
4. Replace “N” direction with “X” direction where applicable
5. Add “Direction_Collected” field to indicate direction of travel while collecting data
6. Add data for route 39 from 2009 file to 2010 file (was not collected in 2010)

Number of Lanes

1. Convert KML file to Layer file
2. Export layer file data to Excel
3. Extract route, direction, milepoint, and number of lanes from popup_text field
4. Rename headings to match data uniformity recommendations
5. Add DIRECTION and LABEL fields
6. Copy all Interstate segments and replace “P” with “X” for direction
7. Remove ramps and local roads from file
8. Remove duplicated records

Urban Code

1. Convert KML file to Layer file
2. Export layer file data to Excel
3. Extract route, direction, milepoint, and urban code from popup_text field
4. Rename headings to match data uniformity recommendations
5. Add ROUTE_ID, DIRECTION and LABEL fields
6. Copy all Interstate segments and replace “P” with “X” for direction
7. Remove ramps and local roads from file
8. Add nominal urban code definition to match existing numerical code

Crash Data (2006-2010)

VB Macros used are listed in parentheses

1. Join Crash Database and location database based on Crash_ID (CMD_Join_Database)
2. Pull direction of first vehicle from vehicle database and join to #1 (Crash_Direction)
3. Remove ramp crashes and non-state route crashes
4. Add 4-digit ROUTE_ID
5. Convert vehicle direction code from #2 to either “P” or “X” (Direction_Code)
6. Rename and organize columns for uniformity

APPENDIX B. CUSTOM GIS TOOLS DEVELOPED

Custom GIS tools were developed to simplify processes performed regularly in analysis. These are packaged in a toolbox as shown in Figure B-1. This appendix will briefly describe each tool with its input dialog box and workflow from ModelBuilder.

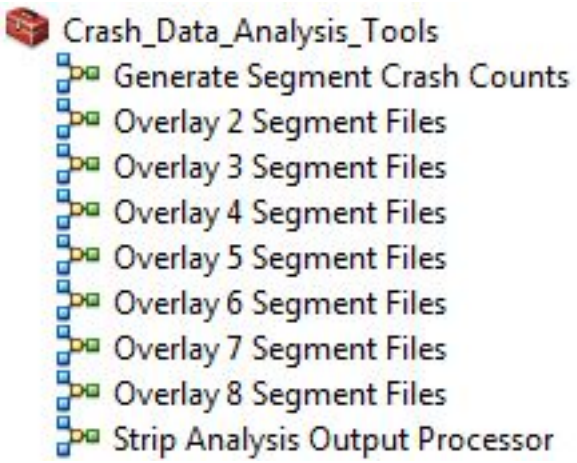


Figure B-1. Crash Data Analysis Toolbox.

"Overlay 3 Segment Files"

This and the other "Overlay" tools all perform the same function. The only difference is the number of datasets that are overlaid. The red "X" circle seen in Figure B-2 disappears upon entering parameters. Figure B-3 shows the flowchart from ArcGIS ModelBuilder.

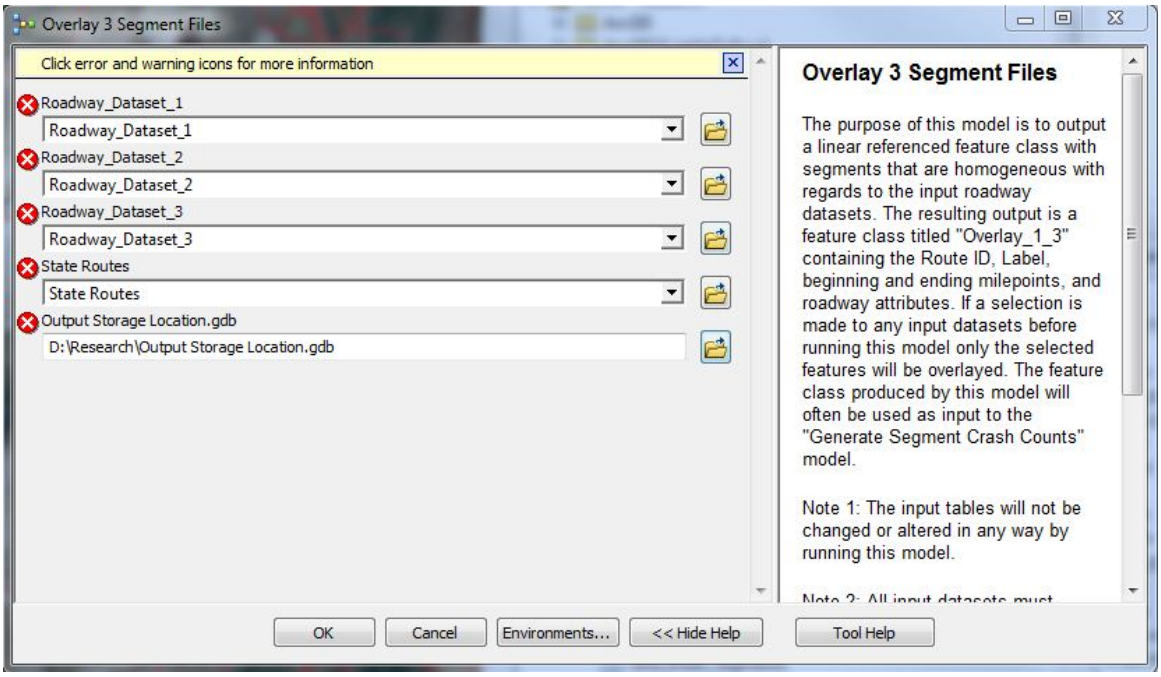


Figure B-2. Overlay Segments Dialog box.

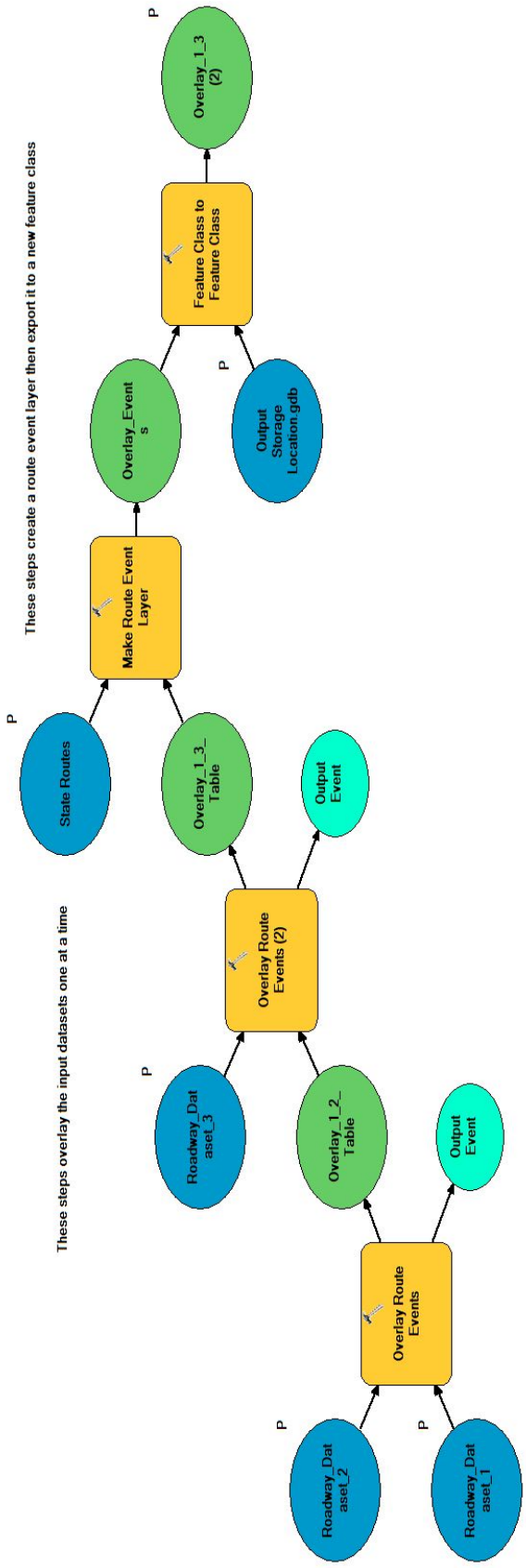


Figure B-3. Overlay Tool Flowchart from ModelBuilder.

“Generate Segment Crash Counts”

This model will count the number of crashes on the input segments. The red “X” circle seen in Figure B-4 disappears upon entering parameters. Figure B-5 shows the crash count tool flowchart from ArcGIS ModelBuilder.

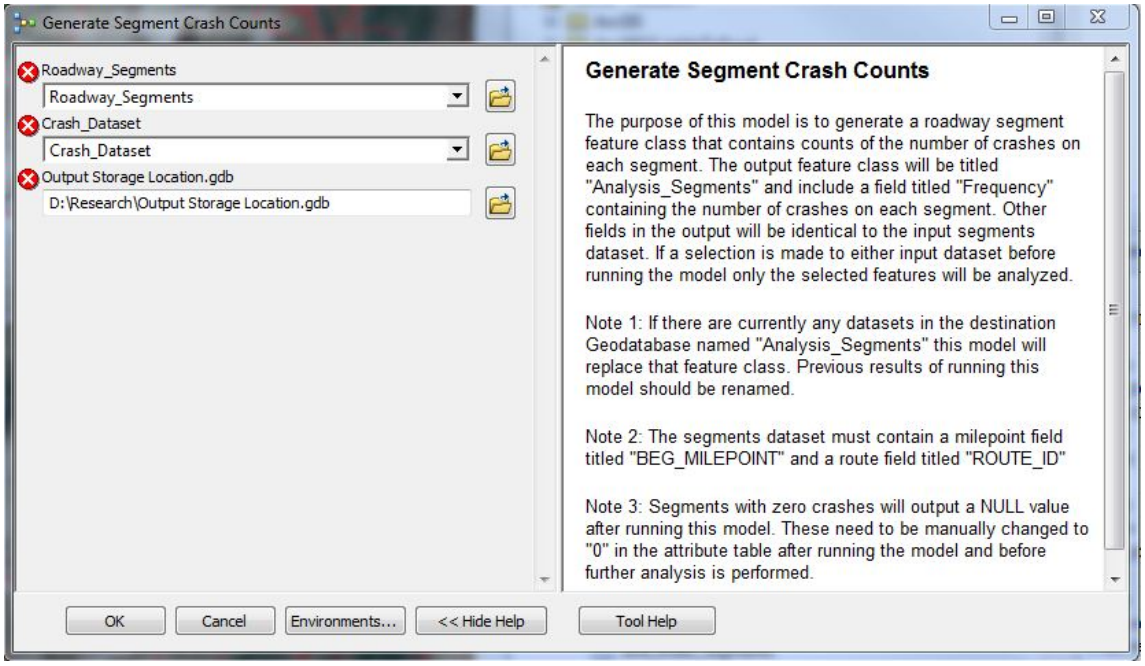


Figure B-4. Generate Crash Counts Dialog Box.

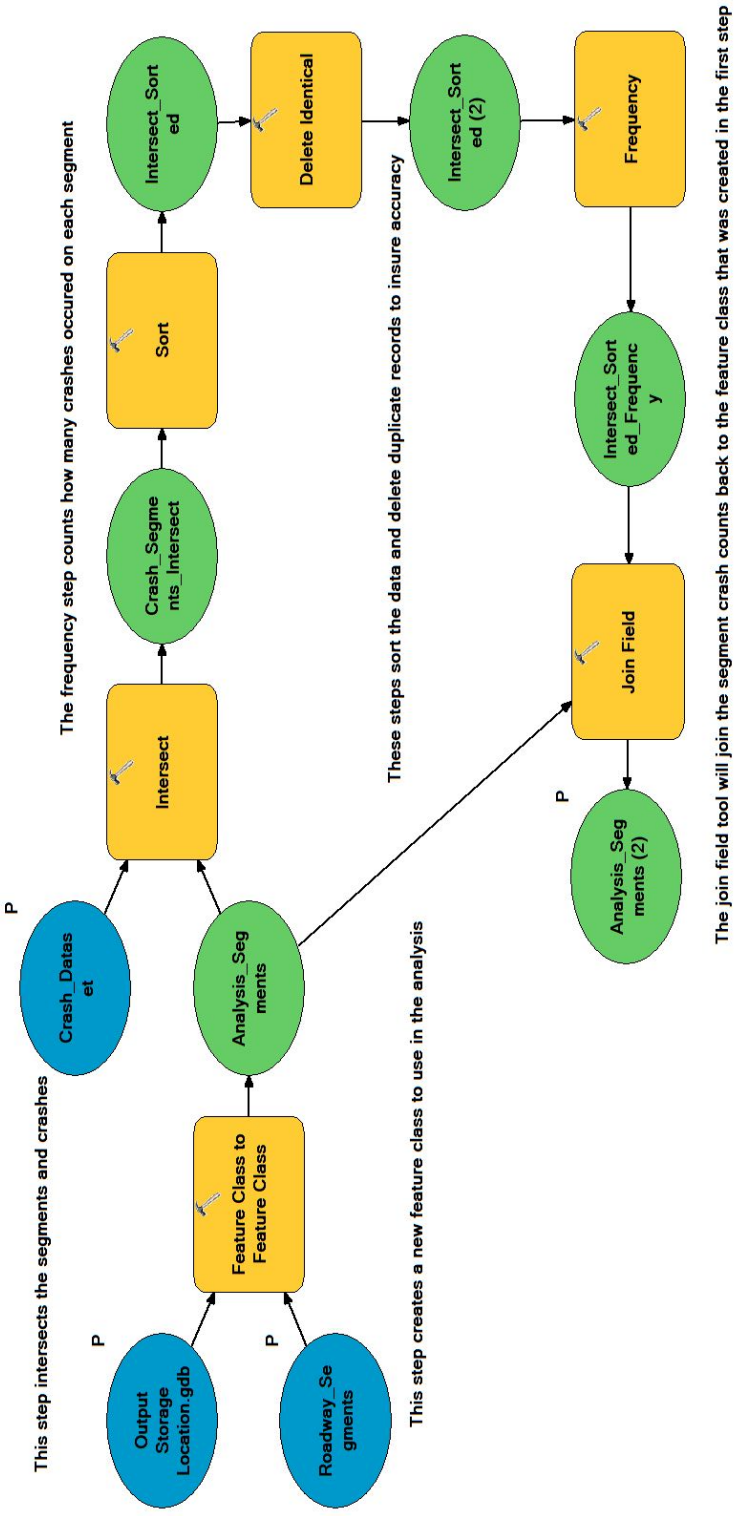


Figure B-5. Crash Count Tool Flowchart from ModelBuilder.

“Strip Analysis Output Processor”

This model prepares strip analysis outputs and prepares them for use in LRS-based analysis. The red “X” circle seen in Figure B-6 disappears upon entering parameters. Figure B-7 through Figure B-11 shows the flowchart in sections because the model is very large. The first and last step of each section is duplicated in each image.

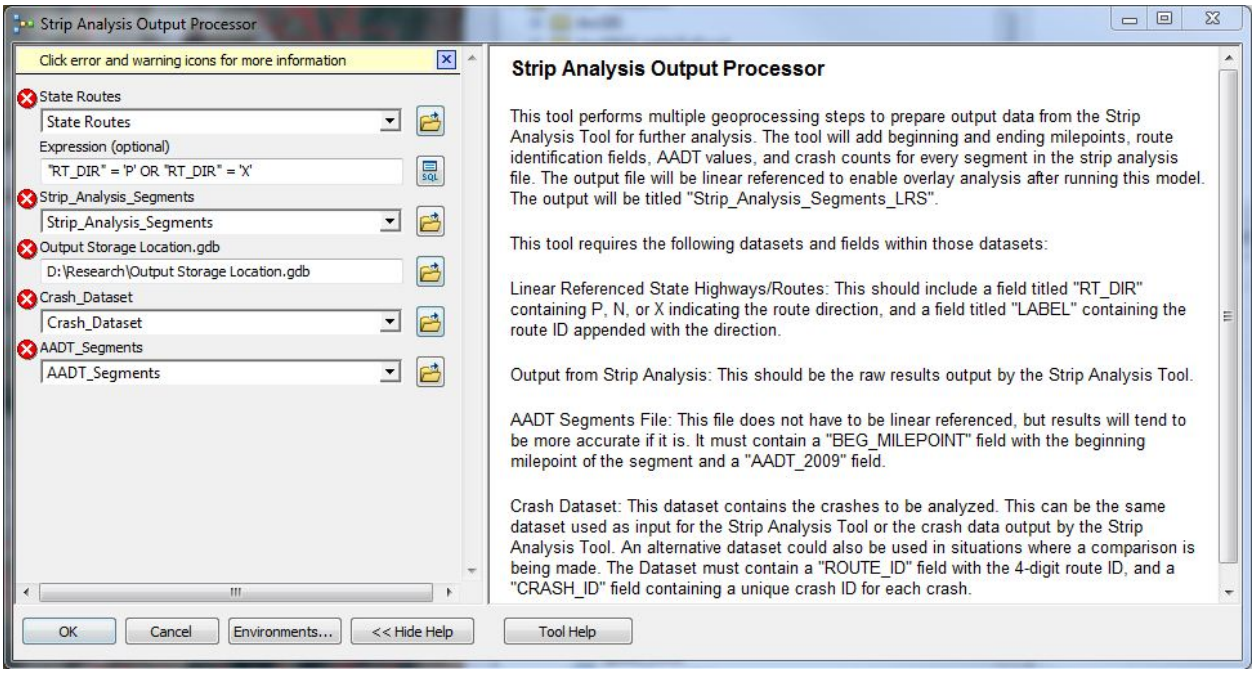


Figure B-6. Strip Analysis Output Dialog box.

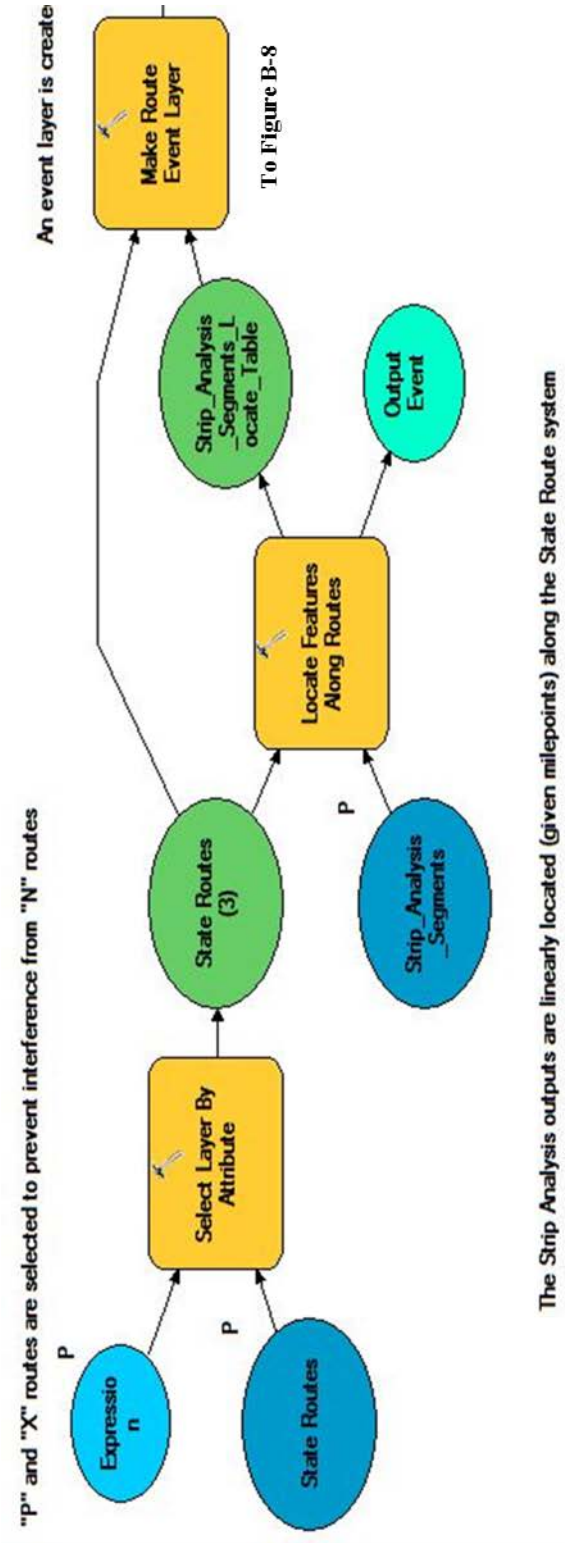


Figure B-7. Strip Analysis Flowchart Section 1.

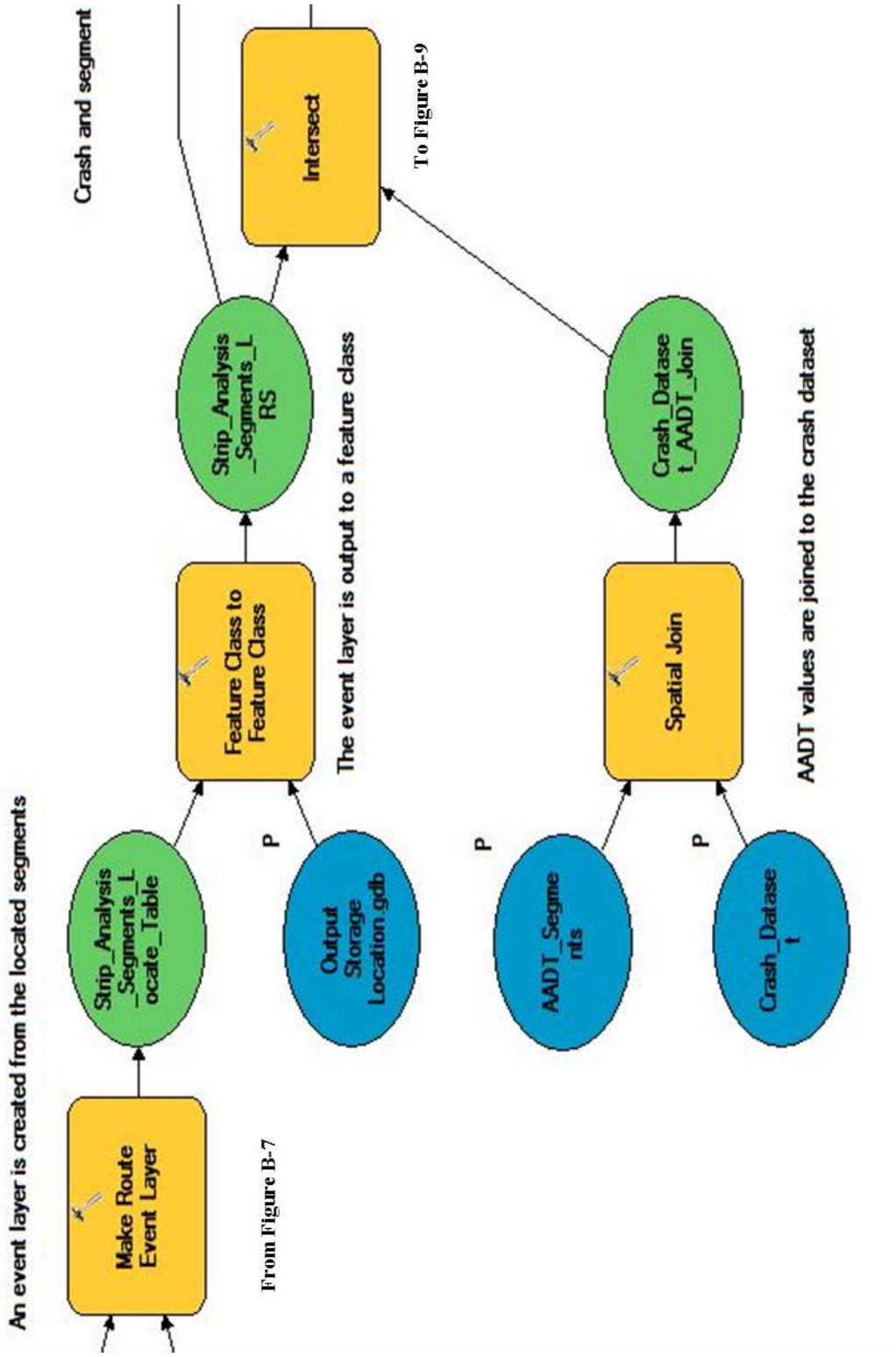


Figure B-8. Strip Analysis Flowchart Section 2.

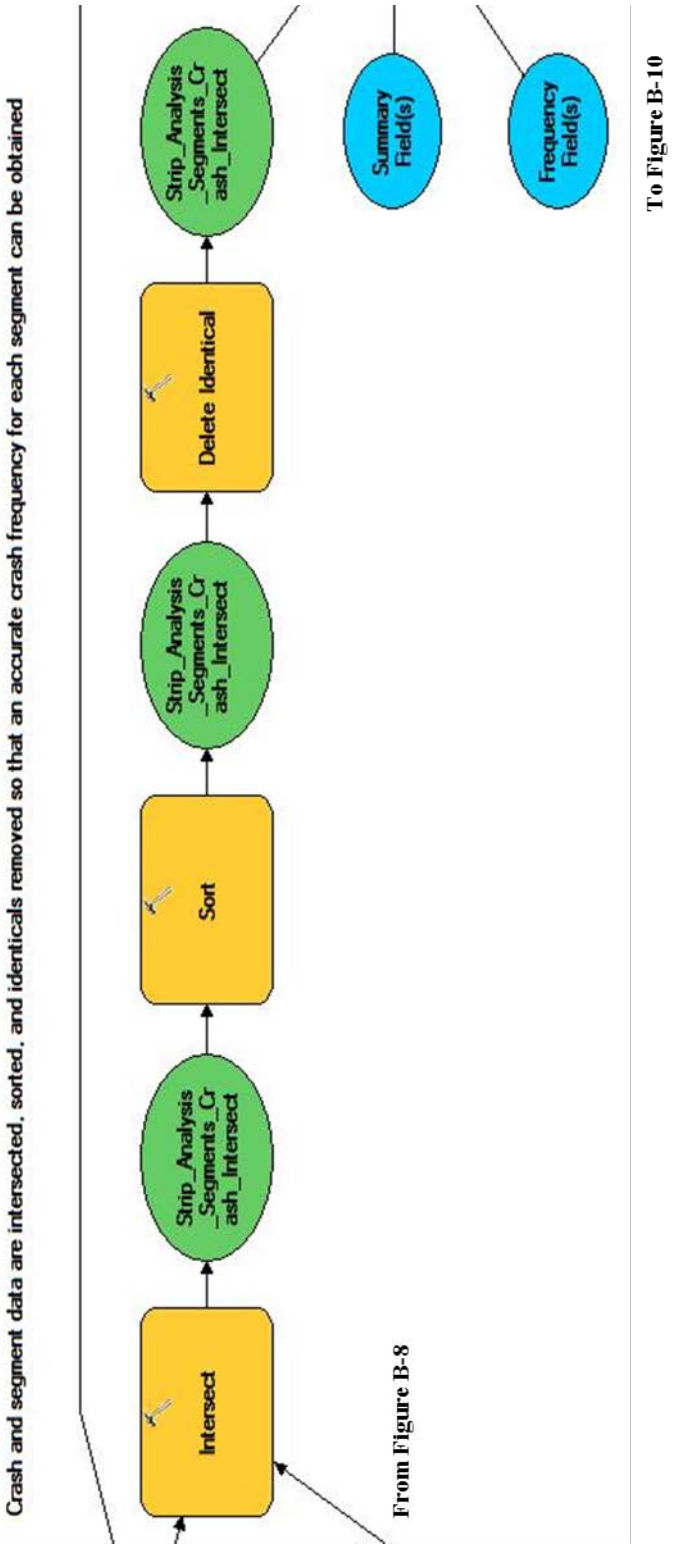


Figure B-9. Strip Analysis Flowchart Section 3.

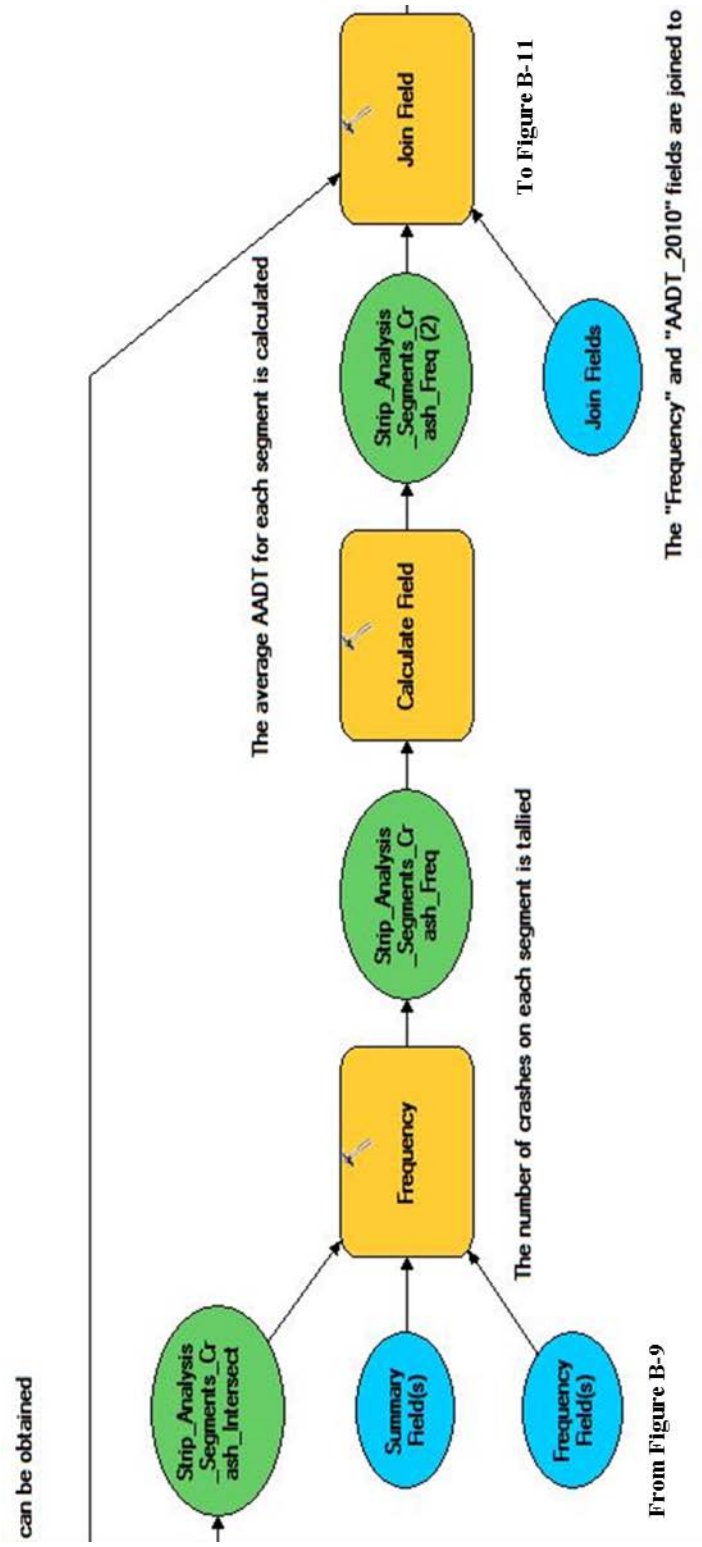


Figure B-10. Strip Analysis Flowchart Section 4.

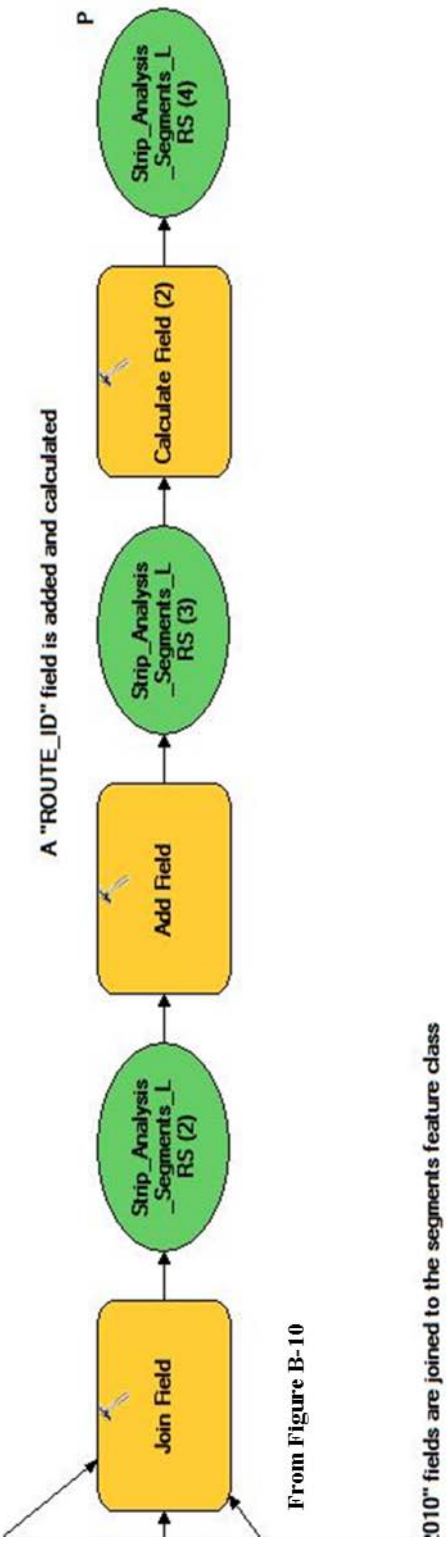


Figure B-11. Strip Analysis Flowchart Section 5.

APPENDIX C. US-89 MILEPOINT 473.7 DETAILED ANALYSIS

The following slides and associated commentary were presented to the UDOT TAC committee as a detailed analysis of what is causing crashes at that location.

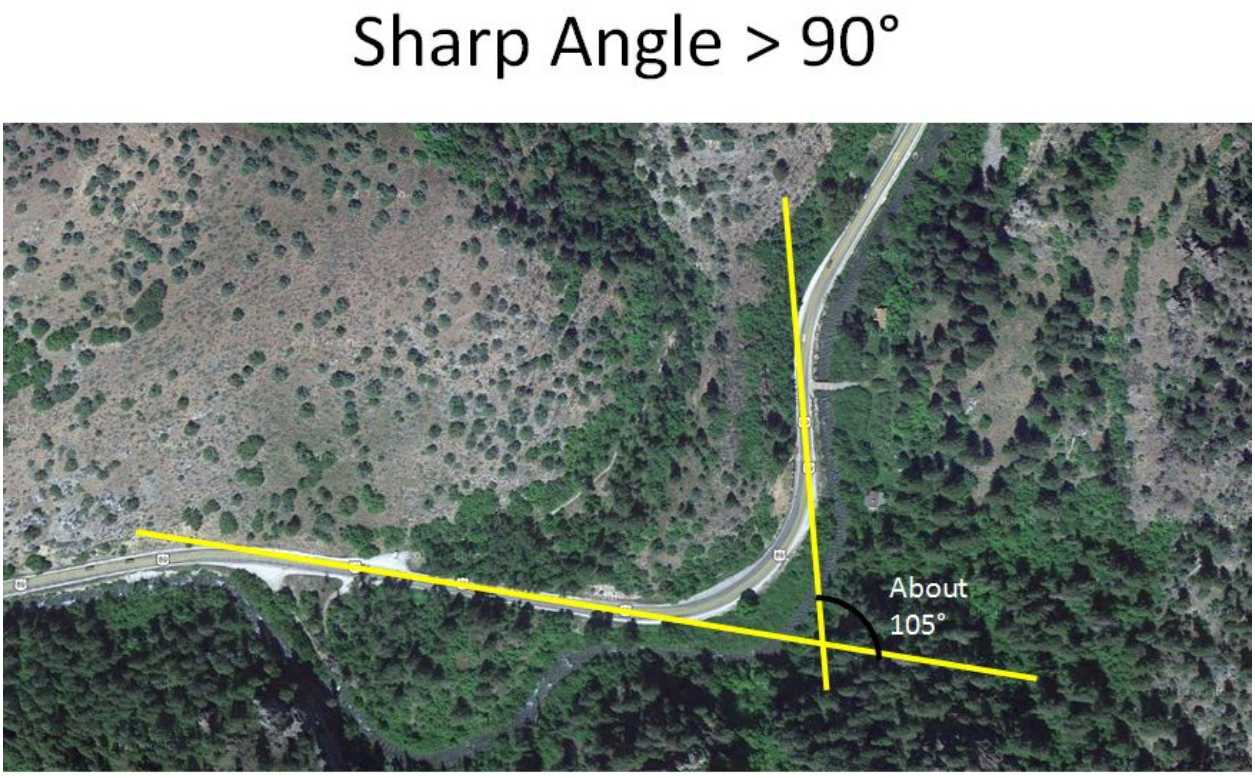


Figure C-1. Slide 1.

Figure C-1 shows that the curve angle is unusual, making it an inherently difficult maneuver. The second curve just after the bridge does not appear to be a factor because most crashes occur northbound before reaching the second curve.

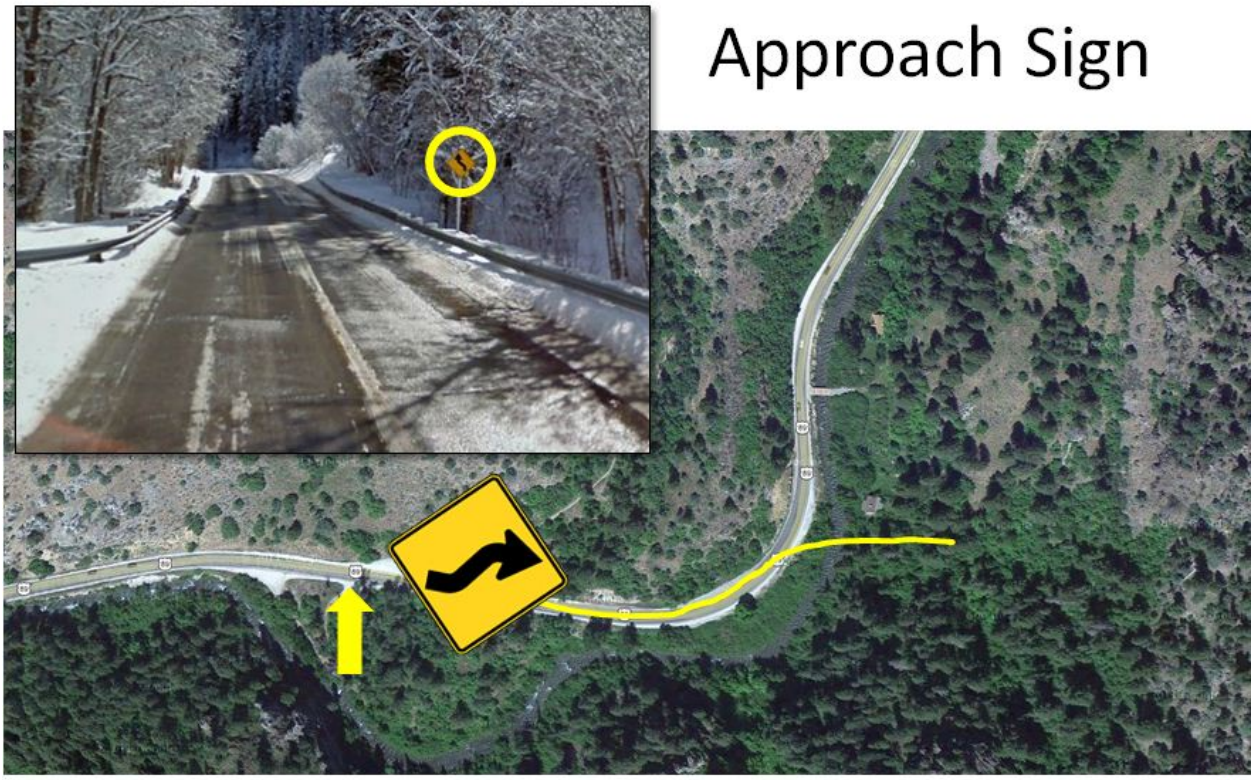


Figure C-2. Slide 2.

Figure C-2 shows that the curve sign on the northbound approach does not match the geometry. Drivers anticipate a mild curve left then curve right and to west. Instead it is a very sharp curve left, then right and they are headed roughly due north. Short sight distance magnifies the problem this misperception causes. Also, no speed reduction warning sign is posted.

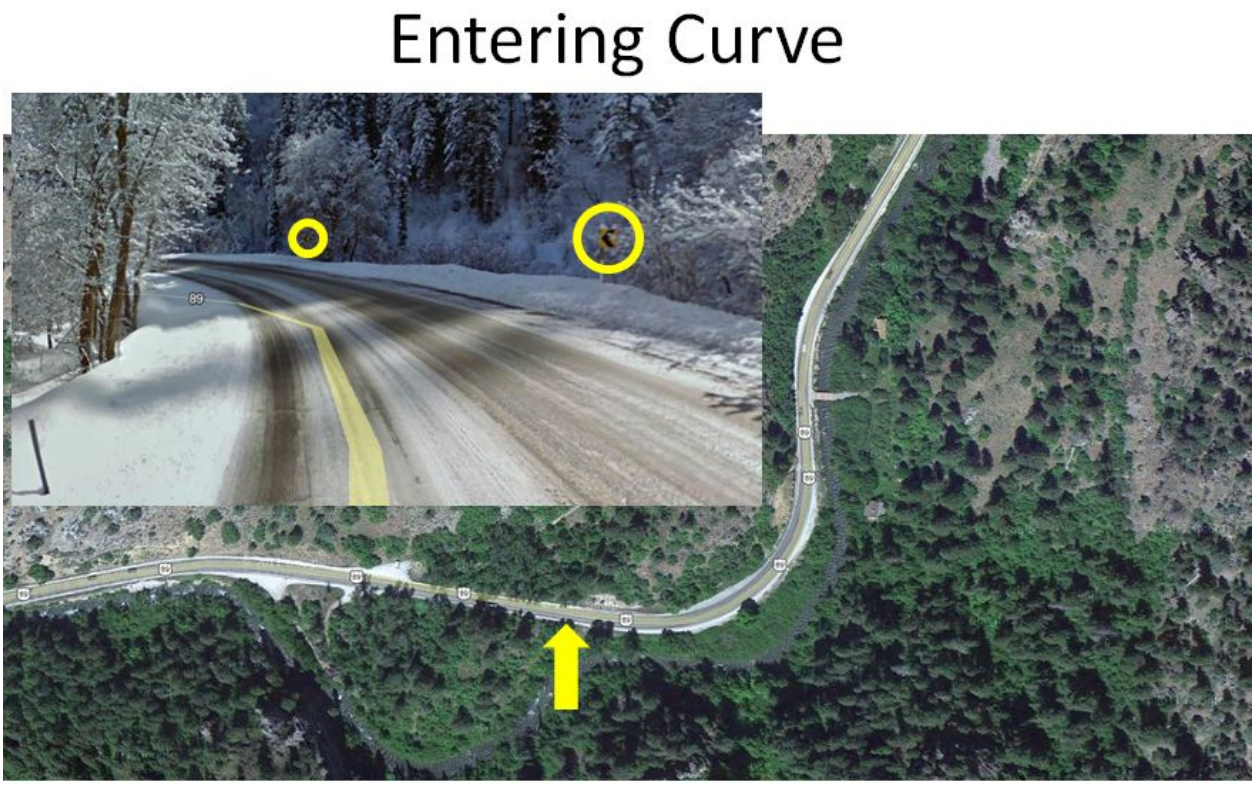


Figure C-3. Slide 3.

Figure C-3 shows that chevrons located to the right of the road are effective at indicating curve presence and direction. The sight distance is still limited.

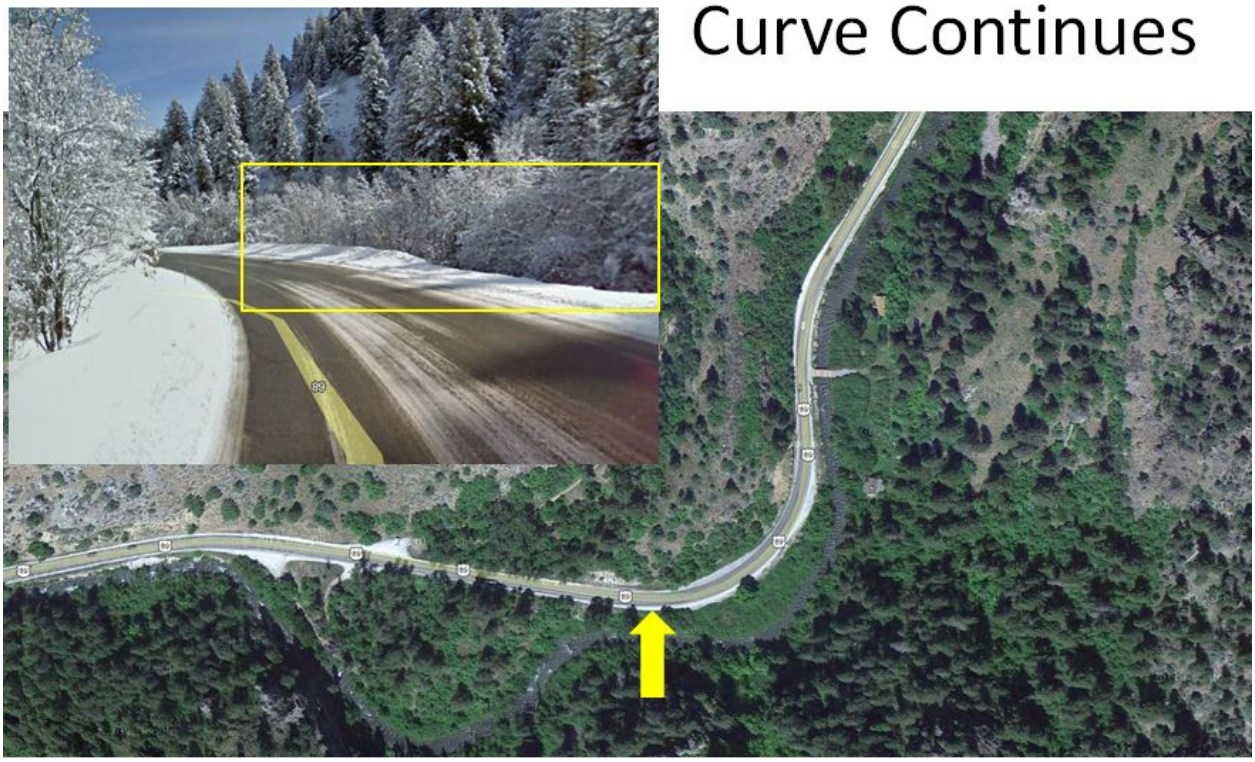


Figure C-4. Slide 4.

Figure C-4 shows that the driver's sight distance is still limited. No more chevrons are on the right side, which leads the driver to believe the curve is ending. Due to this and the initial curve approach sign the driver would probably expect the curve to straighten out immediately at their terminus of vision and then curve back to the right.

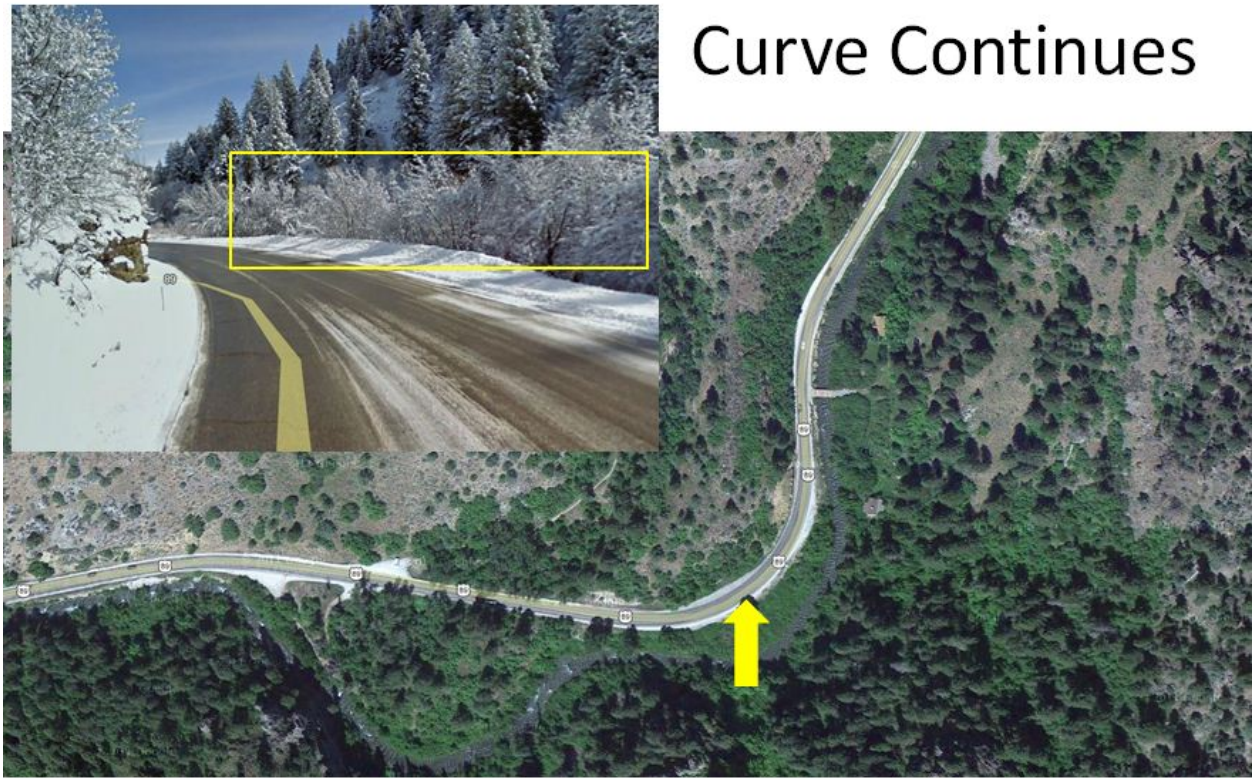


Figure C-5. Slide 5.

Figure C-5 shows that driver's sight distance is still limited. There still are no chevrons on the right side and the driver is again expecting the curve to straighten and curve back to the right.

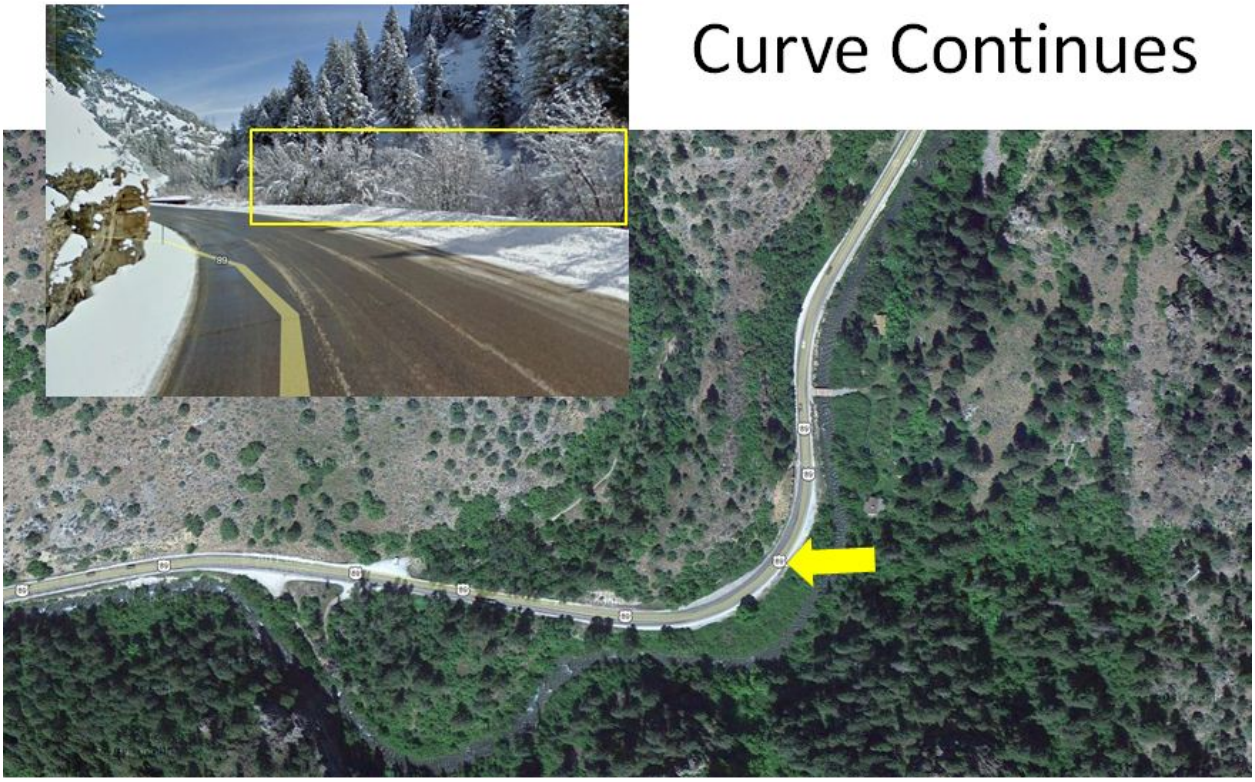


Figure C-6. Slide 6.

Figure C-6 shows that at this point with slightly increased sight distance the driver may begin to realize that the geometry is not what they were expecting. There are still no chevrons present to guide the driver, and very little time to correct the vehicle's trajectory.

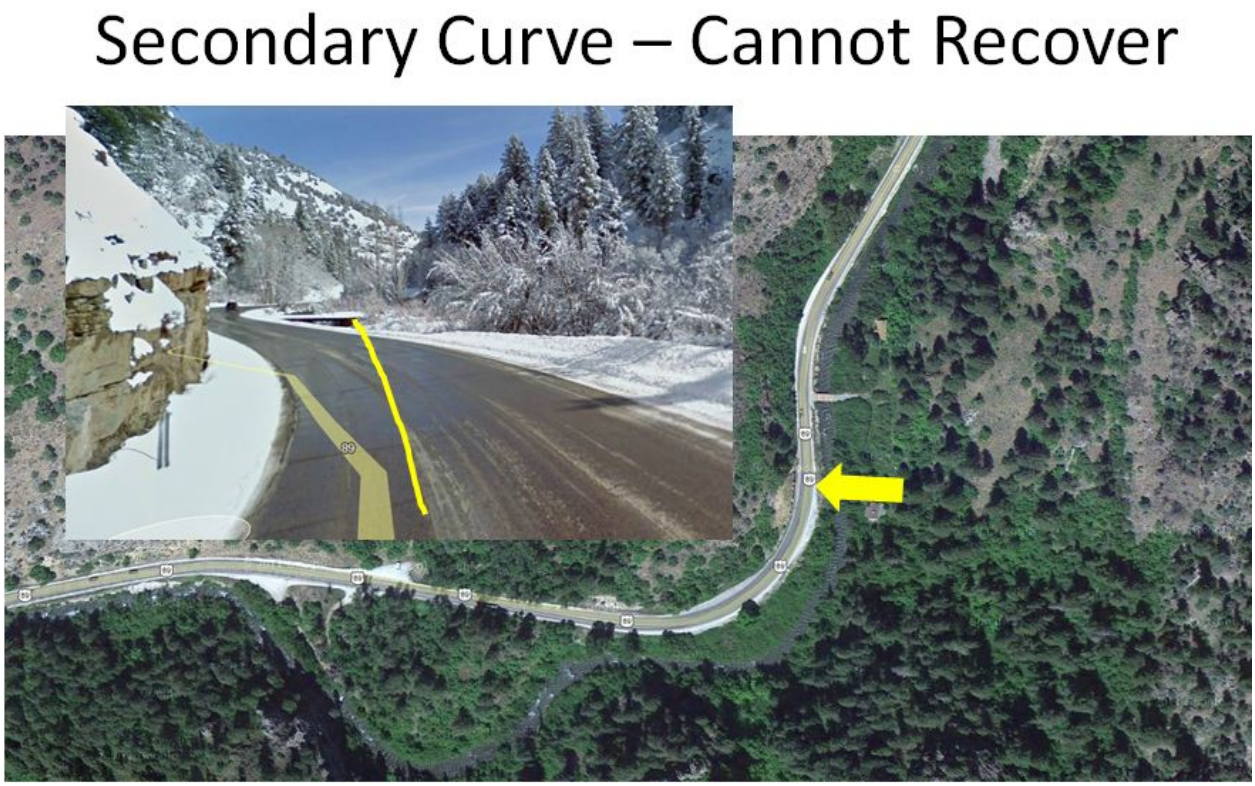


Figure C-7. Slide 7.

Figure C-7 shows that at this point the driver finally sees that there is still more curve ahead and to the left. This would likely cause the driver to steer hard left or press brakes firmly, both of which could contribute to sliding off the road. It is very likely that at this point speeds would be too fast for many people to recover and make the final turn in the curve, as they were expecting that the road would curve back to the right. This would cause many vehicles to slide into the river as shown by the yellow line and as the crash history suggests.