DEVELOPMENT OF SAFETY SCREENING TOOL FOR HIGH RISK RURAL ROADS IN SOUTH DAKOTA

Xiao Qin, Ph.D., PE
Assistant Professor

Adam Wellner, EIT
Graduate Research Assistant

Civil Laboratory for Operations and Safety Engineering In Transportation (CLOSET)
Department of Civil and Environmental Engineering
South Dakota State University

March 2011

Acknowledgements

The authors would like to thank the Mountain Plain Consortium (MPC) and South Dakota State University for providing the funding for this research. Thanks to the South Dakota Department of Transportation and South Dakota Department of Public Safety for providing the data used for the study. Thanks to Jason Anderson for his assistance in reviewing and preparation of this paper. Lastly, thanks to Melissa Mathis from USGS for assistance in troubleshooting the Python scripts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

In South Dakota, an average of 15,000 crashes, including 120 fatal crashes, occurr over 83,744 miles of highway every year. The density of these fatal crashes was barely 0.0015 crashes per mile. It is apparent that hot spot analysis may not be applicable for many locations. A system-wide deployment of safety treatments with substantial coverage may be more appropriate. However, without an effective systematic approach to identifying the boundaries of safety improvement projects, it will be cost prohibitive to repair and retrofit tens of thousands of miles of highways.

Given the sparsely distributed crashes across various highway systems, this study designed an empirical Bayes (EB) based sliding window technique within a spatial context. By examining roadway safety spatially, the safety analysts are able to account for high-risk locations completely within longer predefined segments and locations, which may include multiple predefined roadway segments. Removing the dependence on predefined segmentation can also bring to the forefront safety issues previously ignored. The robustness of the EB method significantly improves the crash estimation accuracy. In conjunction with several different but complementary safety metrics, a complete view of rural highway safety performance can be presented.

To ease the use of such a new technique, the South Dakota GIS Highway Safety Review (GIS-HSR) Tools was developed, which provides a data-driven approach toward identifying high-risk locations. Only very basic user input and interaction is required for the tool, which is implemented on a system-wide basis. SD GIS-HSR is designed to address the rural environment and is tuned specifically for South Dakota. However, the general architecture and design are valid in any location.

1. RURAL HIGHWAY SAFETY

The *Intermodal Surface Transportation Efficiency Act* of 1991 (ISTEA) required state DOTs to establish management systems in six areas, including safety. With ISTEA signed into law, highway safety programs are guaranteed a certain level of funding. This law drastically improved standards for crash data collection, archiving, management, and application. The *Transportation Equity Act for the 21st Century* (TEA-21), ISTEA's successor, continues to promote the overall importance of highway safety in highway programs and projects. As a result, highway safety research has advanced significantly during the past two decades. The methodologies focused on finding locations that have experienced unusually high numbers of crashes (so called "hot spots" or "black spots") have matured. Though crash rates are steadily decreasing, the number of fatal and severe crashes remains at the same level. Crashes resulting in fatalities or serious injuries generate tremendous economic losses, cause permanent damage to people's lives, and have an extensive impact on society. Hence, it is necessary to seek a new approach with a shift in emphasis: the prevention and reduction of the number of fatal and severe crashes, which has become the highlight of the current legislation, *The Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users* (SAFETEA-LU). This prioritization and reinvestment in the area of fatal and severe crashes recognizes that the causes or fatal and severe crashes may be different than the causes of less severe crashes. Re-engineering locations with high crash counts will not automatically reduce the number of fatal and type A injury crashes. It also recognizes that, for a long time, rural area safety has disadvantages compared with its urban counterpart because rural highways carry less traffic, which is normally used as a measure of crash potential.

Although only 23% of the U.S. population lived in rural areas, rural fatalities accounted for 57% of all traffic fatalities in 2007. The fatality rate was 2.5 times higher in rural areas than in urban areas, according to the National Highway Traffic Safety Administration (NHTSA) 2007 Crash Facts (NHTSA 2007). The need to identify and develop new approaches to address areas experiencing fatal and severe crashes is imperative and apparent. However, this task is a considerable challenge because these types of crashes are rare and often random. The current hot spot identification methodologies and related safety analysis tools become less effective when coping with random and widely distributed crash events. For example, in Minnesota in 2007, 352 fatal crashes occurred in rural areas or on the local system, which includes over 45,000 miles of two-lane highways, resulting in an extremely low density of fatal crashes (NHTSA 2007). South Dakota is another rural state that consists of 66 counties and 9 tribal governments. It is located in the Midwestern United States and covers 77,121 square miles. Its cities, towns, and villages are connected by 83,744 miles of highways, most of which are rural two-lane highways. In 2007, 87% of South Dakota's 146 fatal crashes occurred in rural areas. The density of these fatal crashes was barely 0.0015 crashes/mile. The low number of crashes is partly the result of South Dakota's low population of just over 800,000 and a population density of 9.9 persons per square mile, the fifth-lowest population density among U.S. states (US Census Bureau 2010). Given the state's enormous geographic size and long stretches of highways, South Dakotans are more likely to travel a longer distance on roadways that have a speed limit of 55 mph or higher to fulfill their needs (Drake et al. 2005). Considering vehicle miles traveled, the state fatality rate is 2.3, higher than the national average of 1.5, based on the NTHSA 2008 report. In 2003 and 2004, South Dakota's fatality rate ranked second highest in the nation (South Dakota Department of Transportation [SDDOT] 2007). Fatality trends in South Dakota are similar to neighboring rural states, but much higher than its more urban neighbors Minnesota, Iowa, and Nebraska, as shown in Table 1.1. In recent years, both the effort and investment in safety have significantly increased and the total number of crashes has steadily decreased. A better understanding of the cause-effect of a crash occurrence suggests that treating individual locations is less desirable in reducing extensively distributed fatal and severe crashes. A system-wide deployment of safety measures with substantial coverage may be appropriate. A systematic approach can lead to a wider deployment of appropriate low-cost safety measures over longer stretches of roadways and corridors, or throughout the entire highway system, such as with the installation of rumble strips to mitigate roadway departure crashes or the installation of median cable barriers to prevent cross-median crashes.

Table 1.1 South Dakota and Adjacent States

State	2001	2002	2003	2004	2005	2006	2007	2008
South Dakota	2	2.12	2.38	2.24	2.22	2.08	1.62	1.35
Iowa	1.51	1.51	1.48	1.45	1.45	1.4	1.43	1.34
Minnesota	1.51	1.51	1.48	1.45	0.98	0.87	0.89	0.78
Montana	2.3	2.59	2.41	2.04	2.26	2.34	2.45	2.12
Nebraska	1.36	1.64	1.54	1.32	1.43	1.39	1.32	1.09
North Dakota	1.51	1.51	1.48	1.45	1.62	1.41	1.42	1.33
Wyoming	2.16	1.96	1.79	1.77	1.88	2.07	1.60	1.68

Source: NHTSA 2009

An important question for system-wide safety improvements is: how extensive is “system-wide”? Is it state line to state line, entire counties, or all two-lane local rural highways? It will be cost prohibitive to repair and retrofit tens of thousands of miles of highways. Lacking a lucid definition of “coverage,” the goal of system-wide implementation is unattainable. The boundaries for safety treatment need to be quantitatively determined on a manageable scale. Another challenge is that the current systematic approach is less formal and less accurate than the matured hot spots identification method. This is the reason why the hot spot method continues to dominate in safety analysis despite the fact that extremely low numbers of crashes offer limited information for engineers to identify crash patterns and locate crash clusters. In South Dakota, an average of 15,000 crashes and 120 fatal crashes take place over 83,744 miles of highway every year. It is apparent that hot spot analysis may not be appropriate for many locations. If the safety improvements are solely determined based on prior crash locations, it is more likely that safety funding will be applied to chase a moving target and the specific projects may not address the real safety risks, providing a temporary solution to only part of the problem.

Given the sparsely distributed crashes across various highway systems, a methodology designed in a spatial context can assist in establishing the boundaries of safety improvements. By examining roadway safety in a spatial context, we are able to account for high risk locations completely within longer predefined segments and locations, which may include multiple predefined roadway segments. Removing the dependence on predefined segmentation can bring to the forefront safety issues previously ignored. It is hoped that identification and field investigation of these locations will lead to a greater understanding of contributing factors not currently considered visible in the data.

It is implicit that the new methodologies and technologies will not be effective until they are utilized by practitioners. The ease of use of a new technique depends on the data requirements and the ability of users to comprehend the underlying methodological approach. In many respects, the evolution of the methodology due to the complexity of the issues may exceed the limits of practical implementation. An analytical tool is usually developed to bridge the gap between theory and practice. The tool can take advantage of the available safety data, communicate effectively with the users, and present meaningful and informative results to safety stakeholders. The objective of this research is to develop a data-driven traffic safety screening tool for rural highways using a collection of computer and GIS techniques. A sophisticated safety index will be generated to effectively and accurately prioritize rural areas; both based on historical safety performance and when potential risk factors are considered. The key accomplishments of the project are summarized below:

- Given the random nature of a crash occurrence, look beyond the traditional “reactive” approach to the emerging “proactive” approach. A proactive approach is a preventive, forwarding-looking safety measure established on factorial analysis using the principles and causational findings from the past. It is not solely based on the historical crash data, but based on what can potentially be improved to prevent further crashes.
- Utilize and modify the state-of-the-art crash prediction models and safety analysis tools to the needs of South Dakota. Nationwide, traffic safety research has flourished in the past several decades. Our methodologies will synthesize the substantial knowledge from previous studies.
- Allow decisionmakers to explicitly weigh safety investment alternatives based on various safety metrics.
- Develop a GIS data framework to integrate South Dakota safety data, implement GIS techniques to analyze data spatially, and provide a user-friendly interface for practitioners. A design which considers the spatial relationship between crashes without dependence on predefined segments.

2. SOUTH DAKOTA RURAL HIGHWAY SAFETY FEATURES AND STATUS

South Dakota crash patterns are relatively consistent and stable over the years. The same mistakes continue to cause the majority of the state's fatal crashes: failure to wear a seat belt, running off the road, speeding, and aggressive and impaired driving. Of speeding-related crashes, 47% are fatal; 61% of fatal crashes involved a vehicle leaving the road due to factors such as a young or inexperienced driver, failure to negotiate curves, and exceeding the speed limit; and alcohol impaired driving has been the highest contributing factor in traffic deaths for several years (SDDOT 2007). There are numerous collisions with animals, particularly deer. Although it's rarely fatal, collision with an animal is the first harmful event in 30.8 % of all crashes, possibly resulting in a more severe most-harmful event such as roadway departure. These are all substantial problems that plague rural highways and should be addressed with traditional safety improvements.

Additionally, widely varying conditions in rural states complicate the analysis of safety risk. Crashes in South Dakota are distributed in an extremely geographically imbalanced manner. One-third of the crashes occurred in Minnehaha or Pennington Counties. These two counties represent the most urbanized areas in the state, containing the two largest cities: Sioux Falls (Minnehaha) and Rapid City (Pennington) (South Dakota Department of Public Safety 2008). Nevertheless, the issue of serious injury crashes, fatal or incapacitating injury (K and A) crashes, is more prominent in rural areas. 74.3% of fatal crashes and 32.8% of injury crashes occurred on rural highways and roads. Figure 2.1 shows how the crash data vary across South Dakota. As apparent from the map, crash count and rates vary greatly between rural and more urbanized areas (Minnehaha County in the southeast and Pennington County in the west). Even among the more rural counties, crash rates vary greatly between counties.

A number of safety studies have been conducted to investigate the crash causations and factors contributing to high fatality rates in South Dakota. In the report, *Factors Contributing to South Dakota Crash and Fatality Rates* (Drake et al. 2005), the authors concluded that the primary focus areas for the state are underreporting, rollover crashes, restraint use, alcohol, speeding, and young drivers. The authors also mentioned that there is not a systematic approach to capturing and analyzing all of the crashes occurring on reservation lands. In a separate report, *Identification of Abnormal Accident Patterns At Intersections*, Storsteen produced a total of 14 expected value analysis tables for the various intersection types using a statistically valid sampling method to determine whether a specific intersection has an abnormally high number of crashes (Storsteen 1999). The two representative studies provided valuable recommendations to SDDOT about identifying important crash patterns and safety deficiencies.

Safety stakeholders have been investing considerably to improve safety in all aspects. SD DPS has significantly upgraded the Dakota Accident Reporting System (DARS) throughout the entire reporting process, from field data collection to data presentation. The use of Traffic and Criminal Software (TraCS) has been promoted among law enforcement agencies—including tribal police—through funding, extensive training, and technical support. SD DPS has chosen TraCS as the data entry method of preference inside the agency.

SDDOT has implemented comprehensive procedures to plan, select, and finance safety-oriented projects. Current traffic safety improvement projects are identified through the Roadway Safety Improvement (RSI) process in which SDDOT traffic engineers and local agencies jointly identify safety projects across the state based heavily on crash data. Approximately 150 projects are selected and recommended for the HSIP program (SDDOT 2007). The projects can be categorized into roadway segment treatments and intersection improvements, where crashes within a 100-ft radius of the junction area are considered intersection-related crashes. The projects identified through RSI must be submitted by April 1 for inclusion in the State Transportation Improvement Plan (STIP), a five-year transportation plan.

Other approaches considered at SDDOT include reviewing and analyzing safety information during the design process, an effort led by the Office of Road Design; recommending low-cost safety improvements that require no or limited environmental studies for local roads such as signing and pavement marking; conducting Road Safety Audits (RSA) and Road Safety Audit Reviews (RSAR), and informing travelers of real-time road weather conditions to avoid weather-related crashes through the 511 information system. These safety initiatives are proactive-oriented actions reflecting the great effort made by the state safety programs in order to accommodate the rising needs from rural highway safety and increase the level of engagement with local highway agencies.

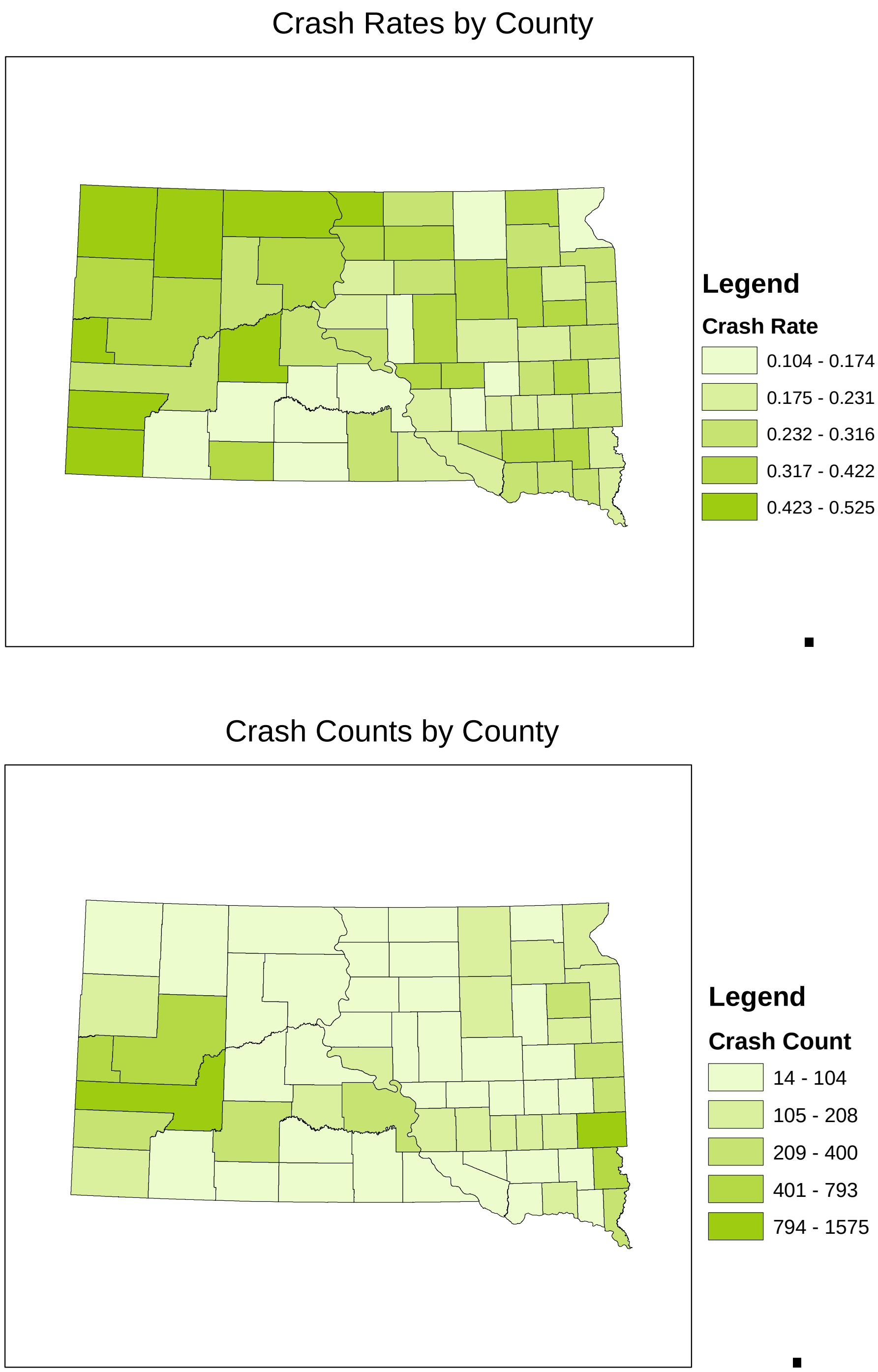


Figure 2.1 South Dakota Crash Map

3. LITERATURE REVIEW

Many methods have been developed and used for identification of high-risk locations or areas over the history of safety analysis. All methods have advantages and disadvantages. Broadly speaking, the two main classes of methods are observational crash analysis and predictive crash analysis.

3.1 Observational Crash Analysis

There are three distinct observational crash analysis methods: crash frequency, crash rate, and crash density. These methods tend to have very simple data requirements and, more importantly, do not require any assumptions. The observed number can be a reflection of the safety performance of a facility if the observation is not fluctuating considerably over a long period of observation. The most popular observational crash analysis statistics are crash counts, crash rates, and crash densities. Some variations or combinations are also available. The crash frequency method analyzes the number of crashes at a location normalized by the time span over which they occurred. It can be used as a preliminary screening tool, to identify locations in need of more detailed analysis. The primary weakness of the crash frequency method is that traffic exposure is not accounted for. When comparing locations with different lengths, crash density is often used. The crash density method analyzes the number of crashes normalized by the segment length. A list of roadway segments can be ranked by the number of crashes, helping to identify potential high-risk locations. This method suffers from the issue of extremely short segments where crash density can be very large as the crash count is divided by the short length of a roadway segment. The crash rate method involves normalizing the crash data by the volume of traffic on a roadway. The crash rate is typically expressed as crashes per million entering vehicles for intersections or crashes per million vehicle miles traveled for segments. The crash rate for an entity can be computed as follows.

Intersections:

$$CR_i = \frac{N*(1{,}000{,}000)}{VT} \tag{1}$$

Where:

CR_i = Crash rate for the intersection (in crashes per million entering vehicles)
N = The number of crashes occurring during the study period
V = Annual average daily entering traffic volume (vehicles/day)
T = The time frame of the study (in days)

Segments:

$$CR_s = \frac{N*(1{,}000{,}000)}{VTL} \tag{2}$$

Where:

CR_s = Crash rate for the segment (in crashes per million vehicle miles traveled)
N = The number of crashes occurring during the study period
V = The AADT of the segment (vehicles/day)
T = The time frame of the study (in days)
L = The length of the segment under consideration (in miles)

The crash rate method has the advantage of taking into account the traffic exposure, a major contributing factor to crashes. In addition, the data needed are typically readily available. While the AADT is needed, most roadways studied will have an AADT measured, or an estimate readily available. Similar to crash density, crash rate measurement is undermined by extremely short roadway segments or segments carrying very low traffic volume.

3.2 Predictive Crash Analysis

Compared with the simple observational crash analysis, regression-based predictive crash analysis produces an expected number of crashes for a site given its characteristics of interest. It is normally based on a distributional form that can be used to describe the probability of a crash occurrence and the assumption for unobserved randomization (unstructured errors). The advantage of using a predictive method for crash analysis is that these methods can separate out variations in observations due to sampling error and identify crucial risk factors that help to predict the outcome levels. As with any parametric-based regression model, selecting an appropriate function form and an error structure will directly affect the accuracy and efficiency of the estimation. Poisson regression is considered one of the most suitable techniques for crash count models because crash data are non-negative and highly right-skewed (positive skew), i.e., the probability of observing a large number of crashes is very low (P. P Jovanis and Chang 1986; Miaou 1993; Miaou and Lum 1993; Washington 2003). If the crash count is considered a random variable following Poisson distribution:

$$y_i|\mu_i \sim Poisson(\mu_i) \;\; i = 1,2,\dots n \tag{3}$$

Where the number of crashes at site i y_i, conditional on mean μ_i , is assumed to follow a Poisson distribution independently over sites.

The general form of the crash model is

$$\mu_i = \left(\frac{AADT \times L \times 365}{1{,}000{,}000}\right)^{\alpha} \exp(\boldsymbol{X_i \beta}) \;\; i = 1,2,\dots n \tag{4}$$

Where:

AADT is the annual average daily traffic for site i
L is the site length
$\boldsymbol{X_i}$ is the vector of variables and $\boldsymbol{\beta_i}$ is the vector of the coefficients for the variables.

In the Highway Safety Manual (HSM), this form is called the Safety Performance Function (SPF). The manual suggests that the general form be implemented using a two-step approach: estimate crash frequency for base conditions and multiple factors by specifications. For a rural two-lane, two-way undivided roadway segment the predictive model is shown in Equation 5.

$$N_{predicted rs} = N_{spf\ rs} \times C_r \times (CMF_{1r} \times CMF_{2r} \times \ldots \times CMF_{12r}) \tag{5}$$

Where:

$N_{predicted\ rs}$ is predicted average crash frequency for an individual roadway segment or a specific year,
$N_{spf\ rs}$ is the predicted average crash frequency for base conditions for an individual roadway segment,
C_r is the calibration factor for roadway segment with a specific type developed for a particular area or jurisdiction, and
CMF_i... are the crash modification factors for rural, two-lane, two-way segments. The 12 factors include lane use, access intensity, roadway geometric characteristics, and others.

The SPF for based conditions is

$$N_{spf\ rs} = AADT \times L \times 365 \times 10^{-6} \times \exp(-0.312) \tag{6}$$

After reviewing the crash prediction methodology provided in the HSM, it was determined that adequate data were not available to implement this methodology, especially the preparation of CMFs. Therefore, for this study, South Dakota crash prediction models were developed in the full form, i.e., including all relevant variables, given in Equation 4.

In practice it is normal that individual locations do not have adequate data for drawing a valid and explicit conclusion because crashes are rare events. In order to obtain a large sample, crash data are often pooled from a wide range of geographic locations and different times. Data collected at the same time and location may exhibit similarities; whereas, data collected at different times and locations may exhibit markedly different characteristics. In statistics, it is called data heterogeneity, meaning the variance of the dependent variable changes in observation. In crash data, it is common to perceive over-dispersed data, i.e., the variance of the dependent variable is higher than expected. The basic assumption of equality of variance and mean in a Poisson model is violated if data present overdispersion. The accuracy and efficiency of model estimates will be compromised. To account for the data overdispersion, the mean expressed was restructured by multiplying an error term to Equation 7 so that

$$\mu_i = \left(\frac{AADT \times L \times 365}{1{,}000{,}000}\right)^{\alpha} \exp(\boldsymbol{X_i \beta}) \exp(\mathrm{e_i}) \quad i = 1,2,\ldots n \tag{7}$$

If $\exp(e_i)$ is assumed to have a gamma distribution with a mean equal to 1 and variance equal to k for all i and independent of all the explanatory variables, the crash count derived from this Poisson-gamma process follows a negative binomial (NB) distribution which can easily handle the crash data over-dispersion. If we use $f(.)$ to represent the function of $\left(\frac{AADT \times L \times 365}{1{,}000{,}000}\right)^{\alpha} \exp(\boldsymbol{X_i \beta})$, the mean and variance of the NB distribution are expressed as $f(.)$ and $f(.)[1+f(.)k]$, with k as the over-dispersion parameter (Caliendo et al. 2007; Lord et al. 2005; Miaou 1993; Shankar et al. 1995; Vogt and Bared 1998; Washington 2003). The NB distribution has a slightly complicated form to estimate the probability of a crash count using the following Equation 8.

$$f(y_i; k, \mu_i) = \frac{\Gamma(y_i + 1/k)}{\Gamma(1/k) y_i!} \left(\frac{1}{\mu_i k + 1}\right)^{1/k} \left(\frac{\mu_i k}{1 + \mu_i k}\right)^{y_i} \tag{8}$$

Regression-based crash analysis is a big step forward from simple observational safety analysis because it predicts the number of crashes given specific values for variables and considers random noise in the crash occurrence from a statistical perspective.

While predictive crash analysis is a powerful tool to explore highway safety, it does require careful use and interpretation of the statistics. It is recommended that the preparation of SPFs be performed by those with proper statistical training. It should be noted that the terms included in SPF do not necessarily indicate a cause and effect relationship. Many variables included are indicators of other factors that are not represented in the data available or not directly measurable. Therefore, the results of the regression should not be interpreted as a list of causational factors, but predictors of crash likelihood.

3.3 Empirical Bayesian (EB) Method

The statistical randomness can be demonstrated by observing the number of crashes occurring at a site for a certain time period, which usually changes from year to year. A closer examination will reveal some trend of this fluctuation. A year with a low number of crashes usually follows a year with a high number of crashes and vice versa. This phenomenon is called *regression-to-the-mean* (*RTM*), when unusually large or small measurements tend to be followed by measurements that are closer to the mean (Bland and Altman 1994). RTM complicates a safety analysis because most "unsafe" locations will, if nothing is done, experience a lower crash rate after time elapses. More importantly, the safety decision is sometimes based on a short-term average that may be different from the actual long-term average (Abbess et al. 1981; Hauer 1997). This tendency frequently misleads the analyst to draw incorrect conclusions as to high risk locations, resulting in applying countermeasures in locations where the crash rate will likely decrease regardless of what countermeasures may be employed. More critically, the areas with potentially high crash risk may be overlooked. Figure 3.1 illustrates a hypothetical example of RTM.

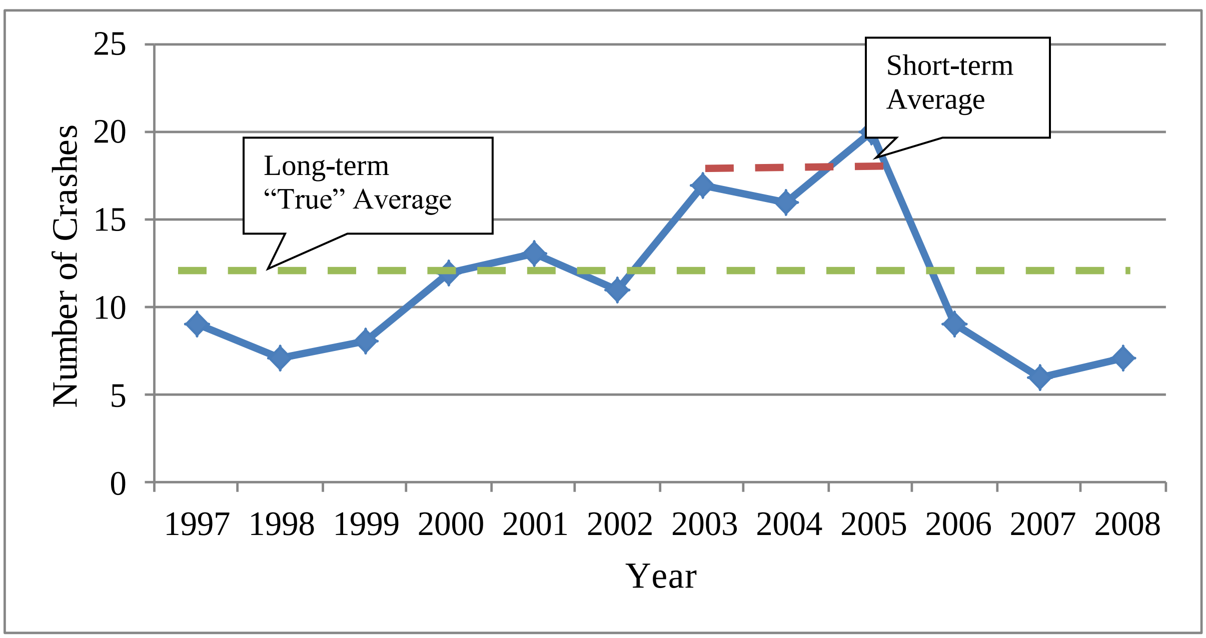


Figure 3.1 Regression to the Mean

Refining the predicted crashes with a site's crash history will overcome the RTM issue and improve the estimation accuracy. The crashes predicted via a SPF yield the same results for the sites having the same values for the characteristics in the model form. These characteristics are by no means an exhaustive list of crash contributing factors. They may be a fraction of all the information yet to be known. Without further knowledge of other crash causational factors, the crash history of a specific site may offer additional clues. Conditional upon its crash history, the estimated number of crashes for a site with known SPF can be simplified in Equation 9 as a weighted average of the actual crashes and the result of a safety performance function. This state-of-the-art method for analyzing crash data is the Empirical Bayes (EB) method (Hauer 1997).

$$E = W \times \mu + (1 - W)N \tag{9}$$

Where:
E is the expected crashes for an entity (the long-term mean of a site),
W is the weight factor that depends on the SPF value and reliability of the crash prediction which is expressed as an over-dispersion parameter k,

$$W = \frac{1}{1 + (\mu Y)k}$$

N is the crash frequency for a site,
μ is the predicted number of crashes/year at the site, the value of a SPF, and
Y is the number of years.

To implement the EB method first a SPF must be determined. After an SPF is determined, the dispersion parameter must be calculated based on the widely accepted negative binomial nature of the SPF. The weight factor is then determined based on the data availability and the quality of the SPF prediction (Powers and Carson 2004). The EB method accounts for regression-to-the-mean by adjusting the actual crash data with an SPF based on the weight. It is generally found to be the most accurate method of hotspot identification (Elvik 2008).

3.4 Existing Safety Analysis Tools

GIS is a powerful tool for the analysis and storage of data. Unlike many traditional methods of data management, it provides both attribute and spatial data. Moreover, GIS allows much advanced analysis to be conducted in addition to the production of maps. Even a new user can produce professional maps using complex symbols representing actual data. While GIS is an effective mapping tool, the real benefit lies in its ability to work with data spatially and based on the associated tabular data (Ormsby 2004).

Recent improvements in computer-based data and GIS techniques make the implementation of a data-driven approach simpler. Traditional methods of managing highway safety programs, however, have not taken advantage of these advancements fully. This is largely due to inaccessible data and difficulty in implementing the methods on a wide scale. Several tools have been developed to facilitate the analysis of safety risk on roadways. Each provides a unique approach for identifying highway safety risks.

HSIS GIS Safety Tools: GIS safety tools provided by the Highway Safety Information System (HSIS), implement several screening techniques to analyze data. The tools include intersection and segment methods, as wells as a sliding-window-based method for analyzing segments. Prototyped by the Federal Highway Administration (FHWA) and implemented in North Carolina, the safety performance measure is primarily based on raw crash count, not crash prediction models (Washington and National Cooperative Highway Research Program; National Research Council (U.S.); American Association of State Highway and Transportation Officials 2006).

PRECIS: A study in Wisconsin was conducted using a sliding window analysis to identify roadway departure crash locations. An algorithm known as PRECIS was developed to aid in the analysis. The analysis was conducted on a continuous corridor in southern Wisconsin. The PRECIS algorithm allowed the production of charts and maps displaying the relative frequency of roadway departure crashes along the highway corridor. These maps can be of great use in identifying the locations of high crash occurrences. The graphical representation can provide insight to potential causes for crashes at high crash locations (Drakopoulos and Ornek 2004).

usRAP (United States Road Assessment Program): a star rating system developed using road and traffic characteristics and does not require crash data. The evaluation is conducted by reviewing approximately 40 key data elements in the areas of highway design and traffic conditions. Crash countermeasures will be recommended and prioritized based on a benefit-cost ratio. The program was initiated by the AAA Foundation for Traffic Safety (AAAFTS) and is being piloted in eight states (Campbell and National Research Council (U.S.); National Cooperative Highway Research Program; American Association of State Highway and Transportation Officials; U.S. 2008).

The Interactive Highway Safety Design Model (ISHDM) provides a suite of modules to analyze various facets of highway safety. This tool provides both proactive and reactive methods to analyze highways. The primary drawback of this tool is the extensive data requirements. Detailed data are needed for a roadway to be analyzed, making the implementation difficult. This tool is primarily for analyzing proposed roadway alignment and improvement projects (Washington and National Cooperative Highway Research Program; National Research Council (U.S.); American Association of State Highway and Transportation Officials 2006).

SafetyAnalyst is by far the most complete, comprehensive, and cutting-edge safety analytical toolbox. It uses extensive roadway data to analyze crash patterns on a section of roadway. It provides a set of computer tools for use by state and local highway agencies for highway safety management. These tools can help improve the programming of site-specific highway safety improvements following the process and procedures in the Highway Safety Manual (HSM). In order to facilitate the implementation for agencies with different levels of data, SafetyAnalyst has the minimum set of data elements required, including crash and traffic data, as well as roadway segment, intersection, and ramp data. However, even the minimum data requirements need considerable data collection effort. In addition, some effort to assemble and format the data will be needed (AASHTO 2010a). According to Minnesota, SafetyAnalyst can be used for improving the identification of black spots, but is not capable of assisting with the identification of candidates for systematic improvements (Campbell and National Research Council (U.S.); National Cooperative Highway Research Program; American Association of State Highway and Transportation Officials; U.S. 2008). The software implementation is progressing in very different stages in several states; some have already been implemented on a statewide basis; some are piloting with a limited number of jurisdictions; and others plan to evaluate the data needs and requirements. The software is now a licensed AASHTOware product and a fee will probably be imposed (AASHTO 2010a).

Although these locally or nationally available safety tools have substantially accelerated the implementation of newly developed methods, systematic safety analysis is either not available or not sufficient to meet South Dakota's needs. A new systematic safety method will be developed using existing data maintained by the State of South Dakota.

4. STUDY DESIGN

After a detailed review of existing South Dakota safety data and the state-of-the-art safety analysis techniques and methodologies, an EB-based sliding window method was selected for this study. Figure 4.1 illustrates the system process and functional requirements.

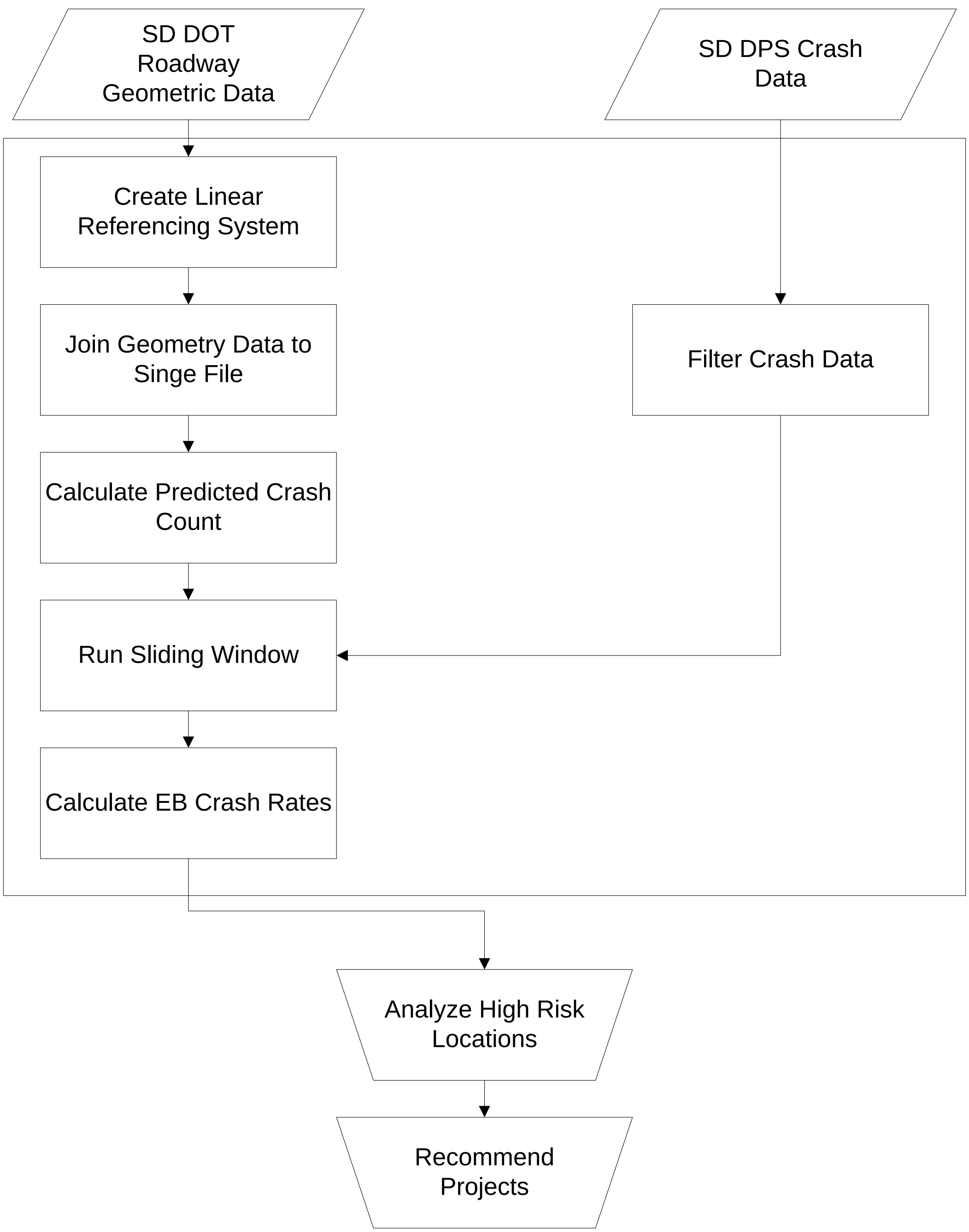


Figure 4.1 Flowchart of Screening Process

Based on the unique challenges faced by rural agencies, the tool includes the following functionalities:

1) A custom integration for South Dakota data and needs
2) Safety performance functions (SPF) for rural highways in South Dakota
3) A data-driven, system-wide method to identify high risk rural roads
4) A presentation of safety outcome in a GIS environment
5) A user friendly interface for practitioners

Developing the SPFs by highway functional classification for South Dakota highways is one of the key requirements of the project. The model accuracy is directly affected by available safety data, which are stored in various locations. A custom integration is required to prepare the data in a single table so that each site will have all the necessary components, including crash data, traffic information, geometric characteristics, and environmental factors. Combined with crash history at each site, the EB crash estimate can be derived. The EB method is considered the basic modeling technique for this study because it not only mitigates the statistical variability caused by a relatively short time period of crash data, but also produces more accurate crash estimation.

From a system perspective, crashes are not isolated chance events but are effects of the causal factors over some spatial extent. A crash occurrence may be affected by the factors upstream or downstream of the location. Therefore, it is important to review crash risk locations on a continuous basis, which presents a complete view of the surrounding environment. A sliding window technique calculates the crash intensity on a dynamic scale irrespective of predetermined segment lengths. With this methodology, a "window" of a designated length incrementally advances along the highway by a shorter incremental distance. As the window progresses along the length of the segment, analyses are conducted for each designated highway length via a window-based EB crash estimate. While the sliding window on top of the highway line features may consist of heterogeneous segments and, therefore, a weighted average, based on the length of a segment included in a window, will be employed to predict crashes (*SPFwin*) for the window section. The EB estimate is then calculated as the weighted average of *SPFwin* and the historical crash frequency within the window boundary, which changes as the window moves along a route. In this way, the crash location, as well as its proximity, can be reviewed in detail. The spatially weighted average *SPFwin* can be calculated with the following formula, Equation 10.

$$SPFwin = SPF_1 \times \frac{l_1}{L} + SPF_2 \times \frac{l_2}{L} \dots \tag{10}$$

Where:
SPF_{win} = SPF value for window,
SPF_1= SPF value for first segment,
l_1= Length of first segment within window,
L = Window Length.

A hypothetical example is illustrated in Figure 4.2. Assume two adjacent roadway segments. One is 2 miles long with two crashes and the other is 3 miles long with three crashes. The segment-based crash count per mile is calculated and displayed with a dashed line; the continuous window-based (one-mile long) crash count is calculated and displayed with a solid line. It is apparent that the segment-based measure fails to identify the crash cluster around the segment boundary while the window-based measure captures the crash risk variation along the roadway with higher resolution. This example demonstrates the strength of the continuous crash risk window-based technique.

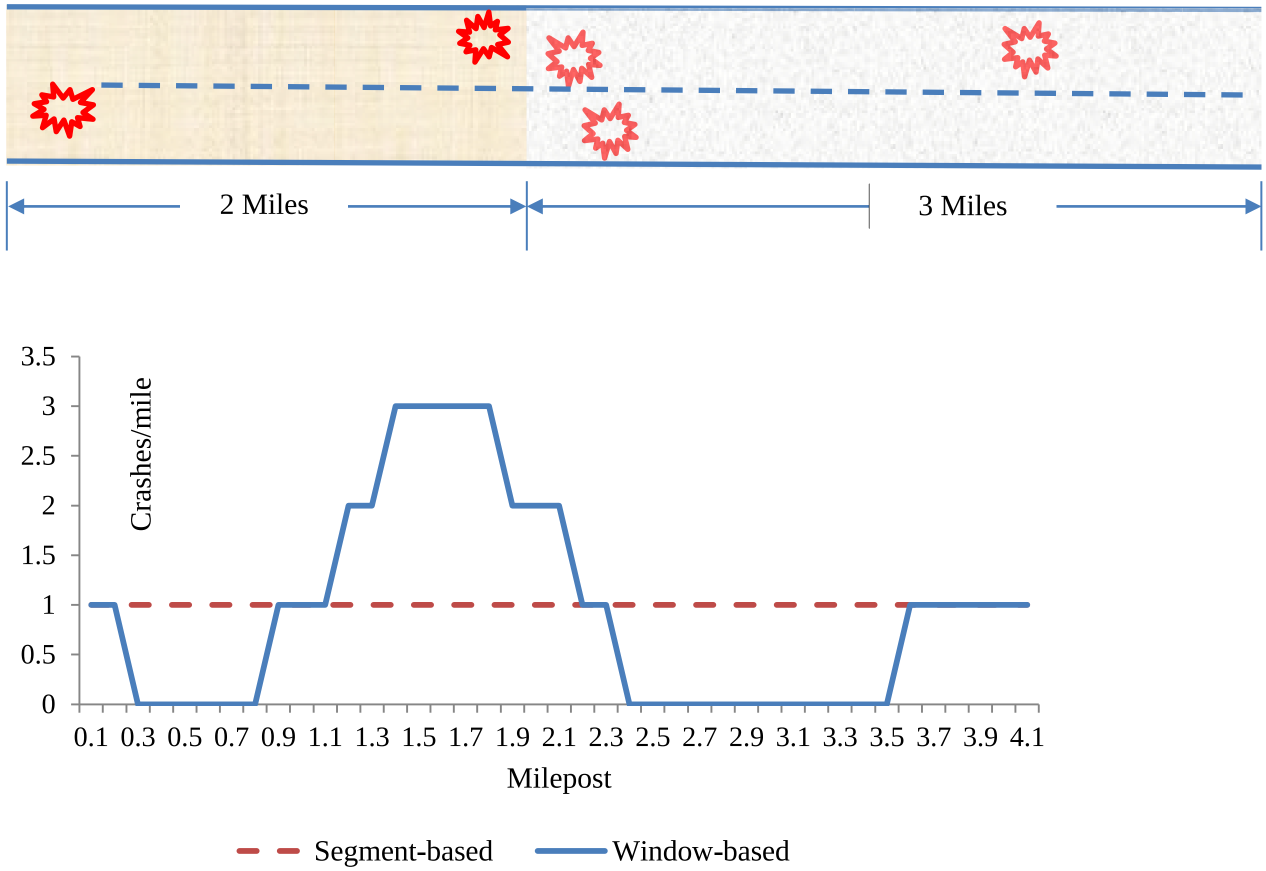


Figure 4.2 Sliding Window

Reviewing crash locations on a continuous basis helps to connect the dots and discover additional information that may be masked by looking at discrete segments. Furthermore, the window size can be adjusted according to the field conditions and needs. Multiple short segments can be contained within a large window in order to smooth out the impacts of short segment length or low volume. On the other hand, very long segments can be scanned via a shorter window size to obtain higher resolution outcome. Finally, the scanning boundaries can be set at any beginning and ending points on a route. Although the sliding window technique holds a great deal of promise, its benefits are somewhat offset by more complex computation and graphical presentation requirements, in comparison with a simple spot map. Fortunately, it is no longer a concern with the fast development of GIS. Subsequent sections present details of data collection and processing, crash prediction modeling, safety metrics development, and the screening tools: *South Dakota GIS Highway Safety Review (GIS-HSR) Tools*.

5. DATA COLLECTION AND PROCESSING

Implementing a data-driven approach presents several challenges, primarily associated with the data availability and quality. The most detailed data available will provide the most accurate safety screening results. Motor vehicle crash data, traffic volumes, and geometric and other roadway attributes were requested for the development of predictive crash models.

5.1 Data Requirements

The various data sources are maintained by the SDDOT and by the South Dakota Department of Public Safety (SDDPS). The data acquired to generate safety performance functions include:

- Traffic
- State Roads
- Vert_Curve
- Horiz_Curve
- Accident

Additional data sources were used for data management, but were not included in the regression analysis. These datasets include intersection location, used to identify intersection associated crashes, and DOT administrative boundaries, which allowed the analysis to be split into sections, improving performance.

Traffic volume, general road conditions, and vertical and horizontal alignment data are stored in the *Traffic, State Roads, Vert_Curve, and Horiz_Curve* tables provided by SDDOT. Specifically, both the *Traffic* and *State Roads* tables are line features shapefiles in which general roadway cross-sectional design features, access information, rumble strip, and ADT are associated with individual roadway segments. Roadway alignment data, including curve speed, degree of curvature, K_value, etc., are stored as point features in a shapefile in *Vert_Curve* and *Horiz_Curve*, respectively. Crash data were provided by SDDPS in a set of shapefiles named "Accident" with one file for each year of crash data. Hence, crash data can be readily related to the South Dakota state trunk network, which facilitates the location-specific safety analysis. The crash data for development of the tool include, but are not limited to, crash location, date, manner of collisions, severity, and relationship to intersection. A five-year crash dataset from the years 2004-2008 was used. A list of selected variables, data sources and descriptions is presented in Table 5.1 and a summary of statistics is displayed in Table 5.2.

Table 5.1 Segment Data Variable Descriptions

Variable	Source	Description
Location Reference		
CITY_NAME	State Roads	Name of city (if in city)
COUNTY_NAM		Name of county segment is within
DATA_CLASS		Used to remove ramps
HIGHWAY		Highway name or number
FROM_MRM		Orignal location system - not valid after processing
FROM_DISPL		
FROM_MILEA		
TO_MRM		
TO_DISPL		
TO_MILEAGE		
ROADNAME		Full Highway Name
RID	Calculated	Unique identifyer for each "route" segment
FMEAS		Starting displacemnt from start of route segment (in miles)
TMEAS		Ending displacement from start of roadway segment (in miles)
Roadway Geometries and Highway Characteristics		
LANES	State Roads	Number of lanes
SURFACE_WI		Width of paved surface (see SD pratices for more detailed information)
SURFACE_TY		Pavment Type
SPEED_LIMI		Posted or statatury speed limit
CURB_GUTTE		Curb and gutter type
PRIMARY_SH		Shoulder pavment type
PRI_LEFT_S		Left shoulder width
PRI_RIGHT_		Right shoulder width
FUNC_CLASS		Functional Classification
URBAN_AREA		Name of urban area (if in urban area)
HIGHWAY_SY		Highway system
RUMBLE_STR		Rumble strip type
RUMBLE_S_1		Rumble strip conditon
MEDIAN_TYP		Type of median present (if present)
MEDIAN_WID		Median width
PREF_TRUCK		Truck route
FUNDING_CA		Funding Category (used to determine municipal)
SIM_SURFAC	Calculated	Simplified Surface Type (BITUMINOUS, CONCRETE, or AGGREGATE)
H_Spd_Ave	Horizonatl Curve	Average horizonal curve speed
H_Count		Horizonal Curve Count
V_Sag_Ct	Vertical Curve	Sag Vertical Curve Count
V_Crest_Ct		Crest Vertical Curve Count
V_sp_ave		Average Vertical Curve Speed
V_curve_de	Calculated	Vertical Curve Density

Traffic and Access Control		
ADT	Traffic	Average Dailty Traffic
TRUCK_ADT		Average Daily truck Traffic
ACCESS_CON	State Roads	Is access control present (are ramps present)
VMT	Calculated	Yearly Vehicle Miles Travelled
STUDY_VMT	Calculated	Study Vehicle Miles Travelled
Crash Data		
Total	Accident	Total Study Crash Count
code_1		Number of crashes with severity "1" - Fatal
code_2		Number of crashes with severity "2" - A Severity
code_3		Number of crashes with severity "3" - B Severity
code_4		Number of crashes with severity "4" - C Severity
code_5		Number of crashes with severity "5" - PDO
code_94		Number of crashes with severity "94" - Animal

Table 5.2 Summary Statistics for Variables

Paramater	Major Collector			Minor Arterial			Major Arterial - Interstate			Major Arterial - Other		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
VMT	7.2E+01	6.5E+04	1.3E+06	1.0E+02	3.2E+04	5.8E+05	1.5E+02	5.7E+04	4.7E+05	8.0E+03	5.9E+04	9.9E+05
Lane Width (ft)	10	12	13	6	12	14	12	12	20	6	13	45
Speed Limit (mph)	20	52	65	20	54	65	65	75	70	20	59	70
Shoulder Width (ft)	0.00	3.43	11.00	0.00	4.00	13.00	0.00	7.00	10.00	0.00	5.90	21.00
Percent Trucks	0.014	0.158	0.357	0.013	0.127	0.388	0.020	0.241	0.333	0.015	0.146	0.405
Vertical Curve Density (per mile)	0.00	5.75	29.60	0.00	5.76	71.77	0.00	4.60	36.99	0.00	6.80	56.80
Is Municipal	20.5%			16.0%			0.0%			10.6%		
Has Median	6.1%			11.4%			100.0%			55.7%		
Median Width (ft)	0	2	27	0	11	300	4	27	75	0	17	25
Has Rumble Strip	0.0%			9.1%			61.8%			31.9%		
Has Curb	9.8%			10.6%			0.6%			11.0%		

5.2 Data Processing

Although all the data are in a GIS format, the resolutions, referencing systems, and temporal or spatial gaps between distinct data sources are different, creating hurdles to integrating the data for use in the screening tool. To solve the data issues, the first step was to develop a method to aggregate the data into a single file. ESRI's ArcGIS geo-processing tools were vital in processing the data (ESRI 2010). Data were presented in either a point or a line shapefile and each shapefile type posed a unique challenge.

Curve information is presented as points, located at the point of intersection (PI) for each curve. Frequently, points are not located exactly on the roadway segment. To resolve this issue a spatial join was used. A buffer of 150 ft was applied to ensure points were appropriately associated. The buffer distance of 150 ft was determined arbitrarily, attempting to maximize the number of points properly associated while avoiding false associations.

Both traffic and basic segment data are stored as linear shapefiles. The geometries of these files do not match, complicating the integration of this data. This was resolved by using the linear referencing tools in ArcGIS. A multistep process, as shown in Figure 5.1, was used to aggregate the segment data. A linear referencing route was created using the base data, and an event table was then created for each—the base data and traffic data. The overlay tool could then be applied to create a table containing new segments broken at any endpoint for either shapefile. When the overlay table is plotted as an event layer, the final geometry is plotted along the "route" created.

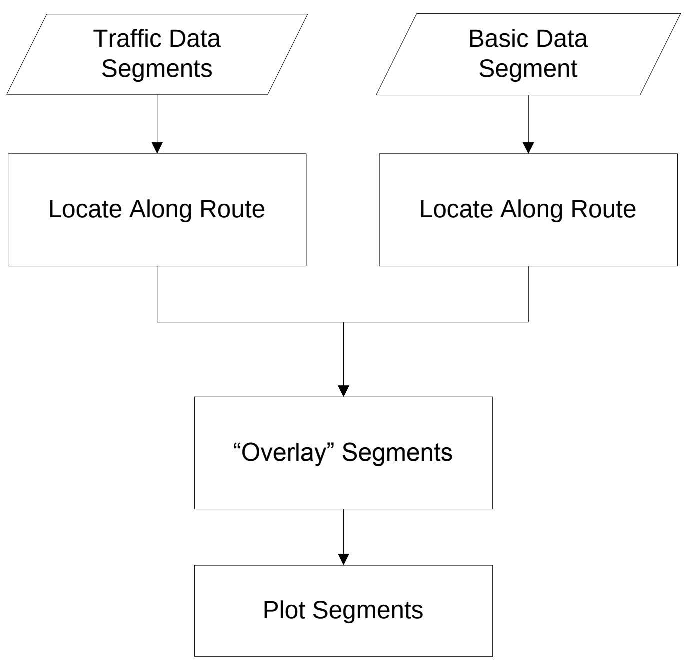


Figure 5.1 Flowchart of "Segment Join" Process

5.3 Exploratory Data Analysis

Data elements are often correlated because of the highway design requirements for each functional class. Including correlated variables in the model may cause a multicollinearity problem that reduces the model efficiency, prediction accuracy, and model stability. Inefficiency means the model can be outperformed by other models with better specifications. Inaccuracy means the variable selection may be biased. In fact, the variables will become less statistically significant because of the high standard errors of the estimated parameters of the collinear variables. Estimated parameters may have large sampling errors when predictor variables are highly correlated, resulting in a wide change in values and signs of estimated parameters from one sample to the next. Furthermore, the standard interpretation of estimated parameters does not apply because it is very difficult to separate correlated variables (Washington 2003).

To mitigate the multicollinearity impact, a correlation study was conducted to identify dependent variables. Variables showing strong dependence was removed or adjusted. Table 5.3 shows the correlation matrix calculated for rural major collectors. Based on this correlation matrix, several variables can be removed. For example, only either PRI_LEFT_S or PRI_RIGHT_ will be included in the final model. In the preliminary models, the following variables were included: logVMT, laneW, SIM_SURFAC, SPEED_LIMI, CURB_GUTTE, PRIMARY_SH, PRI_LEFT_S, MEDIAN_WID, FUNDING_CA, ptrucks, H_Spd_Ave, H_Count, sag_den, crest_den and V_sp_ave. The correlation matrix was repeated for rural minor arterials, rural principal arterial-Interstate, and rural principal arterial-others highway classes for screening the variables used for developing crash prediction models. Please refer to Tables 5.4─ 5.6 for more information.

Table 5.3 Correlation Matrix Between Variables for Rural Major Collectors

	Total	log VMT	LANES	SPEED_LIMI	PRI_LEFT_S	PRI_RIGHT	MEDIAN_WID	H_Spd_Ave	H_Count	V_Sag_Ct	V_Crest_Ct	V_sp_ave	V_curve_de	laneW	ptrucks	sag_den	crest_den
Total	1	0.488	-0.023	0.162	0.315	0.315	-0.068	0.327	0.293	0.499	0.439	0.175	-0.078	-0.106	0.190	-0.073	-0.056
logVMT	0.488	1	-0.029	0.247	0.338	0.338	0.023	0.478	0.434	0.486	0.503	0.422	-0.028	-0.202	0.426	-0.052	0.005
LANES	-0.023	-0.029	1	-0.137	-0.031	-0.031	-0.014	0.031	0.023	-0.034	-0.026	-0.045	0.090	0.067	-0.076	0.012	0.136
SPEED_LIMI	0.162	0.247	-0.137	1	0.146	0.143	-0.526	0.308	0.280	0.271	0.290	0.182	-0.282	-0.371	0.181	-0.245	-0.220
PRI_LEFT_S	0.315	0.338	-0.031	0.146	1	0.962	-0.093	0.064	0.025	0.063	0.033	0.156	0.205	-0.163	0.313	0.187	0.151
PRI_RIGHT_	0.315	0.338	-0.031	0.143	0.962	1	-0.093	0.063	0.023	0.063	0.033	0.155	0.206	-0.163	0.313	0.188	0.151
MEDIAN_WID	-0.068	0.023	-0.014	-0.526	-0.093	-0.093	1	-0.175	-0.180	-0.095	-0.096	-0.006	0.239	0.197	0.012	0.214	0.180
H_Spd_Ave	0.327	0.478	0.031	0.308	0.064	0.063	-0.175	1	0.972	0.390	0.406	0.415	-0.186	-0.050	0.110	-0.190	-0.117
H_Count	0.293	0.434	0.023	0.280	0.025	0.023	-0.180	0.972	1	0.395	0.414	0.408	-0.165	-0.038	0.052	-0.167	-0.104
V_Sag_Ct	0.499	0.486	-0.034	0.271	0.063	0.063	-0.095	0.390	0.395	1	0.960	0.289	-0.041	-0.135	0.064	-0.009	-0.058
V_Crest_Ct	0.439	0.503	-0.026	0.290	0.033	0.033	-0.096	0.406	0.414	0.960	1	0.297	-0.033	-0.133	0.088	-0.070	0.015
V_sp_ave	0.175	0.422	-0.045	0.182	0.156	0.155	-0.006	0.415	0.408	0.289	0.297	1	0.382	-0.082	0.149	0.343	0.285
V_curve_de	-0.078	-0.028	0.090	-0.282	0.205	0.206	0.239	-0.186	-0.165	-0.041	-0.033	0.382	1	0.048	0.003	0.822	0.824
laneW	-0.106	-0.202	0.067	-0.371	-0.163	-0.163	0.197	-0.050	-0.038	-0.135	-0.133	-0.082	0.048	1	-0.115	-0.006	0.084
ptrucks	0.190	0.426	-0.076	0.181	0.313	0.313	0.012	0.110	0.052	0.064	0.088	0.149	0.003	-0.115	1	-0.028	0.032
sag_den	-0.073	-0.052	0.012	-0.245	0.187	0.188	0.214	-0.190	-0.167	-0.009	-0.070	0.343	0.822	-0.006	-0.028	1	0.355
crest_den	-0.056	0.005	0.136	-0.220	0.151	0.151	0.180	-0.117	-0.104	-0.058	0.015	0.285	0.824	0.084	0.032	0.355	1

Note: Darker green values indicate a strong positive correlation, yellow values indicate a weak correlation, and darker red values indicate a strong negative correlation

Table 5.4 Correlation Matrix Between Variables for Rural Minor Arterials

	Total	log VMT	LANES	SPEED_ LIMI	PRI_ LEFT_ S	PRI_ RIGHT	MEDIAN_ WID	H_ Spd_ Ave	H_ Count	V_Sag_ Ct	V_Crest_ Ct	V_sp_ ave	V_curve_ de	laneW	ptrucks	sag_ den	crest_ den
Total	1	0.513	-0.024	0.093	-0.095	-0.098	-0.066	0.175	0.191	0.452	0.445	0.086	-0.108	-0.117	-0.048	-0.089	-0.080
logVMT	0.513	1	-0.021	0.274	0.077	0.082	0.032	0.329	0.314	0.609	0.612	0.260	-0.220	-0.187	0.039	-0.187	-0.156
LANES	-0.024	-0.021	1	-0.169	-0.069	-0.128	-0.286	-0.062	-0.057	-0.030	-0.032	-0.007	0.093	-0.402	-0.087	0.054	0.094
SPEED_LIMI	0.093	0.274	-0.169	1	0.219	0.243	0.058	0.218	0.167	0.270	0.276	0.240	-0.318	-0.241	0.295	-0.246	-0.254
PRI_LEFT_S	-0.095	0.077	-0.069	0.219	1	0.869	0.218	0.049	0.031	-0.129	-0.115	0.010	-0.102	-0.240	0.083	-0.108	-0.048
PRI_RIGHT_	-0.098	0.082	-0.128	0.243	0.869	1	0.334	0.064	0.045	-0.131	-0.119	-0.001	-0.109	-0.165	0.093	-0.099	-0.071
MEDIAN_WID	-0.066	0.032	-0.286	0.058	0.218	0.334	1	0.076	0.076	-0.096	-0.110	-0.124	-0.009	0.252	0.063	0.075	-0.101
H_Spd_Ave	0.175	0.329	-0.062	0.218	0.049	0.064	0.076	1	0.978	0.271	0.272	0.259	-0.059	-0.051	0.076	-0.016	-0.081
H_Count	0.191	0.314	-0.057	0.167	0.031	0.045	0.076	0.978	1	0.271	0.271	0.244	-0.041	-0.054	0.044	-0.001	-0.068
V_Sag_Ct	0.452	0.609	-0.030	0.270	-0.129	-0.131	-0.096	0.271	0.271	1	0.970	0.270	-0.030	-0.153	0.107	-0.005	-0.045
V_Crest_Ct	0.445	0.612	-0.032	0.276	-0.115	-0.119	-0.110	0.272	0.271	0.970	1	0.266	-0.039	-0.160	0.107	-0.077	0.024
V_sp_ave	0.086	0.260	-0.007	0.240	0.010	-0.001	-0.124	0.259	0.244	0.270	0.266	1	0.325	-0.071	0.085	0.282	0.224
V_curve_de	-0.108	-0.220	0.093	-0.318	-0.102	-0.109	-0.009	-0.059	-0.041	-0.030	-0.039	0.325	1	0.205	-0.077	0.817	0.746
laneW	-0.117	-0.187	-0.402	-0.241	-0.240	-0.165	0.252	-0.051	-0.054	-0.153	-0.160	-0.071	0.205	1	-0.022	0.164	0.158
ptrucks	-0.048	0.039	-0.087	0.295	0.083	0.093	0.063	0.076	0.044	0.107	0.107	0.085	-0.077	-0.022	1	-0.068	-0.052
sag_den	-0.089	-0.187	0.054	-0.246	-0.108	-0.099	0.075	-0.016	-0.001	-0.005	-0.077	0.282	0.817	0.164	-0.068	1	0.226
crest_den	-0.080	-0.156	0.094	-0.254	-0.048	-0.071	-0.101	-0.081	-0.068	-0.045	0.024	0.224	0.746	0.158	-0.052	0.226	1

Note: Darker green values indicate a strong positive correlation, yellow values indicate a weak correlation, and darker red values indicate a strong negative correlation

Table 5.5 Correlation Matrix Between Variables for Rural Principal Arterials – Interstate

	Total	log VMT	LANES	SPEED_ LIMI	PRI_ LEFT_ S	PRI_ RIGHT	MEDIA N_ WID	H_ Spd_ Ave	H_ Count	V_Sag_ Ct	V_Crest _ Ct	V_sp_ ave	V_curv e_ de	laneW	ptrucks	sag_ den	crest_ den
Total	1	0.739	-0.059	0.036	-0.030	0.006	-0.058	0.354	0.352	0.766	0.732	0.270	-0.043	0.126	-0.201	0.018	-0.065
logVMT	0.739	1	-0.080	0.023	-0.004	0.017	0.041	0.453	0.450	0.735	0.742	0.496	-0.128	0.061	-0.043	-0.034	-0.122
LANES	-0.059	-0.080	1	-0.078	-0.110	0.079	-0.101	-0.051	-0.051	-0.087	-0.070	-0.147	-0.057	-0.033	-0.038	-0.089	0.008
SPEED_LIMI	0.036	0.023	-0.078	1	-0.011	0.035	-0.278	0.038	0.038	0.055	0.060	0.021	0.019	-0.065	0.156	0.026	0.000
PRI_LEFT_S	-0.030	-0.004	-0.110	-0.011	1	-0.837	-0.042	-0.003	-0.003	-0.012	-0.010	-0.021	-0.003	-0.169	0.016	-0.010	0.005
PRI_RIGHT_	0.006	0.017	0.079	0.035	-0.837	1	0.043	0.021	0.021	0.014	0.017	0.003	0.017	-0.191	0.004	0.004	0.017
MEDIAN_WID	-0.058	0.041	-0.101	-0.278	-0.042	0.043	1	0.111	0.111	0.025	0.023	0.096	-0.069	-0.034	0.095	-0.024	-0.061
H_Spd_Ave	0.354	0.453	-0.051	0.038	-0.003	0.021	0.111	1	1.000	0.385	0.401	0.242	-0.060	-0.052	-0.071	-0.013	-0.061
H_Count	0.352	0.450	-0.051	0.038	-0.003	0.021	0.111	1.000	1	0.384	0.399	0.242	-0.060	-0.052	-0.071	-0.012	-0.061
V_Sag_Ct	0.766	0.735	-0.087	0.055	-0.012	0.014	0.025	0.385	0.384	1	0.839	0.320	-0.005	-0.017	0.001	0.152	-0.133
V_Crest_Ct	0.732	0.742	-0.070	0.060	-0.010	0.017	0.023	0.401	0.399	0.839	1	0.328	-0.001	-0.005	0.003	-0.095	0.078
V_sp_ave	0.270	0.496	-0.147	0.021	-0.021	0.003	0.096	0.242	0.242	0.320	0.328	1	0.476	0.060	-0.029	0.327	0.286
V_curve_de	-0.043	-0.128	-0.057	0.019	-0.003	0.017	-0.069	-0.060	-0.060	-0.005	-0.001	0.476	1	0.072	-0.156	0.550	0.716
laneW	0.126	0.061	-0.033	-0.065	-0.169	-0.191	-0.034	-0.052	-0.052	-0.017	-0.005	0.060	0.072	1	-0.165	0.017	0.070
ptrucks	-0.201	-0.043	-0.038	0.156	0.016	0.004	0.095	-0.071	-0.071	0.001	0.003	-0.029	-0.156	-0.165	1	-0.069	-0.126
sag_den	0.018	-0.034	-0.089	0.026	-0.010	0.004	-0.024	-0.013	-0.012	0.152	-0.095	0.327	0.550	0.017	-0.069	1	-0.189
crest_den	-0.065	-0.122	0.008	0.000	0.005	0.017	-0.061	-0.061	-0.061	-0.133	0.078	0.286	0.716	0.070	-0.126	-0.189	1

Note: Darker green values indicate a strong positive correlation, yellow values indicate a weak correlation, and darker red values indicate a strong negative correlation

Table 5.6 Correlation Matrix Between Variables for Rural Principal Arterials – Other

	Total	log VMT	LANES	SPEED_ LIMI	PRI_ LEFT_ S	PRI_ RIGHT	MEDIAN_ WID	H_ Spd_ Ave	H_ Count	V_Sag_ Ct	V_Crest_ Ct	V_sp_ ave	V_curve_ de	laneW	sag_ den	crest_ den
Total	1	0.732	-0.041	0.126	0.009	0.009	0.066	0.194	0.200	0.562	0.566	0.161	-0.121	-0.116	-0.090	-0.091
logVMT	0.732	1	-0.086	0.184	0.094	0.076	0.071	0.249	0.231	0.730	0.747	0.214	-0.133	-0.144	-0.112	-0.086
LANES	-0.041	-0.086	1	-0.407	-0.084	-0.088	-0.163	-0.144	-0.126	-0.099	-0.104	-0.063	0.041	-0.134	0.046	0.015
SURFACE_WI	-0.107	-0.163	0.708	-0.440	-0.272	-0.268	-0.204	-0.165	-0.144	-0.130	-0.132	-0.117	0.082	0.589	0.085	0.037
SPEED_LIMI	0.126	0.184	-0.407	1	0.210	0.216	0.129	0.210	0.149	0.178	0.181	0.131	-0.228	-0.170	-0.203	-0.138
PRI_LEFT_S	0.009	0.094	-0.084	0.210	1	0.407	-0.052	0.049	0.028	-0.006	-0.004	0.010	-0.169	-0.278	-0.113	-0.139
PRI_RIGHT_	0.009	0.076	-0.088	0.216	0.407	1	-0.038	0.033	0.010	-0.007	-0.004	-0.011	-0.148	-0.265	-0.105	-0.116
MEDIAN_WID	0.066	0.071	-0.163	0.129	-0.052	-0.038	1	0.081	0.074	-0.001	0.008	0.014	0.095	-0.102	-0.008	0.149
H_Spd_Ave	0.194	0.249	-0.144	0.210	0.049	0.033	0.081	1	0.983	0.232	0.239	0.206	-0.040	-0.073	-0.040	-0.021
H_Count	0.200	0.231	-0.126	0.149	0.028	0.010	0.074	0.983	1	0.214	0.221	0.194	-0.023	-0.063	-0.024	-0.011
V_Sag_Ct	0.562	0.730	-0.099	0.178	-0.006	-0.007	-0.001	0.232	0.214	1	0.947	0.232	0.009	-0.079	0.057	-0.042
V_Crest_Ct	0.566	0.747	-0.104	0.181	-0.004	-0.004	0.008	0.239	0.221	0.947	1	0.231	0.010	-0.075	-0.061	0.075
V_sp_ave	0.161	0.214	-0.063	0.131	0.010	-0.011	0.014	0.206	0.194	0.232	0.231	1	0.374	-0.103	0.298	0.261
V_curve_de	-0.121	-0.133	0.041	-0.228	-0.169	-0.148	0.095	-0.040	-0.023	0.009	0.010	0.374	1	0.078	0.744	0.751
laneW	-0.116	-0.144	-0.134	-0.170	-0.278	-0.265	-0.102	-0.073	-0.063	-0.079	-0.075	-0.103	0.078	1	0.077	0.040
sag_den	-0.090	-0.112	0.046	-0.203	-0.113	-0.105	-0.008	-0.040	-0.024	0.057	-0.061	0.298	0.744	0.077	1	0.117
crest_den	-0.091	-0.086	0.015	-0.138	-0.139	-0.116	0.149	-0.021	-0.011	-0.042	0.075	0.261	0.751	0.040	0.117	1

Note: Darker green values indicate a strong positive correlation, yellow values indicate a weak correlation, and darker red values indicate a strong negative correlation

6. METHODOLOGIES

Multiple methodologies were employed in this study. Crash prediction techniques were implemented to generate safety performance functions. Additionally, a procedure was developed to perform the sliding window analysis based on the state highway system geometry. Lastly, various safety metrics which led to informed safety improvements decisions.

6.1 Prediction Models

Traffic safety research studies show that the causal factors of a crash can be identified through a well designed regression model. In other words, the inherent safety performance of a roadway segment or an intersection can be measured through the attributes correlated to the outcome, i.e., crash frequency and consequence. SPFs developed through advanced statistical modeling can identify the reliable correlation between crashes and roadway conditions and their effects from a large sample of entities. SPFs can quantitatively describe the relationship between the number of crashes per year (and per mile if a road segment) and a measure of traffic exposure. Using a weight factor, the long-term mean for site safety performance can be obtained using the EB method as Equation 9.

$$E = W \times \mu + (1 - W)N$$

Where:

E is the expected crashes for an entity (the long-term mean of a site),
W is the weight factor that depends on the SPF value and reliability of the crash prediction which is expressed as an over-dispersion parameter k,
μ is the predicted number of crashes/year at the site, the value of a SPF expressed as:

$$\mu = (traffic\ exposure)^{\alpha} e^{(\beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \cdots)}$$

Where:

Traffic exposure is represented by million vehicle miles traveled (MVMT),
$X_1, X_2, X_3\ldots$ are crash contributing factors,
$\beta_1, \beta_2, \beta_3\ldots$ are unknown coefficients to be estimated,
N is the number of crashes.

The state-of-the-practice distribution considered for modeling crashes is Poisson-gamma (or negative binomial). Poisson-gamma models can easily handle the crash data over-dispersion through its variance expression $V(\mu)=\mu+k\mu^2$, with k as the over-dispersion parameter. In this study, we let N_i denote the number of crashes at site *i*. N_i conditional on its mean μ_i is assumed to follow a Poisson distribution independently over sites (Equation 3).

$N_i|\mu_i \sim Poisson(\mu_i) \;\; i = 1,2,\dots n$

The log link function is defined the same as Equation 4.

$\mu_i = (MVMT)^{\alpha} e^{\boldsymbol{X_i\beta}+\boldsymbol{Z_i\gamma}+\epsilon_i} \;\; i = 1,2,\dots n$

Where:

MVMT: million vehicle miles traveled,
$\boldsymbol{X_i}$: the vector of continuous variables,
$\boldsymbol{Z_i}$: the vector of categorical variables,
α, $\boldsymbol{\beta}$, γ,: regression coefficients (bold represents vector), and
ε_i: is an unstructured random effect independent of $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$.

The Poisson-gamma model is specified by assuming that exp (ε_i) follows a gamma distribution. The unknown parameters and overdispersion factor k can be estimated via statistical software packages such as GENMOD in SAS (SAS 2010) or glm.nb in R (R-Project 2010). Note that the presentation of the overdispersion factor may vary by statistical software package. In SAS, it is overdispersion factor k while in R, it is the inversed overdispersion factor ϕ which is $1/k$. Caution is advised when using different statistical software to compute the overdispersion factor. Usually, the user manual of the software provides relevant information and formulas.

Crash prediction models were generated using GLM regression in R. Variables were selected for each functional classification where a correlation exists with the crash count. Stepwise variable selection was used to keep the statistically significant variable at a level of at least 5% in the model ($p<0.05$). The models were selected based on the smaller Akaike Information Criteria (AIC) which measures the goodness-of-fit and the lower dispersion parameter values. A likelihood ratio test was also preformed after removing a variable to verify a statistical improvement actually existed. Variables were also examined through engineering judgment before any model was accepted. The results of the regression can be seen in Tables 6.1─6.4.

Table 6.1 Regression Results for Rural Major Collector

Variable[1]	Estimate	Std. Error	z value	Pr(>|z|)[2]	
(Intercept)	0.578	0.884	0.654	0.513	
logVMT	0.689	0.050	13.786	< 2e-16	***
laneW	-0.016	0.033	-0.473	0.636	
SPEED_LIMI	0.048	0.011	4.510	0.000	***
PRI_LEFT_S	0.016	0.025	0.631	0.528	
IS_MUNICPATRUE	-0.833	0.334	-2.493	0.013	*
ptrucks	-2.123	0.871	-2.438	0.015	*
V_curve_de	0.007	0.019	0.379	0.705	

1. Dispersion parameter for Negative Binomial is 0.7527
2. Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 1077.35 on 590 degrees of freedom
Residual deviance: 447.65 on 583 degrees of freedom
2 x log-likelihood: -1446.595 AIC: 1464.6

Table 6.2 Regression Results for Rural Minor Arterial

Variable[1]	Estimate	Std. Error	z value	Pr(>|z|)[2]	
(Intercept)	5.640	0.276	20.449	< 2e-16	***
logVMT	0.869	0.027	31.997	< 2e-16	***
laneW	-0.027	0.012	-2.332	0.020	*
V_curve_de	-0.015	0.009	-1.704	0.088	.
PRI_RIGHT_	-0.086	0.013	-6.840	0.000	***
SPEED_LIMI	-0.021	0.003	-6.311	0.000	***
ptrucks	-1.404	0.460	-3.054	0.002	**
HAS_MEDIANTRUE	-0.595	0.131	-4.547	0.000	***
HAS_RUMBLETRUE	-0.296	0.119	-2.481	0.013	*
IS_MUNICPATRUE	-1.235	0.149	-8.275	< 2e-16	***

1. Dispersion parameter for Negative Binomial is 2.0662
2. Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 3776.2 on 1846 degrees of freedom
Residual deviance: 1632.8 on 1837 degrees of freedom
2 x log-likelihood: -4345.106 AIC: 4367.1

Table 6.3 Regression Results for Rural Principal Arterial- Interstate

Variable[1]	Estimate	Std. Error	z value	Pr(>|z|)[2]	
(Intercept)	4.955	0.291	17.015	< 2e-16	***
logVMT	0.888	0.029	30.863	< 2e-16	***
ptrucks	-1.350	0.480	-2.812	0.005	**
V_curve_de	-0.029	0.009	-3.273	0.001	**
laneW	-0.038	0.012	-3.130	0.002	**
MEDIAN_WID	-0.003	0.001	-3.191	0.001	**
SPEED_LIMI	-0.013	0.003	-4.082	0.000	***
HAS_RUMBLETRUE	-0.430	0.124	-3.472	0.001	***

1. Dispersion parameter for Negative Binomial is 1.6231
2. Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 3413.0 on 1846 degrees of freedom
Residual deviance: 1654.1 on 1839 degrees of freedom
2 x log-likelihood: -4493.872 AIC: 4511.9

Table 6.4 Regression results for Rural Principal Arterial- Other

Variable[1]	Estimate	Std. Error	z value	Pr(>|z|)[2]	
(Intercept)	0.390	0.894	0.436	0.663	
logVMT	0.705	0.048	14.676	< 2e-16	***
SPEED_LIMI	0.057	0.011	4.972	0.000	***
laneW	-0.032	0.034	-0.935	0.350	
ptrucks	-1.833	0.856	-2.141	0.032	*
HAS_CURBTRUE	1.087	0.524	2.077	0.038	*
IS_MUNICPATRUE	-0.864	0.318	-2.720	0.007	**

1. Dispersion parameter for Negative Binomial is 0.7464
2. Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 1072.22 on 590 degrees of freedom
Residual deviance: 442.15 on 584 degrees of freedom
2 x log-likelihood: -1442.7670 AIC: 1458.8

6.2 Continuous Sliding window

A continuous sliding window algorithm was developed for this study. The algorithm was developed to run within ESRI’s ArcGIS software. The basic sliding window algorithm was developed to be a completely flexible tool for analysis of any linear feature. Tools were generated for both points and segments located along a route.

A continuous sliding window algorithm was implemented to provide screening along the route. The algorithm screens highways with a window of 1 mile. Crashes occurring within each window are counted and categorized by crash severity. The window is then shifted 1/10th mile and the crashes in the new window are counted as illustrated in Figure 6.1.

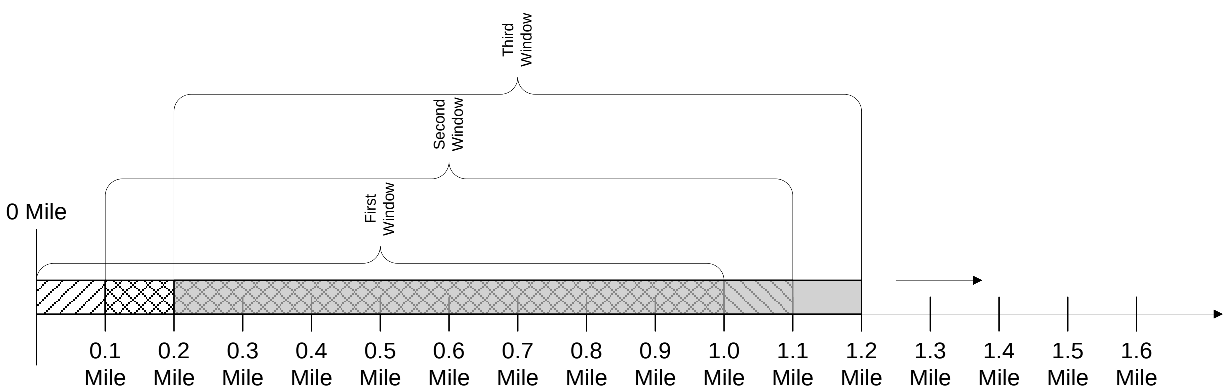


Figure 6.1 Sliding Window Visualized

The first sliding window counts the number of crashes in each category (such as severity) for each window. The categories are generated based on the uniquely coded values in a specified field. After the initial sliding window has been applied to the crash data, segment-based data are linked to implement the EB method to determine an estimate of the true expected number of crashes. A second python script tool was developed to add a weighted average value stored in the underlying segments. A weighted average based on length within the window was chosen to provide proportional weighting based on the exposure of each expected value within the window being considered. The script was used to add the SPF predicted number of crashes and calculate the EB value.

6.3 Safety Metrics

Three classes of performance measures are proposed in Table 6.5 for identifying high-risk locations. The raw crash rate is included to maintain consistency with existing procedures. The EB method was used as a more accurate method. The excessive crash count is provided as a supplemental risk identification factor. Each measure can be calculated either on a segment (predetermined and homogeneous segment) basis or a window (sliding window of one-mile long) basis.

The raw crash rate for an individual window is the first performance measure. Crash rates are a historical performance measure. Those locations with high crash rates can be identified as “high-risk.” The simplistic crash count (or rate) is included because it is a traditional method used by many agencies.

Another performance measure is the EB crash count. Locations with high EB crash counts or crash rates can be targeted for improvement. The EB crash rates should indicate locations with high risk based on both history and predicted crash conditions. The performance measure is applied to both individual windows and predefined segments. The third performance measure is the excessive crash count. All the safety performance measures based on two different roadway sections are visualized in Figure 6.2. Note that Figure 6.2a is for segment-based safety metrics and Figure 6.2b is for window-based safety metrics. The number of crashes occurring beyond the predicted crashes can be an indication of a unique situation requiring further investigation. The difference between the actual crash count and the SPF is an indication of the unexplained risk at a location. However, it does not provide an indication of how many crashes actually occur on a roadway segment. Therefore, it is not a substitute for the EB, but a supplement to encourage more effective use of highway funding. Using these methods, lists of high-risk locations can be generated. Inclusion on multiple lists indicates a greater confidence that a location is high risk. These locations can then be slated for further engineering studies or safety improvements.

Table 6.5 Safety Metrics

Metric	Segment	Window
Raw Crashes	Raw crash count normalized by traffic exposure	Crash count normalized by spatially weighted average of ADT
EB Crashes	Calculated based on segment crash count and SPF value	Calculated based on crash count within window and spatially weighted average of SPF value
Excess Crashes	Raw difference between crash count along segment and predicted crashes along segment. This metric should be multiplied by the segment length	Difference between actual crashes within the window and spatially weighted average of predicted crash count.

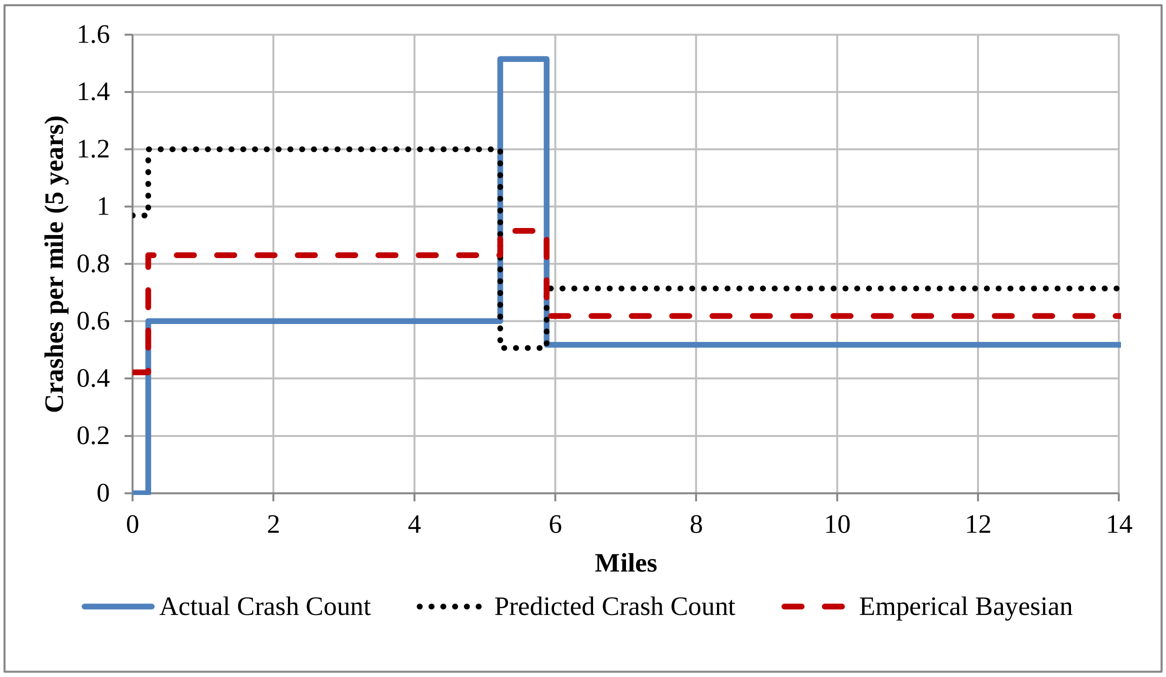


Figure 6.2a Segment-Based Safety Metrics for Roadway Section A

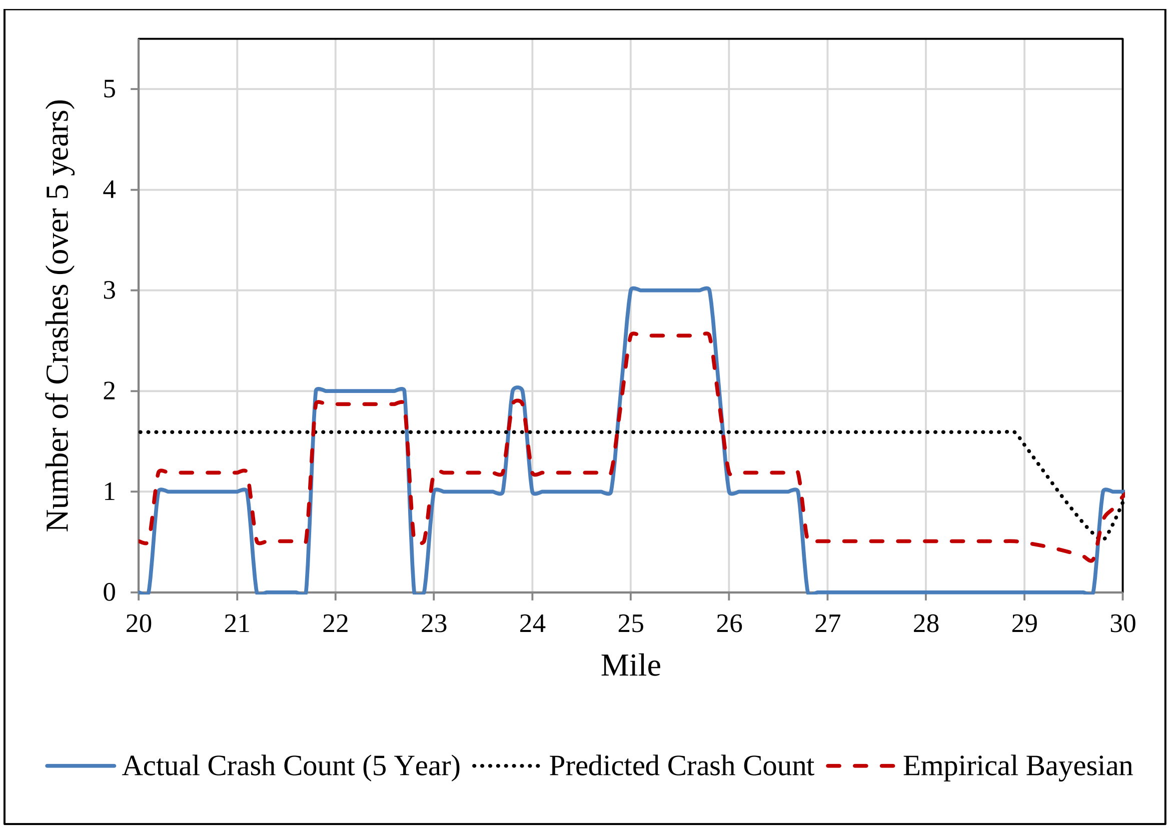


Figure 6.2b Window-Based Safety Metrics for Roadway Section B

7. GIS HIGHWAY SAFETY REVIEW (GIS-HSR) TOOLS

7.1 GIS Highway Safety Review (GIS-HSR) Tools Process

The GIS Highway Safety Review Tools (GIS-HSR Tools) were implemented in a Python script, using ESRI's ArcGIS geo-processing framework. The tool is designed to be a comprehensive tool to quickly identify high-risk locations, using a data-driven methodology. The tool's functions perform four distinct steps.

1. Join and filter geometric and traffic data to a single file. The data provided by SDDOT were in a variety of formats and files. All information needed to be in a single table to calculate predicted crash counts. The data required spatial joins to link the data to the basic segments. Traffic data are stored in a line work shapefile, however, and they do not match the underlying geometric shapefile. In many locations, the line work does not match the actual surface conditions, and endpoints are frequently not the same as the ends of the basic geometric segments. To account for this, the basic segments need to be split where either line work ends. The curves (both vertical and horizontal) are stored as point shapefiles located at the PI. These data needed to be associated with a unique link by a spatial join (run though ESRI's geo-processor). Crash data were also filtered in this step. Animal and intersection-related crashes were removed during this step. Animal-related crashes were removed as they are difficult to prevent through engineering improvements. They were removed based on a "TRUE" value in the "Is Wild Animal Related" field. Intersection crashes were removed by spatial selection of crashes. Crashes were removed within a 150-ft buffer of intersections. Intersection-related crashes were removed because the analysis techniques applied are not designed for intersections. The crash data were then spatially joined to the segment for calculating the EB statistics.

2. Predicted crash counts were calculated for each homogeneous segment using the appropriate equation as in Table 7.1.

   These values are based on the geometric and traffic conditions along a given link. The MVMT (Million Vehicle Miles Traveled) is the most important variable in crash prediction. It is widely accepted that the traffic exposure is a major contributing factor to the number of crashes. It is interesting to note that percent trucks are always the largest negative coefficient. This is in agreement with the results found in a previous study (Miaou 1993). The indication that percent trucks has the largest coefficient may be deceiving as percent trucks being represented as a number between zero (0) and one (1), while many other variables are integers, or fractions potentially greater than one. For example, a rural Major Collector, a change in percent trucks of 5% (a fairly large change in percent trucks), will change the exponent by 0.1061, while a change in speed limit of 5 mph (the smallest change in speed limit possible) will change the exponent by .24. It shows that the percent trucks has a relatively minor impact when compared with other factors even though the coefficient is the largest.

Table 7.1 Safety Performance Functions

Functional Classification	Predicted Crash Count	Dispersion Factor
Rural Major Collector	$= MVMT^{0.689} exp(0.577 - 0.015\ Lane\ Width + 0.048\ Speed\ Limit + 0.015\ Left\ Shoulder\ Width - 2.122\ Percent\ \ Trucks + 0.07.07\ Vertical\ Curve\ Density - 0.833\ If\ Municipal\ Funding\ Catagory)$	1.328
Rural Minor Arterial	$= MVMT^{0.869} exp(5.640 - 0.027\ Lane\ Width - 0.021\ Speed\ Limit - 0.015\ Vertical\ Curve\ Density - 0.085\ Right\ Shoulder\ Width - 1.404\ Percent\ Trucks - 0.594\ If\ Has\ Median - 0.296\ If\ Has\ Rumble\ Strip - 1.234\ If\ Municipal\ Funding\ Catagory\ )$	0.483
Rural Principal Arterial-Interstate	$= MVMT^{0.888} exp(4.954 - 0.038\ Lane\ Width - 0.013\ Speed\ Limit - 1.350\ Percent\ Trucks - 0.029\ Vertical\ Curve\ Density - 0.002\ Median\ Width - 0.430\ If\ Has\ Rumble\ Strip)$	0.616
Rural Principal Arterial-Other	$= MVMT^{0.704} exp(0.389 - 0.031\ Lane\ Width + 0.056\ Speed\ Limit - 1.832\ Percent\ Trucks + 1.087\ If\ Has\ Curb - 0.863\ If\ Municipal\ Funding\ Catagory)$	1.339

Speed limit is also found to be an important predictor variable, however, the sign of the coefficient changes depending on the functional classification. The functional classifications where speed limit has a negative coefficient (higher speed relates to fewer crashes) are “Rural Minor Arterial” and “Rural Principal Arterial Interstate.” This is a classic example that statistical correlation does not necessarily imply causation. Higher speed does not contribute to fewer crashes. These roadways are typically built to a higher design standard than collectors, and could indicate the roadways with a higher speed limit are built safer than the lower speed limit roadways. Typically, Interstate and arterial roadways will be posted with a speed limit of 75 mph or 65 mph respectively, and any deviation would indicate that conditions are not appropriate for those speeds and a lower speed limit would be implemented. However, when examining “Rural Major Collector” and “Rural Principal Arterial-Other” there is much more variation in the design standards of these roadways. It is much more likely that the speed limits on these roadways should be individually examined but likely have been determined through a “blanket implementation.”

The variable “Is Municipal” is believed to not directly affect the safety of a highway, but is an indicator of other unaccounted for factors such as maintenance, shelter provided by buildings, and general urban roadway characteristics. While the exact impact of this predictor variable is unclear, the fit of the model is substantially better when included. An alternative which could be considered is developing a separate classification for roadways falling within any city’s limits; however, for the purpose of this study, and because of the fairly small dataset where “Is Municipal” is true, it was decided to be included as a logical variable within the equations.

Vertical curve density is chosen as a predictor variable, instead of vertical curve count, due to a strong correlation between VMT and the curve count. The density of the curves tends to be uniform, with longer segments containing more curves. To avoid including the length of the segment as part of multiple parameters, the density was chosen to represent vertical curve presence. When the coefficient is examined, it shows that a higher vertical curve density relates to fewer crashes, for two of the three functional classifications where vertical curve density is a significant variable. This is counter to intuition and results found by Zhang and Ivan (Zhang and J. Ivan 2005). The influence of vertical curves on crashes requires additional exploration in the future, using more detailed curve data not available for this study.

The coefficients for lane width are fairly consistent across all functional classifications. The negative coefficient (wider lanes result in fewer crashes) is logical, as the space between vehicles is larger, and vehicles have to travel farther to enter opposing traffic or leave the roadway. This result is consistent with the methods presented in the Highway Safety Manual (AASHTO 2010b).

Shoulder width appears as a coefficient in two functional classifications. Only one shoulder (left or right) was considered as typically both are the same width for most undivided roadways. The coefficient for “Rural Minor Arterial” matches the intuitive relationship. The data show that for minor arterial,s a wider shoulder will reduce the number of crashes, as the shoulder provides space for broken down vehicles and additional recovery space for vehicles that depart the travel surface. These results are in agreement with prior studies conducted, showing that wider shoulders decrease crash occurrence. While the amount of reduction is different in many studies, the basic result that a wider shoulder reduces crash occurrence is constant (Gross and Paul Jovanis 2007; Harwood et al. 2000; Zegeer et al. 1988).

The presence of rumble strips appeared as a significant variable for two functional classifications. The coefficient for both is negative (fewer crashes). This is logical and shows the implementation of rumble strips improves safety as is expected, and is shown in many other studies. The effectiveness of shoulder rumble strips indicates that they can reduce overall crashes by 14-17% (FHA 2010). In fact, a memorandum issued by FHWA recommends rumble strip or rumble stripe as a proven safety countermeasure. Shoulder rumble strips have been documented to reduce run-off-road crashes by 7% to 41% (FHA 2010). The presence in only two functional classifications is most likely due to limited implementation thus far for the other classifications as can be seen in Table 5.3.

The median characteristics, both presence and width, show a reduction in crashes. The median provides a physical barrier and is typically a safety feature; the negative coefficient indicates that medians improve safety.

The presence of curb is another variable of interest. The coefficient shows that for “Rural Principal Arterials-Other” curb and gutter increase the number of crashes. It is suspected that the presence of curb is likely to be an indicator variable for narrow right of way, as a ditch-based drainage system is preferred in a rural environment. It is well understood that adequate clear zone is an important factor in crash reduction. Future work is needed to identify more clearly the causational factors indicated by the variable “has curb,” including a detailed review of these site characteristics.

It is important to note that the correlation between two variables does not necessarily relate to a causational relationship. While the statistical tests find that some features increase crash risk and others decrease the risk, it is not guaranteed that those characteristics by themselves are directly responsible. When analyzing the results of the SPFs, it is important to keep in mind that these parameters are merely representative of the conditions on a highway, and do not necessarily explain the entire situation.

3. A continuous sliding window was used to calculate crash data by severity and the predicted crash count. A one-mile window (excepting the end of a roadway, where the window length will decrease) is incremented 0.1 miles. The sliding window code was designed to work within the existing infrastructure provided in ArcGIS. Information about the roadway was added to the sliding window through a spatially weighted average. A spatially weighted average was used to account for heterogeneous stretches of roadway. When the expected crash rate changes due to changing roadway or traffic conditions, this change needs to be reflected in the window. The spatially weighted average of the SPF values was used to determine the EB prediction of safety in the next step.

   The continuous sliding window algorithm uses two tables to create windows: 1) a "Located Event Table" containing a route identifier, a "measure" (displacement from the start of the roadway) and the crash data; 2) a table containing the route IDs, the endpoint "measures" needed to establish the limits of analysis, and the roadway data. The algorithm dynamically creates windows based on the input. This allows any size window or increment to be used by modifying a parameter (a default is used for the basic tool, and is set in the code). The sliding window algorithm calculates the limits of each window and then counts the crashes within the window. In the basic tool, crashes are categorized by severity but can be readily extended to any other field containing a coded value, such as crash type, by using the sliding window tool, also included in the HSR-Tools suite.

4. Additional performance indices were calculated based on both the link and window data. Crash counts and rates were calculated to provide a historical view of the location and provide continuity to previously implemented techniques. An EB prediction of the crash count for each location was also calculated. The EB method is the primary, most reliable performance index. The deviation from the predicted crash count was calculated, with the intent of identifying locations where the SPFs are unable to accurately capture safety risks. The deviation is calculated as both the segment difference and the window difference. Large positive values indicate locations which may be of interest.

   During tool development, the computational performance of the tool became a major concern. Early tests indicated the tool would require more than a week to process the state of South Dakota using a 5-year crash dataset. To improve the performance, it was decided to analyze the state in smaller portions. When the state was divided into 12 areas, which coincided with SDDOT's existing administrative areas, tool performance was improved, requiring only a single day to perform analysis on the test computer. It is believed to be associated with a nonlinear processing time requirement when large datasets are passed to the ESRI geo-processing tools within Python.

7.2 GIS Highway Safety Review (GIS-HSR) Tools Interface

The interface of the tool was implemented through the ArcGIS ArcToolbox (ESRI 2010). This interface provides an appearance which is consistent with other tools in ArcGIS as shown in Figure 7.1. The use of the tool requires a minimal familiarity with the ArcGIS Desktop from a user perspective Figure 7.2 shows the dialog interface. Each individual component of the tool is included in the ArcToolbox. Most users will interact with the "Basic Tool by Zone."

- First, the location the tool will use to store the output and intermediate data must be specified in the "Output Folder" field.
- The "Roadway Features" variable is an ESRI shapefile containing segment data with basic geometric data as maintained by SDDOT.
- SDDPS-maintained crash data are stored in a shapefile with points for each crash. Each individual year is contained in a single file. The tool counts the number of files to determine the study period and uses this to calculate VMT and crash rates.

- The “Intersection Features” variable requires a point shapefile used to remove intersection-related crashes.
- The variable “Vertical Curve Features” uses a point shapefile maintained by SDDOT.
- “Horizontal Curve Features” is also a point shapefile maintained by SDDOT.
- The “Traffic Data Features” file contains ADT and truck ADT stored in a linear shapefile. This file is maintained by SDDOT and contains line work drawn to provide a symbolic representation of the roadway.
- “Zone Features” is a polygon shapefile used to break apart the study. Additionally, the “Zone Name Field” is a field from this file containing the name of each polygon, used to identify the output files. This is used to improve the performance of the algorithm. This file is typically the SDDOT regions, used for administration.

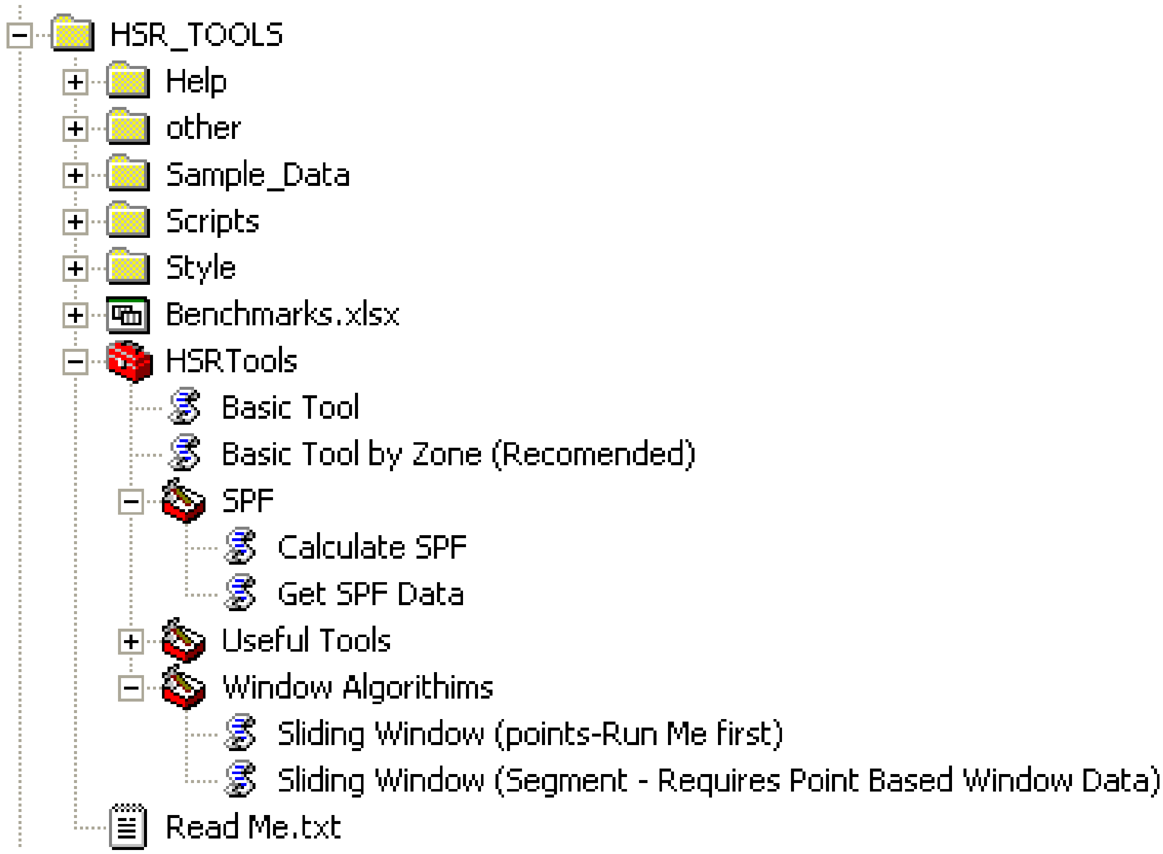


Figure 7.1 GIS-HSR Tools Toolbox

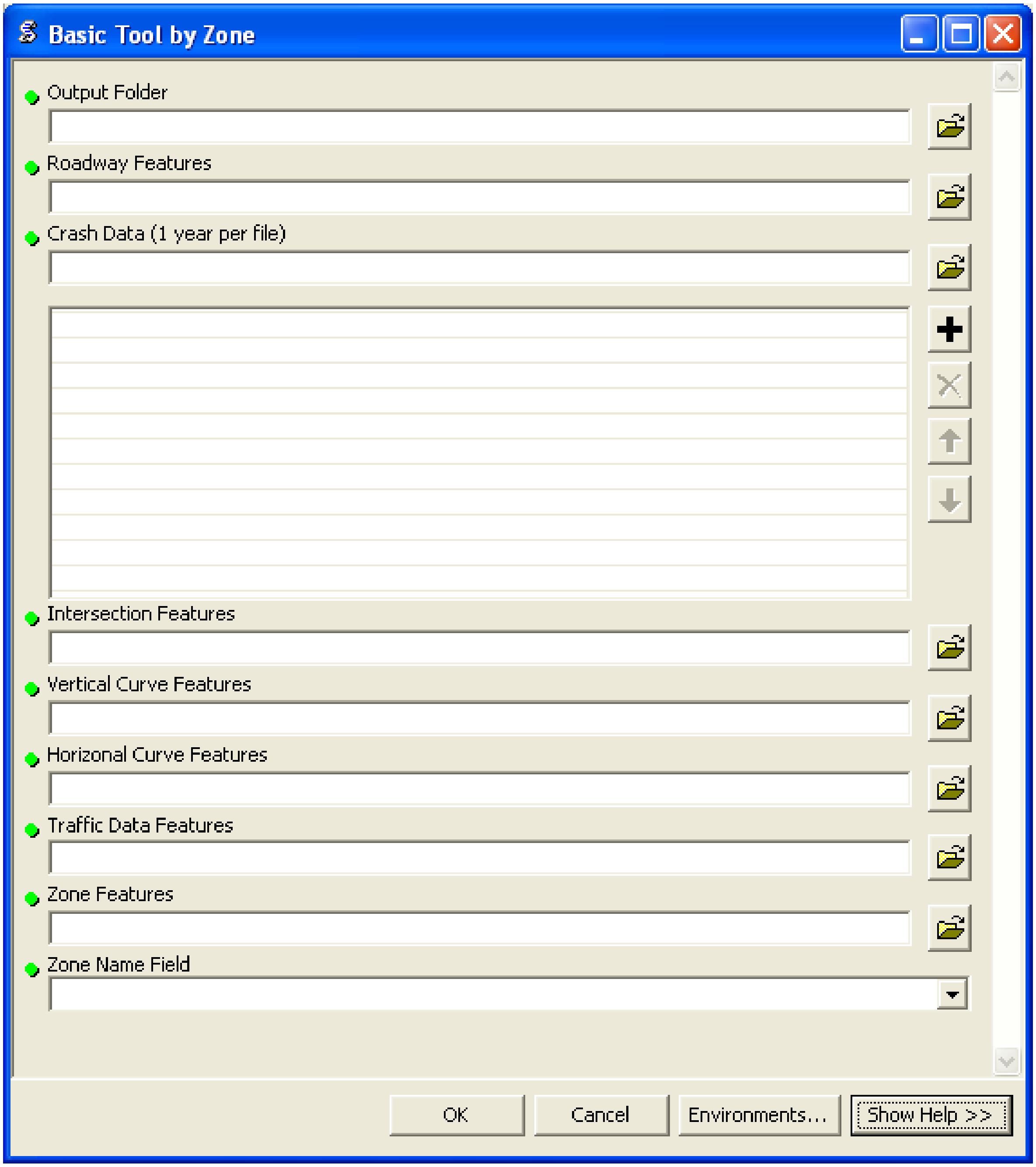


Figure 7.2 GIS-HSR Tools “Basic Tool by Zone” Interface

The tool was designed to simplify user interaction as much as possible. Although complex processing is performed, after a user specifies the input files, simply clicking the OK button will perform the analysis without further user interaction before interpretation of the results.

Additional “advanced” components are also included in the GIS-HSR Tools. These advanced components are actually just smaller portions of the code, which are useful for additional analysis not planned for in the basic tool. The tool “Get SPF Data” performs the data processing necessary to calculate predicted crash counts using the SPF’s provided. This tool is intended to allow easy access to the data necessary for future calibration of the crash prediction models. The “Calculate SPF” tool within the SPF toolset runs the same calculations as the basic tool; however, it will stop after the SPF values for each segment have been calculated.

The “Window Algorithms” toolset includes the two sliding window scripts developed. These tools only perform the sliding window analysis. The point-based sliding window will perform a count based on any user-defined field in the point event table, and using any set of window length and increment. The segment-based sliding window will perform a spatially weighted average underlying line segments. This script requires that the window extents be defined by the point-based sliding window algorithm first.

7.3 Output

The output of the algorithm is a set of ESRI personal geodatabases. Figure 7.3 shows how the output data are organized. The input data are split and placed in individual folders in the output folder. All interim data and the final output are contained within the geodatabase for each “Zone.” Three feature classes included in each Geodatabase contains most of the information a user will require. Those feature classes are:

- *Windows:* This table contains the output from the sliding window algorithm. This table contains one record per window generated in the sliding window analysis. Data contained in this table include the spatially weighted average of the SPF, the EB crash count, the actual crash count, and the crash count by severity.
- *Segments:* This file contains data based on each homogeneous segment. Each individual homogeneous segment will be a separate record. Additionally, these segments are not necessarily the same as the segments provided by SDDOT. Segments are broken at either a change in ADT (as provided by SDDOT) or a change in roadway geometry. Actual crash counts, SPF values, and EB values are included in this table.
- *Route:* This file contains the linear referencing information needed for ArcGIS to plot the windows on a map.

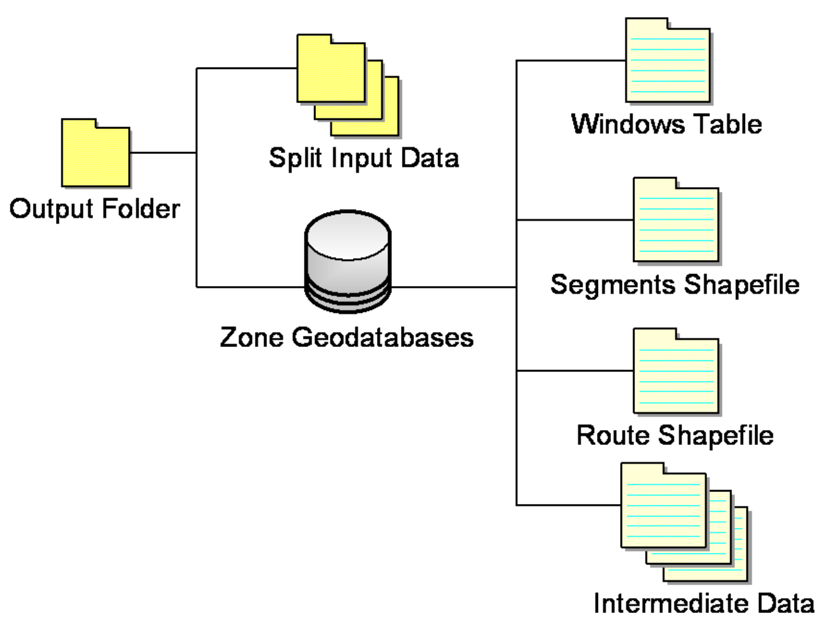


Figure 7.3 Output File Structure

8. APPLICATION

The output produced by SD GIS-HSR Tools provides substantial information, which can be interpreted based on the safety metrics proposed above. The three classes of proposed safety metrics provide substantially different results. The results were mapped on a statewide level in Figure 8.1 to Figure 8.6. When comparing the window-based and homogeneous segment-based maps, the results are fairly similar. However, when an individual segment is examined, the sliding window provides a higher resolution for crash rate calculation. This is useful for identifying specific locations in need of improvement. Additionally, the window-based analysis ignores changes in the roadway geometry.

When examining the sample South Dakota data, the raw crash rate calculation (crash rate with no statistical modification) indicates that the area near Rapid City (in west-central South Dakota) is an area of concern. Additionally, the northwestern corner of the state experiences higher crash rates than the majority of the state. This would lead the analyst to examine those areas in greater detail.

When comparing the raw crash rate map with the EB crash rate map, it becomes apparent that many locations may be overlooked by the crash rate method. The EB method is generally accepted as a more accurate method of determining the safety risk at a location. Examining the sample dataset shows that most of far western South Dakota contains high safety risks.

Several additional high-risk locations are identified in central and eastern South Dakota. These locations can be identified by the darker spots on the appropriate map. The excess crash count results are an independent measure of safety. The window-based excess crash count reveals shorter segments where the numbers of crashes are much higher than the predicted crash count. Additionally, the excess crash count method indicates a safety risk along the eastern edge of South Dakota; however, this location did not attract as much attention when either of the crash rate methods was applied. This method also indicates that several major corridors are performing better than expected. Despite this result, these locations should not be ignored if other methods indicate these are problem locations, such as I-90 near Rapid City.

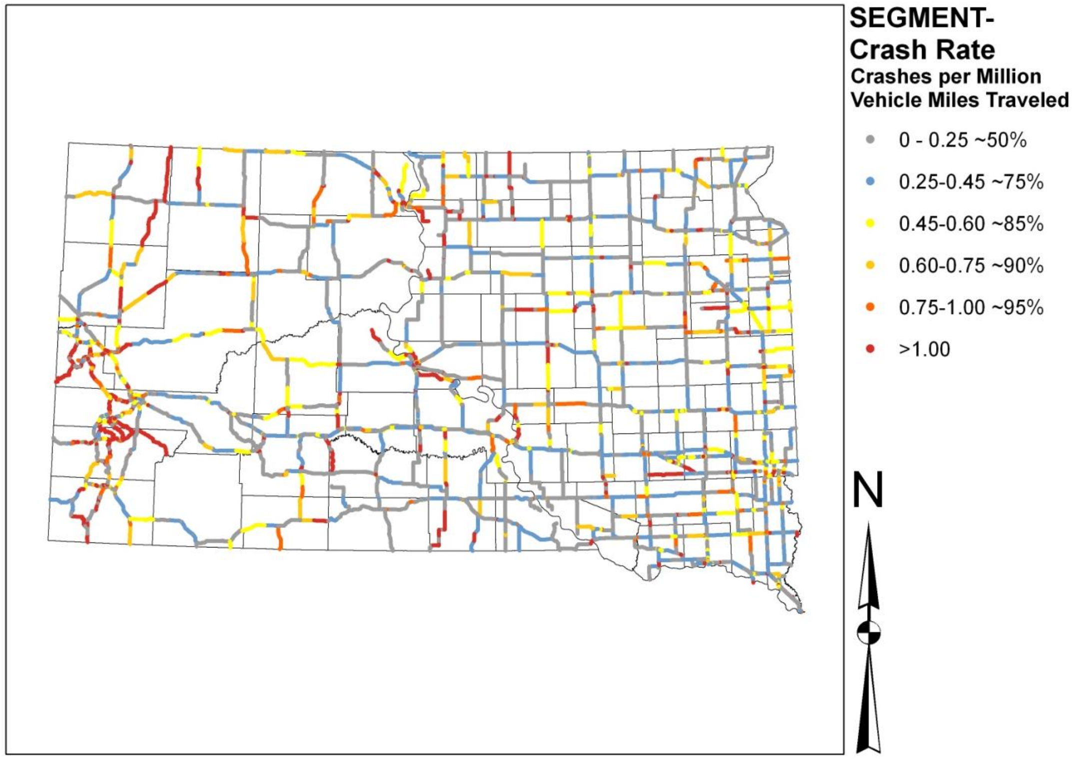


Figure 8.1 Crash Rate Segments

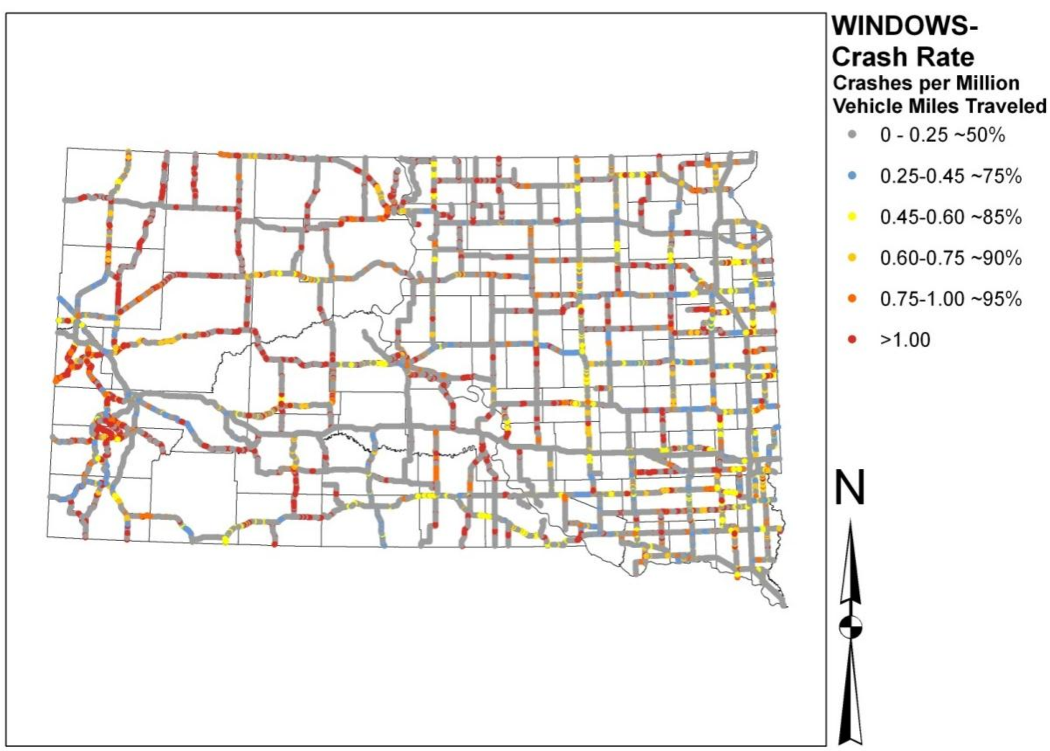


Figure 8.2 Crash Rate Windows

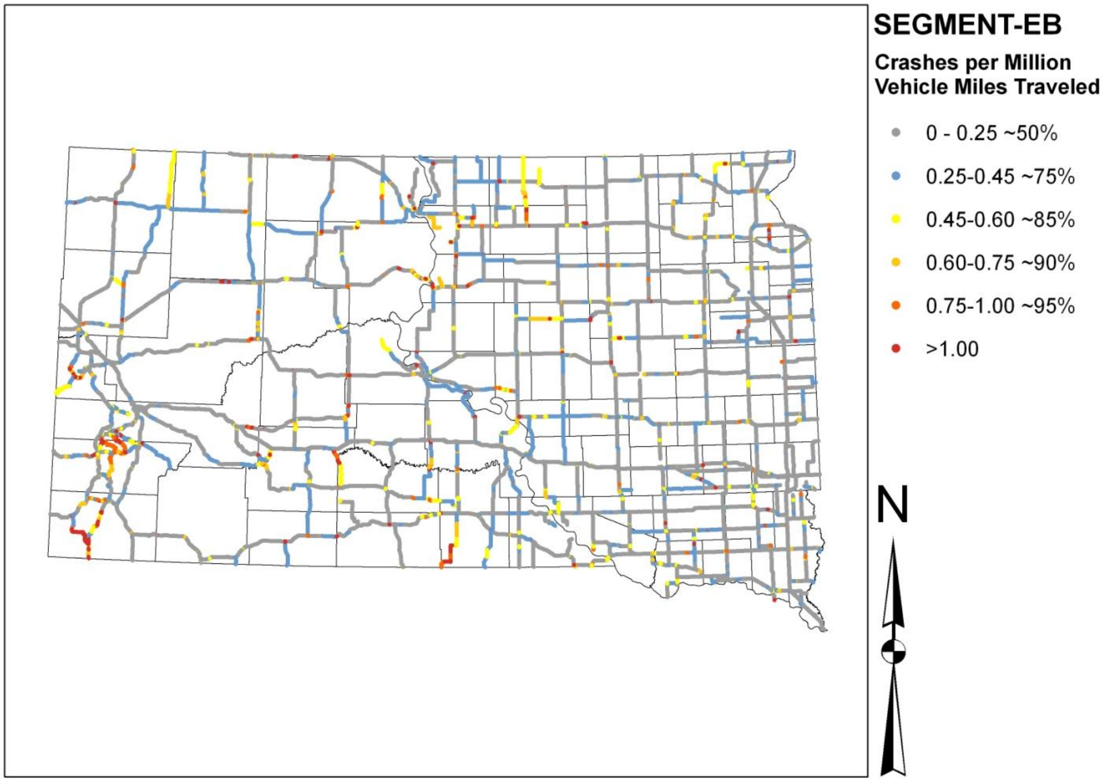


Figure 8.3 EB Crash Rate Segments

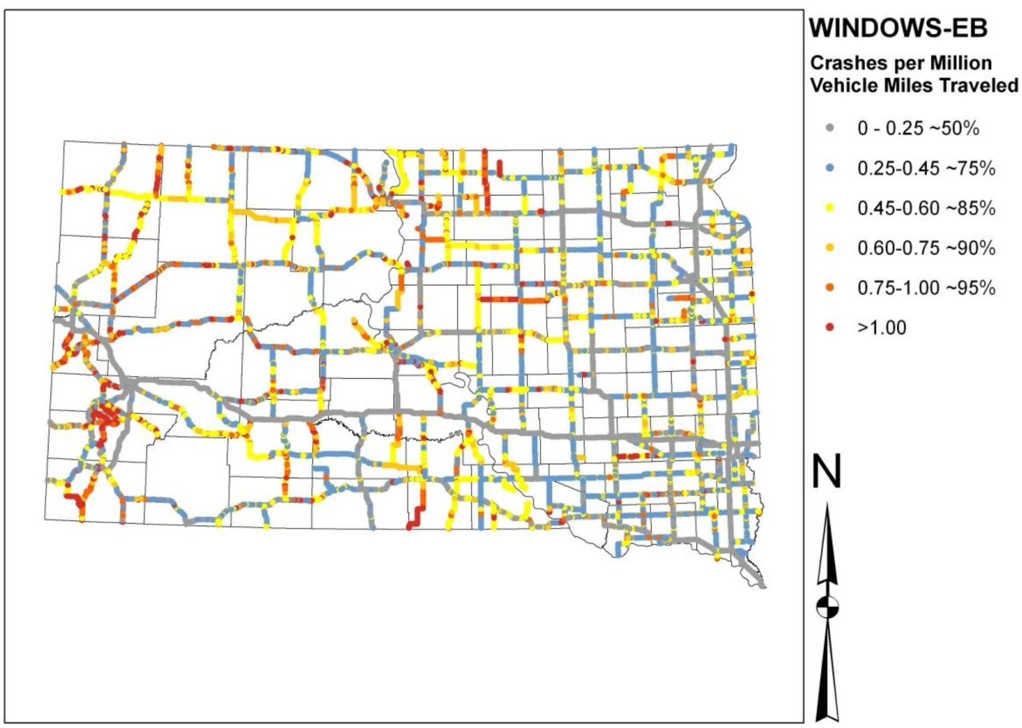


Figure 8.4 EB Crash Rate Windows

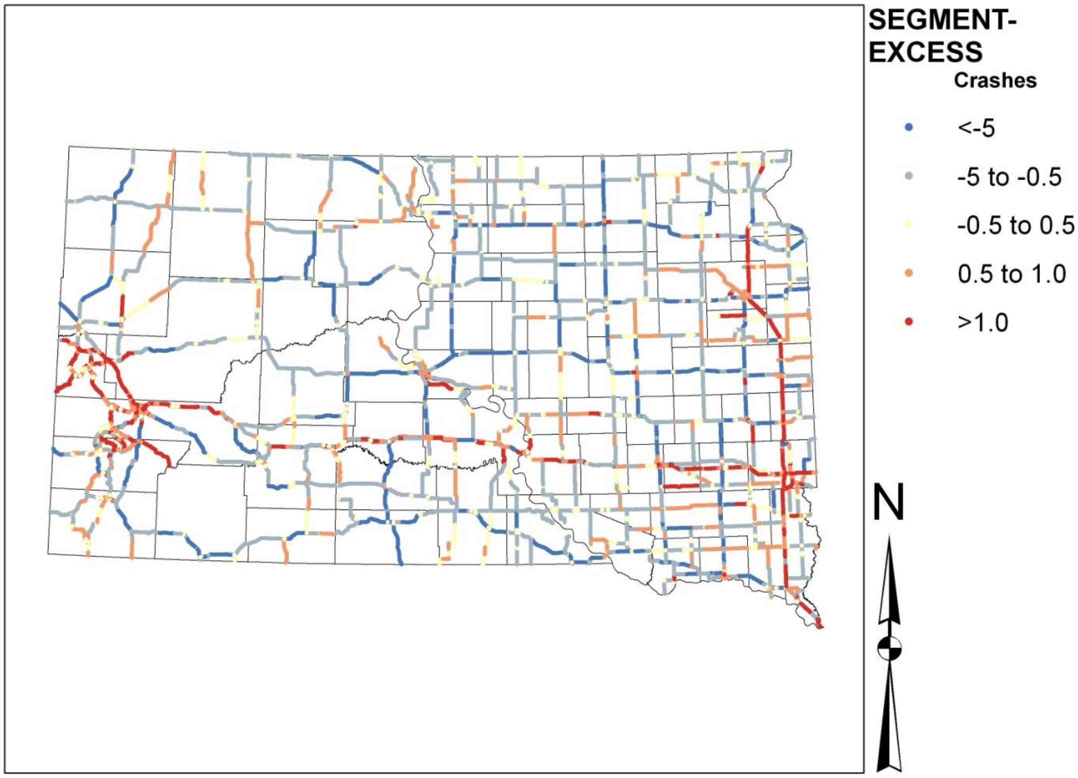
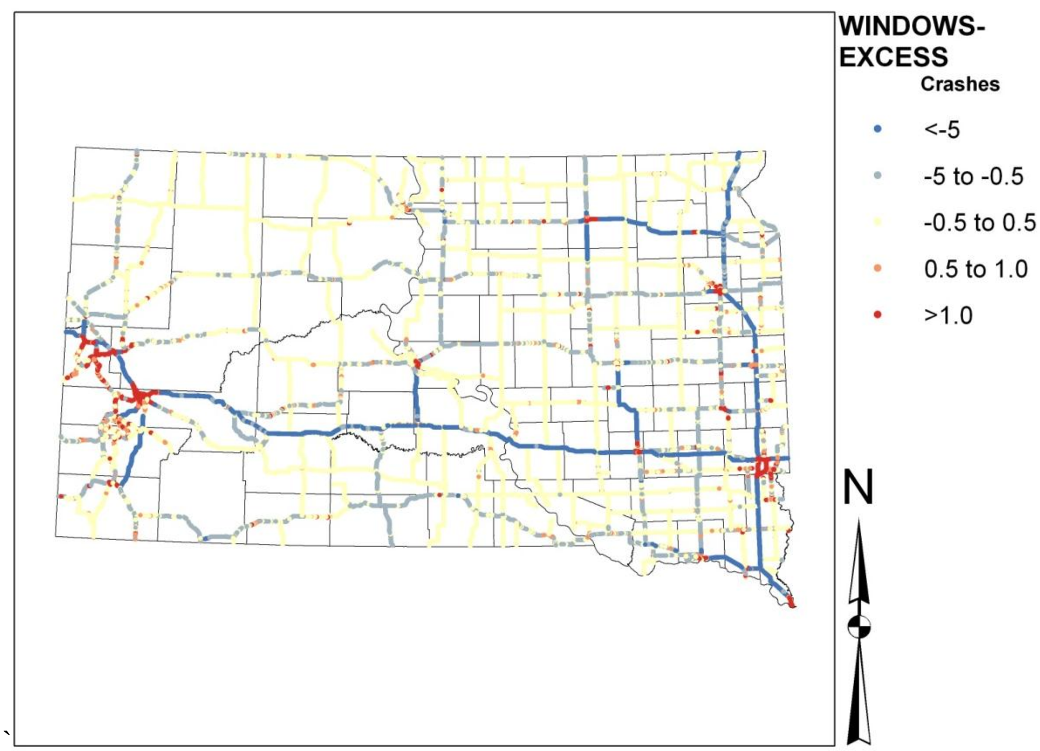


Figure 8.5 Excessive Crashes Segments



`

Figure 8.6 Excessive Crashes Windows

With the crash severity information, the fatal and severe injury high-risk locations can be mapped on a statewide basis. As illustrated in Figure 8.7, crash clusters with F & A can be identified visually in several areas. However, given the relatively low VMT, a single F or type A injury crash may lead to a high crash rate. Caution is advised when using the F & A crash rate map.

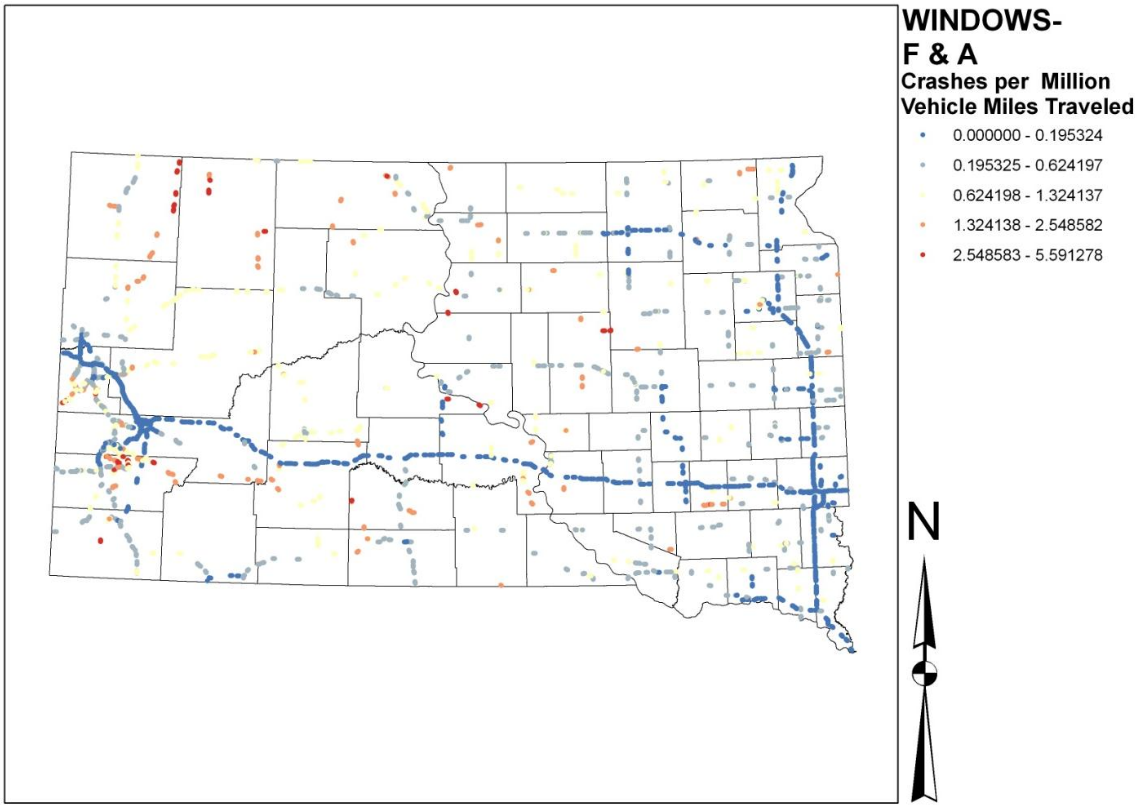


Figure 8.7 Fatal and Incapacitating Injury Crash Rate Windows

While the GIS–HSR Tools is useful for identifying high-risk areas on a statewide scale, it is also useful for identifying specific high risk areas. The sliding window approach can provide not readily available information which can guide engineers in defining the limits of a highway improvement project. When the output is mapped at a smaller scale, detailed information regarding the crash risk along a corridor becomes apparent. Frequently, the homogeneous segments, on which the roadway data are based, have long stretches of roadway several miles long. Many high-risk locations exit completely within a single long segment of highway. A segment-based method is unable to identify these locations. Figure 8.8 shows a location on SD Highway 212 where a several-miles-long stretch of highway experiences a high crash rate; however, the segment does not, as few crashes occur elsewhere along the segment. This short stretch of roadway could benefit from improvements, but those will never be identified without the sliding window analysis.

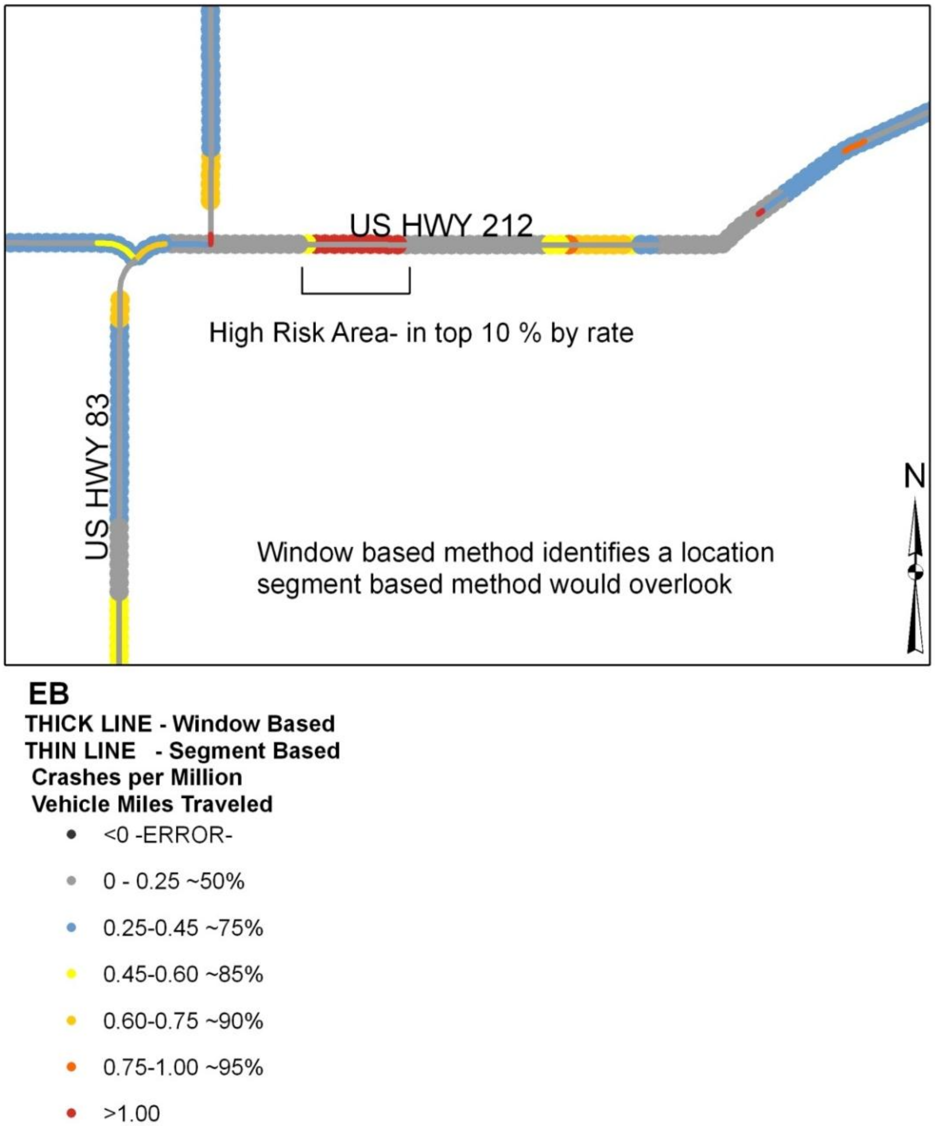


Figure 8.8 US 212 Window Based High Risk

Another situation where a sliding window method can potentially identify high-risk locations, and provide additional guidance where a segment-based approach lacks is where several short segments exist near each other. This is a common situation in towns where speed limits and roadway geometry progressively change on the approach to a town. These short segments may not be able to appropriately capture the impact of the conditions in the area. Figure 8.9 shows a stretch of SD Highway 25 where a series of short segments fail to capture the safety risk in the area. The window-based analysis identifies a stretch of approximately five miles where crash risk is among the top 10% statewide, however, only a short series of six segments in an approximately ¾-mile stretch appear as a high risk.

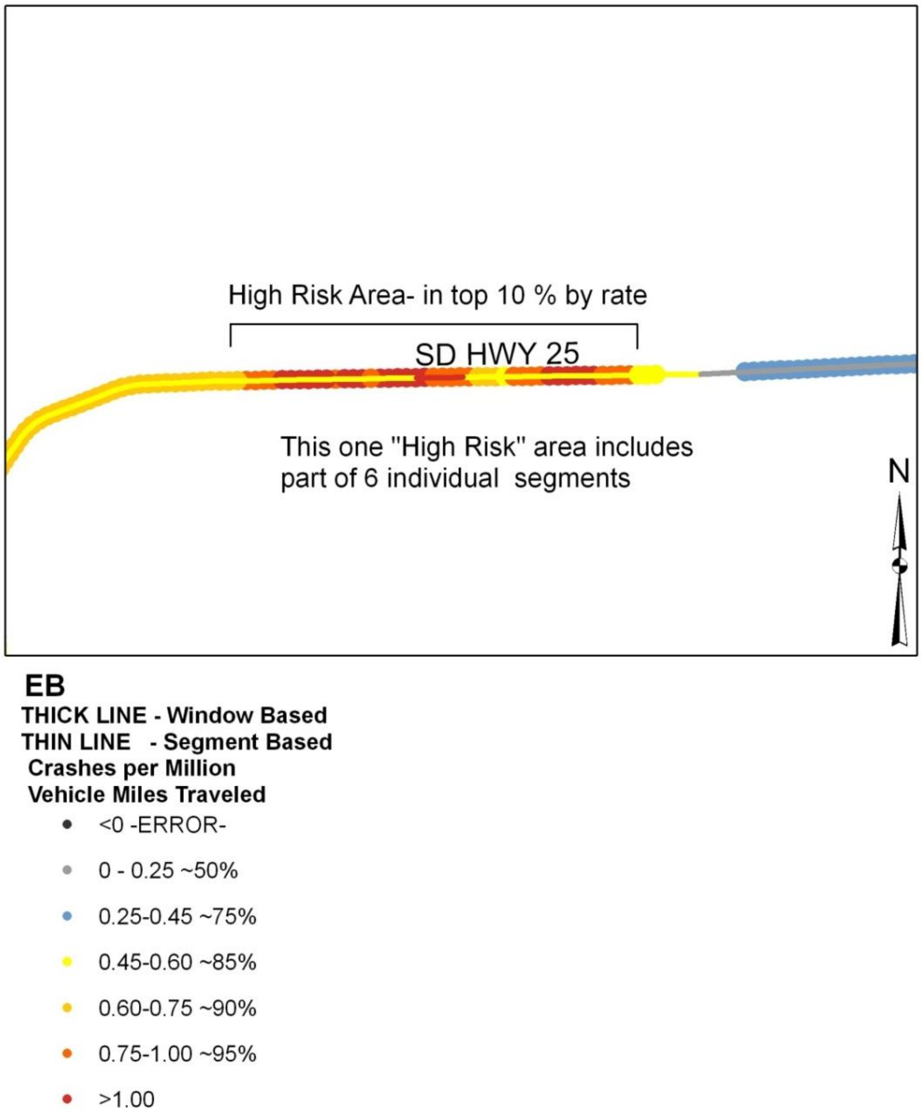


Figure 8.9 SD 25 Where A Short Segment Influences A Longer Stretch

While the window-based method is the primary approach to analyzing safety using the GIS-HSR Tools, the segment approach can frequently identify locations of high risk but the window-based method may not identify. Figure 8.10 shows a location on US Highway 212 where a segment falls in the top 10% for EB crash rates, while the window-based method through the smoothing applied shows only a slight risk in this location. These locations are typically short segments where geometric conditions are likely to contribute to higher crash frequencies than the surrounding roadways. Improving these roadways to standards similar to the surrounding roadways has the potential to reduce crashes at these locations.

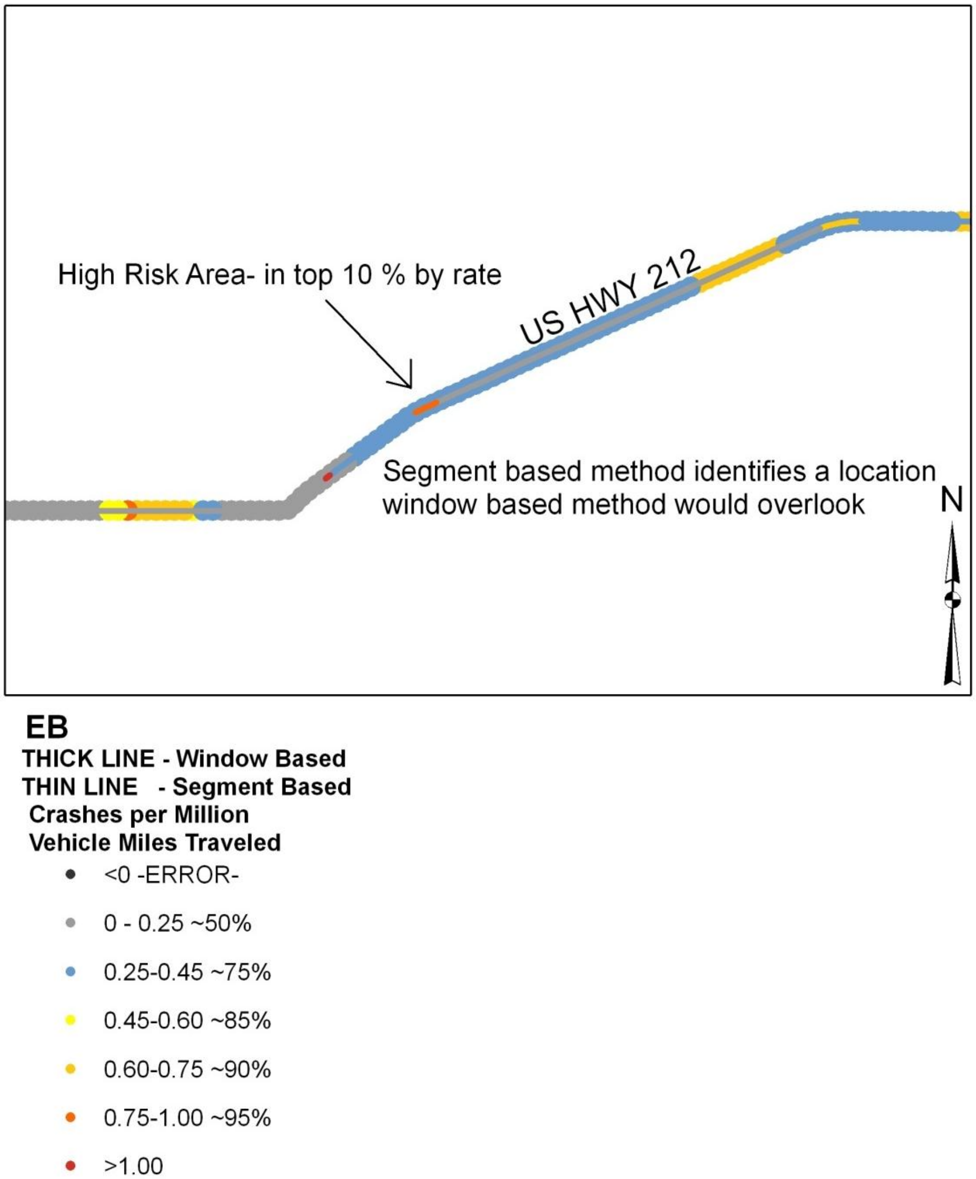


Figure 8.10 US 212 Segment Based High Risk

Each performance indice also provides unique indicators of safety risk. Figure 8.11 shows one location in which the different performance metrics produce different results when identifying high risk locations. The map at the top of the figure shows the EB crash rate, corresponding to the chart below. Examining the chart by historical crash data only, the first location of interest is near mile 12.5. However, this location is not as severe when the EB crash rate is examined. The same location does exhibit a fairly high excessive crash count, as the historical crash rate is much higher than the predicted crash rate based on the SPF. By comparison, the area near mile 12.0 has a lower excessive crash count, as the predicted crash count is larger than near mile 12.5. When examining the location near mile 10.2, the EB crash rate identifies a somewhat high crash safety risk, in the top 10% of high-risk locations. This location would not be identified if a historical crash risk were used, as no crash history exists in this location. However, based on the crash history and characteristics in the remainder of the state, this location is likely to experience a high crash rate in the future.

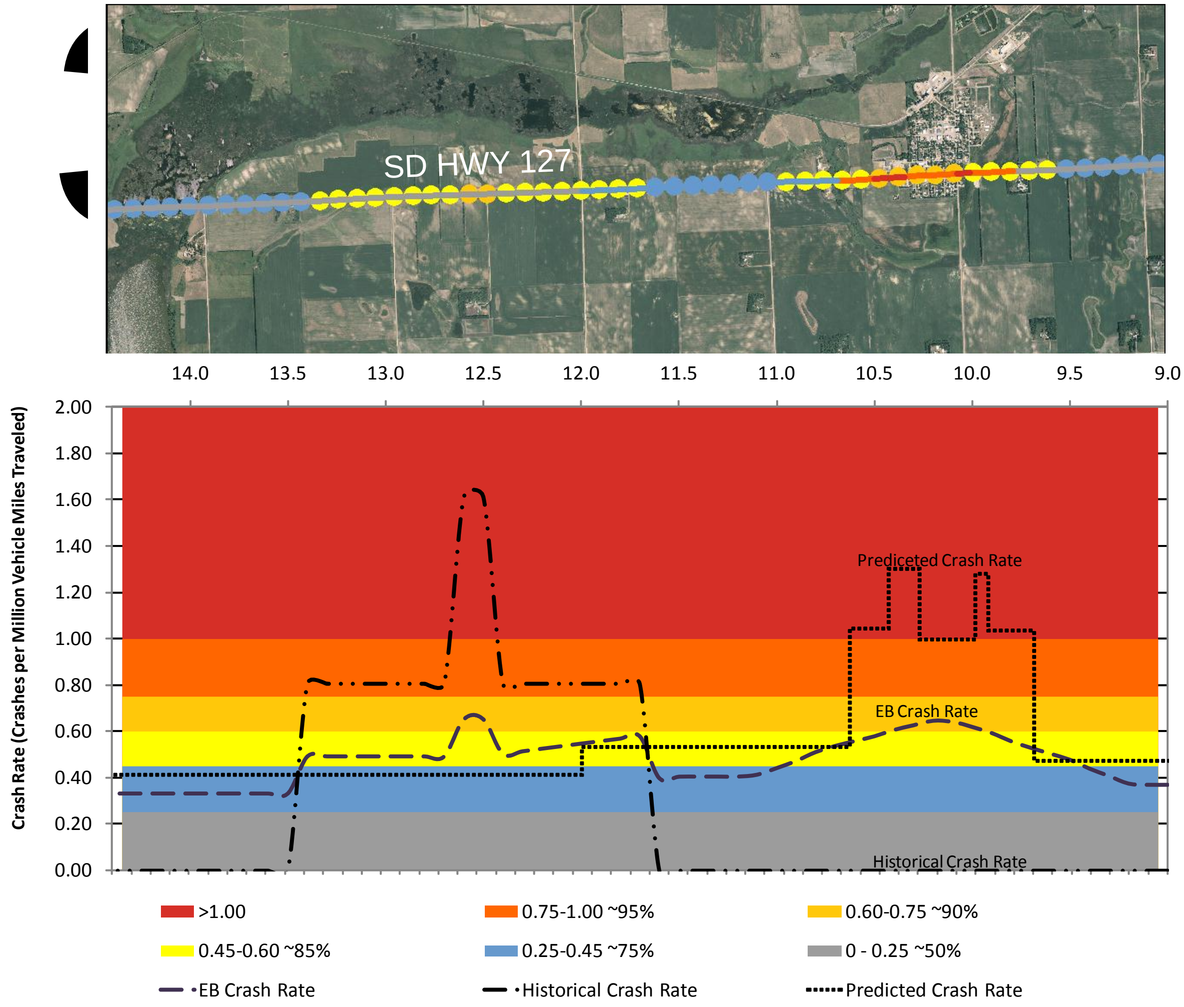


Figure 8.11 Performance Indices For SD HWY 127

9. CONCLUSIONS AND FUTURE WORK

The South Dakota GIS Highway Safety Review (GIS-HSR) Tools provides a data-driven approach toward identifying high-risk locations. Only very basic user input and interaction is required. Requiring minimal data and implementing a system-wide application, the tool was designed to address the rural environment and is tuned specifically for South Dakota. However, the general architecture and design are valid in any location.

GIS-HSR Tools implement a sliding window method to provide a fine resolution in identifying specific crash locations. This method identifies hotspots completely independent from the existing segments, allowing for any portion of a segment, or multiple segments, to be identified as high-risk locations.

The toolbox provides multiple methods of identifying high-risk locations. Crash rate data are provided to maintain consistency with many existing techniques. Predictive techniques are provided. The EB method is used to calculate a statistically-based estimate of the safety in a location. This method accounts for roadway characteristics in determining the safety of a location. The excessive crash frequency is calculated to identify locations where the provided models are generating poor estimates of the crash frequency. When applied to the South Dakota dataset, the algorithm performs as expected and produces results which are useful for identifying high-risk locations.

This study focused on fairly basic implementation of the GIS-HSR Tools. This tool was initially intended to address the rural safety problem, however, only limited data are available for non-state roadways in the rural environment. An additional implementation of the basic tool is proposed, with intent to accommodate the very limited data available on the state highway system. This will involve the development of new SPFs for the local roadway systems.

Additional work is needed on high severity crash analysis. The causational factors affecting high severity crashes are not the same as other crashes. Additional regression and calibration of SPFs for high severity crashes have the potential to greatly improve the accuracy of high severity crash prediction and prevention.

Many specific types of crashes are more frequent in rural settings than urban settings. Crashes such as run-off-road and animal-related collisions are much more frequent in the rural environment, and are usually not affected by the same factors as other crashes. Further analysis using GIS-HSR Tools and research into causational factors shows potential to provide new insight to the causes of these crashes.

While the tool currently is a user-friendly application, analysis of the results requires an analyst to manually interpret all results. Additional tools are proposed to improve the interpretation. A report creation tool is proposed to provide a ranked list of high-risk locations. This tool would need to be able to understand the spatial relationship between windows and segments to propose the limits of any proposed highway improvement project. This would further aid the ease of use for the GIS-HSR Tools, reducing the effort required for result interpretation.

REFERENCES

AASHTO. (2010a). “AASHTOWare.” <http://www.aashtoware.org/Pages/default.aspx> (Dec. 16, 2010).

AASHTO. (2010b). *Highway Saftey Manual*.

Abbess, C., D. Jarrett and C. Wright (1981). “Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment with Special Reference to the “Regression-to-Mean” Effect.” *Traffic Engineering and Control*, 22(910), 535-542.

Bland, J.M., and D.G. Altman (1994). “Statistic Notes: Regression Towards the Mean.” *British Medical Journal*, 308(1499).

Caliendo, C., M. Guida, M and A. Parisi (2007). “A Crash-Prediction Model for Multilane Roads.” *Accident Analysis and Prevention*, 39(4), 657-670.

Campbell, J. and National Research Council (U.S.). National Cooperative Highway Research Program. American Association of State Highway and Transportation Officials. United States. (2008). *Human factors guidelines for road systems*. Transportation Research Board, Washington D.C.

Drake, M.L., E.A. Sparks and J.E. Thomaz (2005). *Factors Contributing to South Dakota Crash and Fatality Rates*. Final Report, Purdue University.

Drakopoulos, A. and E. Ornek (2004). *Systematic Evaluation of Run Off Road Crash Locations in Wisconsion*. Final Report, Wisconsion Department of Transportation.

Elvik, R. (2008). “Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads.” *Transportation Research Record*, 2083, 72-75.

ESRI. (2010). “ArcGIS Desktop Help 9.2.” <http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=welcome> (Dec. 16, 2010).

“Federal Highway Administration.” (2010). <http://www.fhwa.dot.gov/> (Dec. 16, 2010).

Gross, F. and P. Jovanis (2007). “Estimation of the Safety Effectiveness of Lane and Shoulder Width: Case-Control Approach.” *Journal of Transportation Engineering*, 133(6), 362-369.

Harwood, D.W., F. Council, E. Hauer, W.E. Hughes and A. Vogt (2000). *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. Final Report.

Hauer, E. (1997). *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon, Oxford OX U.K. Tarrytown, NY U.S.A.

Jovanis, P.P., and H. Chang (1986). “Modeling the Relationship of Accident to Miles Traveled.” *Transportation Research Record*, 1068, 42-51.

Lord, D., S. Washington and J. Ivan (2005). “Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory.” *Accident Analysis and Prevention*, 37, 35-46.

Miaou, S. (1993). "The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions." *Accident Analysis and Prevention*, 26(4), 471-482.

Miaou, S. and G, Lum (1993). "Statistical Evaluation of the Effects of Highway Geometric Design on Truck Accident Involvements." *Transportation Research Record*, 1407, 11-23.

NHTSA. (2007). "Traffic Safety Facts 2007 Data."

Ormsby, T. (2004). *Getting to Know ArcGIS Desktop: Basics of ArcView, ArcEditor, and ArcInfo*. ESRI Press, Redlands Calif.

Powers, M. and J. Carson (2004). *Before-After Crash Analysis: A Primer for Using the Empirical Bayes Method*. Final Report, Montana State University.

R-Project. (2010). "R: Fit a Negative Binomial Generalized Linear Model." <http://stat.ethz.ch/R-manual/R-patched/library/MASS/html/glm.nb.html> (Dec. 16, 2010).

SAS. (2010). "The GENMOD Procedure: MODEL Statement." *SAS/STAT(R) 9.2 User Guide, Second Edition*, <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_genmod_sect022.htm> (Dec. 16, 2010).

Shankar, V., F. Mannering and W. Barfield (1995). "Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies." *Accident Analysis and Prevention*, 27(3), 371-389.

South Dakota Department of Public Safety. (2008). *2008 South Dakota Motor Vehicle Traffic Crash Summary*.

South Dakota Department of Transportation. (2007). *South Dakota Strategic Highway Plan 2007*. Highway Safety Plan.

Storsteen, M. (1999). *Identification of Abnormal Accident Patterns at Intersections*. Final Report, .

US Census Bureau. (2010). "Census Bureau Home Page." <http://www.census.gov/> (Dec. 16, 2010).

Vogt, A. and J. Bared (1998). "Accident Models for Two-Lane Rural Segments and Intersections." *Transportation Research Record*, 1635, 18-29.

Washington, S. (2003). *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman & Hall/CRC, Boca Raton.

Washington, S., and National Cooperative Highway Research Program. National Research Council (U.S.). American Association of State Highway and Transportation Officials. (2006). *Incorporating Safety into Long-Range Transportation Planning*. Transportation Research Board, Washington D.C.

Zegeer, C.V., J. Hummer, D. Reinfurt, L. Herf and W. Hunter (1988). "Safety Effects of Cross-Section Design for Two-Lane Roads." *Transportation Research Record*, 1195, 20-32.

Zhang, C. and J. Ivan (2005). "Effects of Geometric Characteristics on Head-On Crash Incidence on Two-Lane Roads in Connecticut." *Transportation Research Record*, 1908, 159-164.