Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs

Final Report

Prepared for Federal Highway Administration
September 2013

TECHNICAL DOCUMENTATION PAGE

1. Report No. FHWA-SA-14-005	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs		5. Report Date September 2013
		6. Performing Organization Code
7.Author(s) Raghavan Srinivasan and Karin Bauer		8. Performing Organization Report No.
9. Performing Organization Name and Address Highway Safety Research Center University of North Carolina 730 Martin Luther King Jr Blvd Chapel Hill, NC 27599-3430		10. Work Unit No. 11. Contract or Grant No. TPF-5(255)
12. Sponsoring Agency Name and Address Federal Highway Administration Office of Safety 1200 New Jersey Ave., SE Washington, DC 20590		13. Type of Report and Period 14. Sponsoring Agency Code FHWA
15. Supplementary Notes The contract manager for this report was Ms. Esther Strawder. A special note to all the States participating in the technical advisory group and the HSM Implementation Pooled Fund Study: your time and contributions were extremely valuable to the project team as we developed the Safety Performance Function Guides.		
16. Abstract This guidebook is intended to provide guidance on developing safety performance functions (SPFs) from the Highway Safety Manual (HSM) (AASHTO, 2010). The guidebook discusses the process to develop jurisdiction specific SPFs. It is intended to be of use to practitioners at state and local agencies and to researchers.		
17. Key Words: Safety Performance Functions, Highway Safety Manual, Pooled Fund Study TPF-5(255)	18. Distribution Statement No restrictions.	

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 45	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed pages authorized

Notice

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the use of the information contained in this document.

The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

Contents

Executive Summary

This is a "how-to" guidebook for states that are developing jurisdiction-specific safety performance functions (SPFs). The guidebook discusses the issues associated with the development of jurisdiction-specific SPFs and provides a step-by-step procedure that states can use to develop jurisdiction-specific SPFs.

The guidebook starts with a brief overview of other documents being developed by FHWA and NCHRP to facilitate the implementation of the HSM. This is followed by a brief discussion of "What are SPFs". This is then followed by a discussion on how SPFs are used for different applications, i.e., network screening, project level analysis, and determining the safety effect of improvements – examples are provided to illustrate these three applications.

Next, there is a discussion of the statistical issues associated with the development of jurisdiction-specific SPFs. This section was written for readers with more than a basic understanding of statistics. We feel that such a discussion is essential and serves as a useful introduction to the next section that discusses the steps involved in developing SPFs. The following statistical issues are discussed:

- Overdispersion
- Selection of Explanatory Variables
- Functional form of the Model and the Explanatory Variables
- Overfitting of SPFs
- Correlation among Explanatory Variables
- Homogenous Segments and Aggregation
- Presence of Outliers
- Endogenous Explanatory Variables
- Estimation of SPFs for Different Crash Types and Severities
- Goodness of Fit

This is followed by a step-by-step approach that can be used to develop jurisdiction-specific SPFs. The steps are presented in the context of the purpose (or application) of the SPF, i.e., network screening, project level analysis, and determining the safety effect of improvements. The following steps are discussed:

- Step 1 – Determine use of SPF
- Step 2 – Identify facility type
- Step 3 – Compile necessary data
- Step 4 – Prepare and cleanup database

- Step 5 – Develop the SPF
- Step 6 – Develop the SPF for the base condition
- Step 7 – Develop CMFs for specific treatments
- Step 8 – Document the SPFs

Recent developments in SPF development are discussed next. The following topics are discussed:

- Variance of Crash Estimates Obtained from SPFs
- Temporal and Spatial Correlation
- Other Model Forms
- Generalized Additive Models
- Random-parameters Models
- Bayesian Estimation Methods

A brief overview of software tools is provided in the last section. The guidebook concludes with a list of references.

1. Background and Context

This is a "how-to" guidebook for states that are developing jurisdiction-specific safety performance functions (SPFs). The guidebook discusses the issues associated with the development of jurisdiction-specific SPFs and provides a step-by-step procedure that states can use to develop jurisdiction-specific SPFs.

This document is part of a series of documents currently being developed by the Federal Highway Administration (FHWA) and the National Cooperative Highway Research Program (NCHRP) to facilitate the implementation of the HSM by the States. The other documents being currently prepared as part of the series include:

- *Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs* (hereafter referred to as the ***SPF Decision Guide***) (Srinivasan et al., 2013). This guidebook is intended to provide guidance on whether an agency should calibrate the safety performance functions (SPFs) from the Highway Safety Manual (HSM) (AASHTO, 2010) or develop jurisdiction-specific SPFs. The guidebook discusses the factors that need to be considered while making the decision. It is intended to be of use to practitioners at state and local agencies and to researchers.
- *User's Guide to Develop Highway Safety Manual Safety Performance Function Calibration Factors* (hereafter referred to as the ***SPF Calibration Guide***). This guide is being developed through NCHRP Project 20-7 (Task 332) by Dr. Geni Bahar of NAVIGATS. This document will provide guidelines to assist an agency in developing statistically sound calibration factors. This document will also provide guidance for assessing the quality of a calibration factor after it is developed.
- *SPF Needs Assessment*: This project, led by Volpe did a needs and alternatives assessment to determine the set of potential resources that would best satisfy the future needs of the states using SPFs. Part of this project involved conducting interviews with selected States to better understand their needs and requirements regarding SPFs. Further information about this effort can be obtained from http://safety.fhwa.dot.gov/rsdp/.

Organization of this Document

The next section of this document gives a brief discussion of "What are SPFs". This is followed by a discussion on how SPFs are used. Next, there is a discussion of the statistical issues associated with the development of jurisdiction-specific SPFs. This is followed by a step-by-step approach that can be used to develop jurisdiction-specific SPFs. Recent developments in SPF development are discussed next followed by a brief overview of statistical tools. The report concludes with a list of references.

2. What Are SPFs?

SPFs are crash prediction models. They are essentially mathematical equations that relate the number of crashes of different types to site characteristics. These models always include traffic volume (AADT) but may also include site characteristics such as lane width, shoulder width, radius/degree of horizontal curves, presence of turn lanes (at intersections), and traffic control (at intersections). One example is the following SPF from *Safety Analyst* for predicting the total number of crashes on rural multilane divided roads:

$$P = L \times e^{-5.05} \times (AADT)^{0.66}$$

Where:

P is the total number of crashes in 1 year on a segment of length L.

The primary purpose of this SPF from *Safety Analyst* is to assist an agency in their network screening process, i.e., to identify sites that may benefit from a safety treatment. This is a relatively simple SPF where the predicted number of crashes per mile is a function of just AADT. On the other hand, Bauer and Harwood (2012) provide a more complex prediction model for fatal and injury crashes on rural two lane roads:

$$N_{FI} = \exp\left[b_0 + b_1 \ln(AADT) + b_2 G + b_3 \ln\left(2 \times \frac{5730}{R}\right) \times I_{HC} + b_4 \left(\frac{1}{R}\right)\left(\frac{1}{L_C}\right) \times I_{HC}\right]$$

Where:

N_{FI}	=	fatal-and-injury crashes/mi/yr
AADT	=	veh/day
G	=	absolute value of percent grade; 0 percent for level tangents; ≥ 1 percent otherwise
R	=	curve radius (ft); missing for tangents
I_{HC}	=	horizontal curve indicator: 1 for horizontal curves; 0 otherwise
L_C	=	horizontal curve length (mi); not applicable for tangents
ln	=	natural logarithm function
$b_0,\ldots,b_4$	=	regression coefficients

This prediction model was estimated using data from Washington. The goal of this model was to examine the safety aspects of horizontal and vertical curvature.

3. How Are SPFs Used?

The Highway Safety Manual (HSM) outlines at least three different ways in which SPFs can be used by jurisdictions to make better safety decisions. One application is to use SPFs as part of network screening to identify sections that may have the best potential for improvements (i.e., Part B of the HSM). The second application is to use SPFs to determine the safety impacts of design changes at the project level (i.e., Part C of the HSM). The third application is the use of SPFs in determining the safety effects of engineering treatments. Following is a brief discussion of these applications.

Network Screening Level: Identifying Locations with Promise

SPFs may be used to identify locations with promise, which are locations that may benefit the most from a safety treatment. This application is also referred to as network screening. Here, SPFs can be used to estimate the predicted number of crashes for a particular facility type with a particular traffic volume. SPFs that could be used for network screening are available in *Safety Analyst*, a set of software tools that can be used by state and local agencies for safety management. An example of how to apply an SPF in networks screening is shown here.

Example calculation for using an SPF in network screening

Suppose the goal is to determine if a section (with the following characteristics) should be selected as a "site with promise" for further review as part of network screening:

- Roadway type: Rural 2-lane road
- Section Length = 2 mi
- Average AADT in the last 3 years = 5,000 vpd
- Total crashes in 3 years = 31

The first step is to select the appropriate SPF for this roadway type. One example is the following SPF from *Safety Analyst* for predicting total crashes on rural two- lane roads:

$$P = C \times L \times e^{-3.63} \times (AADT)^{0.53}; \quad k = 0.50$$

where P is the total number of crashes in 1 year on a segment of length L, C is the calibration factor, and k is the overdispersion parameter of the SPF[1].

This SPF was estimated using HSIS data from the state of Ohio. Estimating the calibration factor C through the calibration process is necessary because "the general level of crash frequencies may vary substantially from one jurisdiction to another for a variety of reasons including crash reporting thresholds and crash reporting system procedures" (HSM, page C-18). Within *Safety Analyst*, calibration is done automatically (as discussed in Section 1, the ***SPF Calibration Guide***, being developed under

[1] Here, the overdispersion parameter k is a constant. k can also be estimated on a per-mile basis. Further discussion of the overdispersion parameter and other statistical issues associated with SPF development are provided in Section 4.

NCHRP Project 20-07, will provide guidelines to assist an agency in developing statistically sound calibration factors).

For this example, let us assume that the calibration factor is 1.1. Since 3 years of data are available for network screening, the predicted number of crashes (P) in 3 years using the SPF is:

$$P = 3 \times 1.1 \times 2 \times e^{-3.63} \times (5000)^{0.53} = 15.98$$

Part B of the HSM discusses different ways in which the predicted SPF value can be used in network screening. One approach is to rank sites in decreasing order based on the difference between the observed number of crashes (31) and the predicted number of crashes from the SPF (15.98) – this is done in the *level of service* method.

Another approach is to estimate the expected number of crashes in the section while accounting for possible bias due to regression to the mean (RTM). One way to do this is to use the empirical Bayes (EB) method. To implement the EB, the first step is to estimate the weight, w_1, as follows:

$$w_1 = \frac{1}{1 + k \times P} = \frac{1}{1 + 0.5 \times 15.98} = 0.111$$

The EB estimated number of crashes is then calculated as:

$$E = P \times w_1 + A \times (1 - w_1)$$

where A is the observed number of crashes for the section under consideration. In our case A = 31, and thus E becomes:

$$E = 15.98 \times 0.111 + 31 \times (1 - 0.111) = 29.33$$

Once the EB estimate of the expected crashes is obtained, sites could be ranked based on the expected crashes per mile or the difference between the expected and predicted number of crashes per mile. Further discussion of the use of SPFs in network screening can be found in Part B of the HSM.

Instead of calibrating existing SPFs, an agency may choose to develop their own SPFs in order to improve the accuracy of the predictions. The advantages of jurisdiction-specific SPFs are also discussed in previous studies (e.g., Lu et al., 2012; Sacci et al., 2012) and in the HSM. As discussed earlier, this guidebook provides guidance for agencies who want to estimate jurisdiction-specific SPFs. This guidance is available from Sections 4, 5, 6, and 7 of this guidebook.

Project Level: Determining the Expected Safety Impacts of Design Changes

When SPFs are used in project-level decision making, they are used for estimating the *average predicted/expected crash frequency* for existing conditions, alternatives to existing conditions, or

proposed new roadways. Part C of the HSM provides methods for estimating the average predicted/expected crash frequency of a site or project.

The HSM provides prediction methods for the following road types:

Roadway Segments

- Rural two-lane roads
- Rural four-lane divided and undivided roads
- Two-lane, three-lane (with center TWLTL), four-lane divided, four-lane undivided, and five-lane roads (with center TWLTL) in urban and suburban arterials

Intersection Types

- Three- and four-leg minor road stop-controlled and four-leg signalized intersections on rural two-lane roads
- Three- and four-leg minor road stop-controlled and four-leg signalized intersections on rural four-lane roads
- Three- and four-leg minor road stop-controlled and signalized intersections on urban and suburban arterials

The predictive method in Part C of the HSM is an 18-step procedure to estimate the average expected crash frequency at a site. A site in the HSM is defined as an intersection or a homogenous roadway segment. The predictive method utilizes SPFs that were developed from observed crash data for a number of similar sites. The method uses three components to predict the average expected crash frequency at a site:

(1) the base model, called a safety performance function (SPF);

(2) the crash modification factors (CMFs) to adjust the estimate for additional site specific conditions, that may be different from the base conditions; and

(3) a calibration factor to adjust the estimate for accuracy in the state or local area (as mentioned earlier, the procedure to estimate a calibration factor is provided in the Appendix to Part C of the HSM and will be further discussed in the upcoming ***SPF Calibration Guide***).

These components are combined in the general form below:

$$N_{predicted} = N_{spf} \times (CMF_{1x} \times CMF_{2x} \times \ldots \times CMF_{yz}) \times C_x \quad (3.1)$$

where:

$N_{predicted}$ = predicted average crash frequency for a specific year for site type x;
N_{spf} = predicted average crash frequency determined for base conditions of the SPF developed for site type x;
CMF_{nx} = crash modification factors specific to SPF for site type x; and
C_x = calibration factor to adjust SPF for local conditions for site type x.

As indicated, each predictive model is specific to a facility or site type (e.g., urban four-lane divided segments) and a specific year. It should be noted that the predictive method can be used to predict crashes for past years based on observed AADT or for future years based on forecast AADT.

The steps for the predictive method are presented in detail in Section C.5. of Volume 2 of the HSM. In short, they are:

- Decide which facilities and roads will be used in the predictive process and for what period of time (Steps 1 and 2)
- Identify homogenous sites and assemble geometric conditions, crash data, and AADT data for the sites to be used (Steps 3 through 8)
- Apply the safety performance function, any applicable crash modification factors, and a calibration factor if available (Steps 9 through 11)
- Apply site- or project-specific empirical Bayes method if applicable (Steps 12 through 15)
- Repeat for all sites and years, sum, and compare results (Steps 16 through 18)

An example of how to apply an SPF, CMFs, and calibration factor for the predictive method is shown next.

Example calculation of average expected crash frequency using HSM predictive method

This example demonstrates how to use the HSM predictive method to calculate the expected average crash frequency for a rural four-lane divided roadway segment with the following characteristics (this example was taken from Section 2 of Srinivasan and Carter, 2011):

- 1.0-mi segment
- 12-ft lane
- 6-ft paved right shoulder
- AADT of 15,000 vpd
- 80-ft traversable median with no barrier
- No roadway lighting
- No automated enforcement

All tables, equations, and page numbers in the example below refer to Chapter 11 of the HSM.

Steps 1 through 8

Since this example is directed at applying the predictive method to a single pre-selected segment with existing data, steps 1 through 8 are not necessary.

Step 9: Apply the appropriate safety performance function (SPF)

The SPF for a rural divided roadway segment is presented in Equation 11-9 (p. 11-18) in the HSM with coefficients listed in Table 11-5.

$$N_{spf\ rd} = e^{(a + b \times \ln(AADT) + \ln(L))}$$

where:

$N_{spf\ rd}$ = base total number of roadway segment crashes per year;

AADT = annual average daily traffic (vehicles/day) on roadway segment;
L = length of roadway segment; and
a, b = regression coefficients (appropriate values to be selected from Table 11-5)

Using the SPF for this example yields the following prediction:

$N_{spf\ rd} = e^{(-9.025 + 1.049 \times \ln(15000) + \ln(1.0))} = 2.892$ crashes per year

Step 10: Apply the appropriate crash modification factors

The HSM procedure for rural divided roadways involves five CMFs.

Lane Width (CMF_{1rd})
Based on Table 11-16 for a lane width of 12 feet, $CMF_{1rd} = 1.0$.

Right Shoulder Width (CMF_{2rd})
Based on Table 11-17 for a right shoulder width of 6 feet, $CMF_{2rd} = 1.04$.

Median Width (CMF_{3rd})
Based on Table 11-18 for a median width of 80 feet, $CMF_{3rd} = 0.95$.

Lighting (CMF_{4rd})
Since there is no roadway lighting at this location, $CMF_{4rd} = 1.0$ (the base condition for CMF_{4rd} is absence of lighting).

Automated Enforcement (CMF_{5rd})
Since there is no automated enforcement at this location, $CMF_{5rd} = 1.0$ (the base condition for CMF_{5rd} is absence of automated enforcement).

Combined CMF
The combined CMF value is calculated below.
$CMF_{comb} = 1.0 \times 1.04 \times 0.95 \times 1.0 \times 1.0 = 0.99$

Step 11: Apply a calibration factor if available

For this example, the calibration factor (C_r) for the local area is assumed to be 0.96.

Calculation of Average Expected Crash Frequency

$N_{predicted\ rd} = N_{spf\ rd} \times CMF_{comb} \times C_r$
$= 2.892 \times 0.99 \times 0.96 = 2.75$ crashes per mile per year

As discussed earlier, instead of calibrating existing SPFs, an agency may choose to develop their own SPFs in order to improve the accuracy of the predictions. The HSM indicates that jurisdiction-specific SPFs "are likely to enhance the reliability of the Part C predictive method" (HSM, page A-9).

Appendix A to Part C of the HSM outlines two possible approaches for developing jurisdiction-specific SPFs for project level analysis to make use of the Part C prediction methodology. One option is for the

SPF to be developed using only data that represent the base conditions (defined for each SPF in Chapters 10, 11, and 12). However, in many cases, there may not be a sufficient number of sites with the specific base conditions to estimate an SPF. Under those circumstances, SPFs are estimated with "all variables that are part of the applicable base-condition definition, but have non-base-condition values". Then, "the initial model should be made applicable to the base conditions by substituting values that correspond to those base conditions into the model" (HSM, page A-10). As discussed earlier, this guidebook provides guidance for agencies that want to estimate jurisdiction-specific SPFs.

Evaluating the Effect of Engineering Treatments

Researchers commonly conduct safety evaluation studies to determine the effect on crashes (e.g., estimate CMFs) from implementing some safety countermeasure. Observational studies to develop CMFs can be broadly classified into before-after studies and cross-sectional studies. Before-after studies include "all techniques by which one may study the safety effect of some change that has been implemented on a group of entities (road sections, intersections, drivers, vehicles, neighborhoods, etc.)" (Hauer, 1997, p. 2). On the other hand, cross-sectional studies include those where "one is comparing the safety of one group of entities having some common feature (say, STOP controlled intersections) to the safety of a different group of entities not having that feature (say, YIELD controlled intersections), in order to assess the safety effect of that feature (STOP versus YIELD signs)" (Hauer, 1997, p. 2-3). Since in a typical before-after study, one is dealing with the same roadway unit located in a particular location used by probably the same users in both the before and after periods, it is less likely to be prone to confounding (Elvik , 2011). Hauer (2010) discussed the use of before-after and cross-sectional studies to estimate CMFs in the setting of a case study, namely railroad crossings where crossbucks were replaced by flashers. Hauer (2010) found that unlike in before-after studies, the results of CMFs estimated from cross-sectional studies were not consistent with each other, and concluded that at this time, cross-sectional regression cannot be relied upon to capture cause and effect, and hence the CMFs from these types of studies are not very reliable. Further discussion of the issues associated with the development of CMFs from cross-sectional and before-after studies can be found in Elvik (2011) and Carter et al., (2012). Following is a brief discussion of how SPFs can be used in estimating CMFs from before-after and cross-sectional studies.

Use of SPFs in before after studies

Many engineering treatments are implemented at locations that may have a higher than normal crash count, therefore, before-after studies need to account for potential bias due to RTM. One way to address this bias is to make use of the empirical Bayes (EB) procedure developed by Hauer (1997). SPFs are an integral part of implementing the EB method.

Following are the steps used to conduct a before-after evaluation using the EB method:

1. Identify a reference group of sites that are similar to the sites that are treated, but without the treatment.
2. Estimate an SPF using the data from the reference sites

3. Estimate the EB expected crashes for the before period of the treatment group by combining the observed crashes from the before period of the treatment sites with the predicted crashes for the before period based on the SPFs
4. Estimate the EB expected crashes in the after period of the treatment group (had the treatment not been implemented) and the variance of this estimate
5. Using the observed number of crashes experienced by the treatment sites in the after period along with EB expected crashes in the after period (and its variance), estimate the CMF and the standard error of the CMF.

Further discussion of this application can be found in Hauer (1997) and Gross et al. (2010). An example illustrating the use of SPFs in implementing the EB method for before-after evaluation can be found in Section 5 of Srinivasan and Carter (2011).

Estimating CMFs directly from SPFs

The coefficients of the variables from SPFs can be used to estimate the CMF associated with a particular treatment. For example, suppose the intent is to estimate the CMF for shoulder width based on the following SPF which was estimated to predict the number of crashes per mile per year on rural two-lane roads in mountainous roads with paved shoulders (Appendix B of Srinivasan and Carter, 2011):

$$Y = \exp[0.8727 + 0.4414 \times \ln(AADT/10000) + 0.4293 \times (AADT/10000) - 0.0164 \times SW]$$

where AADT is the annual average daily traffic and SW is the width of the paved shoulder, in feet. If the intent is to estimate the CMF of changing the shoulder width from 3 to 6 ft, then the CMF can be estimated as the ratio of the predicted number of crashes when the shoulder width is 6 ft to the predicted number of crashes when the shoulder width is 3 ft:

$$CMF = \frac{\exp[0.8727 + 0.4414 \times \ln(AADT/10000) + 0.4293 \times (AADT/10000) - 0.0164 \times 6]}{\exp[0.8727 + 0.4414 \times \ln(AADT/10000) + 0.4293 \times (AADT/10000) - 0.0164 \times 3]}$$

This ratio simplifies to:

$$CMF = \exp[-0.0164 \times (6-3)] = 0.952$$

4. Statistical Issues Associated with the Development of Jurisdiction-Specific SPFs

This section provides a discussion of the statistical issues associated with the development of jurisdiction-specific SPFs. This section was written for readers with more than a basic understanding of statistics. It assumes that the reader is familiar with the basic principles of modeling and associated assumptions in general and most importantly understands how the selection of a group of sites (i.e., their characteristics, traffic volumes, and crash experience) affects the estimation and appropriateness of an SPF. We feel that such a discussion is essential and serves as a useful introduction to the next section that discusses the steps involved in developing SPFs.

As discussed earlier, SPFs are crash prediction models that relate crash frequency to site characteristics. In other words, the dependent variable in the equation is the number of crashes of a specific type. Crashes are examples of "count data" and are properly modeled using a specific family of statistical models called count data models. The most popular count data models for rare events are Poisson and negative binomial regression models.

To illustrate the principles of a Poisson regression model, consider the number of crashes occurring per year at a site (i.e., roadway segment or intersection) (Washington et al., 2011). In a Poisson regression model, the probability of site *i* having y_i crashes per year is given by:

$$P(y_i) = \frac{\exp(-\lambda_i) \times \lambda_i^{y_i}}{y_i!} \quad (4.1)$$

where: λ_i is the Poisson parameter for site *i*, which is equal to site *i*'s expected number of crashes per year, $E(y_i)$. In Poisson regression models, the intent is to express the expected number of crashes as a function of site characteristics. In other words, $\lambda_i = f(\beta X_i)$, where f is a function, X_i is a vector of explanatory variables, and β is a vector of estimable parameters (coefficients of X_i). The most common relationship between the explanatory variables and λ_i is the following:

$$\lambda_i = \exp(\beta X_i) \text{ or } \ln(\lambda_i) = \beta X_i \quad (4.2)$$

This relationship is also called a log-linear model. One reason the log-linear model for counts is popular is because it ensures that the Poisson parameter (i.e., expected number of crashes during a time period) is always positive. Another reason is that taking the log on both sides of the equation results in a linear combination of the predictor variables (i.e., the Xs) on the right-hand side. This type of model form belongs to a category of models called generalized linear models (GLM). In a GLM, the regression coefficients and their standard error are typically estimated by maximizing the likelihood or log likelihood of the parameters for the data observed; this is called estimating by the maximum likelihood

method (the well-known monograph by McCullagh and Nelder, 1989, is a standard reference for generalized linear models).

Overdispersion

The Poisson distribution restricts the mean and variance to be equal, i.e., $E(y_i) = VAR(y_i)$. Often, with crash data, $VAR(y_i) > E(y_i)$, leading to overdispersion (there have been a few examples with crash data where $VAR(y_i) < E(y_i)$, leading to underdispersion; interested readers can refer to Lord and Mannering, 2010, for further discussion of this issue). One way to account for overdispersion is to model crash counts using a negative binomial regression model, which can be written as follows:

$$\lambda_i = f(\beta X_i) \times \exp(\varepsilon_i) \tag{4.3}$$

where ε_i is a gamma-distributed disturbance term.

If a log-linear model is assumed, then

$$\lambda_i = \exp(\beta X_i) \times \exp(\varepsilon_i) = \exp(\beta X_i + \varepsilon_i) \tag{4.4}$$

By introducing the disturbance term, the variance is now larger than the mean, and can be shown to be:

$$VAR(y_i) = E(y_i) + k \times [E(y_i)]^2 \tag{4.5}$$

This form of the negative binomial regression model has been called as the NB2 model by Cameron and Trivedi (1998). In the above equation, k is the overdispersion parameter. Some studies (e.g., Hauer et al., 2002) prefer to deal with the reciprocal of the overdispersion parameter rather than the overdispersion parameter. If ϕ is used to denote the reciprocal of the overdispersion parameter, then $\phi = 1/k$. In that case, equation 4.5 can be rewritten as:

$$VAR(y_i) = E(y_i) + \frac{[E(y_i)]^2}{\phi} \tag{4.6}$$

If k is zero, then the negative binomial regression model reduces to a Poisson regression model. As in the case of the Poisson regression, the coefficients of X_i can be estimated using standard maximum likelihood methods.

The most common approach is to assume that the overdispersion parameter is a constant. However, when roadway segments are modeled, Hauer (2001) maintains that "if one assumes that the same overdispersion parameter applies to all road sections in the data base, then, the maximum likelihood estimate of parameters will be unduly influenced by very short road sections and insufficiently influenced by long road sections", and suggested that "a way to avoid both problems is to estimate an overdispersion parameter that applies to a unit length of road". In other words, $k = \frac{k_1}{L}$, where k_1 is the

overdispersion parameter for a unit length of road and *L* is the length of a particular segment. Note that when the reciprocal of the overdispersion parameter is used (instead of *k*), then $\phi = \phi_1 L$. More recently, Cafiso et al. (2010) found the overdispersion parameter to be inversely related to segment length for rural two-lane roads. Alternatively, the overdispersion parameter itself can be modeled as a function of site characteristics including section length (e.g., see Miaou and Lord, 2003; Mitra and Washington, 2007).

Selection of Explanatory Variables

The selection of explanatory (independent) variables is an important step in the development of SPFs. The list of explanatory variables may depend on the proposed application of the SPF. As discussed in Srinivasan et al. (2013), for SPFs that are used for network screening, for each facility type, the number of crashes for each unit (intersection, segment, or ramp), along with the traffic volume (AADT) associated with that unit are required. For roadway segments and ramps, the segment length will be required as well. For intersections of any type, it is recommended that AADT for both major and minor roads be available (there has been some discussion in the highway safety research community on whether SPFs for network screening will be significantly improved by including other explanatory variables in addition to traffic volume and segment length – Srinivasan et al., (2011) provides some discussion on this topic). The list of roadway, ramp, and intersection types that could be considered is available from Appendix A of the ***SPF Decision Guide*** (Srinivasan et al., 2013). If SPFs are to be estimated for a particular crash type or severity, the number of crashes by severity and type will be needed for each unit.

With respect to project-level SPFs, Appendix A to Part C of the HSM outlines two possible approaches for developing SPFs. Following is a quote from page A-10 of the HSM:

> "Two types of data sets may be used for SPF development. First, SPFs may be developed using only data that represent the base conditions, which are defined for each SPF in Chapter 10, 11, and 12. Second, it is also acceptable to develop models using data for a broader set of conditions than the base condition. In this approach, all variables that are part of the applicable base-condition definition, but have non-base-condition values, should be included in the initial model. Then, the initial model should be made applicable to the base conditions by substituting values that correspond to those base conditions into the model".

With either approach, detailed information is necessary about the site characteristics in addition to traffic volume so that it can be determined whether the characteristics of the site correspond to the base condition. A good starting point is the list of variables that are considered required by the HSM for calibrating the Part C prediction models. This list for the different facility types is available in the Appendix to Part C of the HSM.

In practice, it may not be possible to include all the relevant independent variables that could potentially have an impact on safety. If the independent variables that are not included in the model are correlated with independent variables that are in a model, then it can lead to *omitted variable bias*. For example, if the intent is to estimate the safety effects of chevrons on horizontal curves and the curves with

chevrons also have the worst roadside hazards, but the information on the roadside hazards is not included in the model (because it was not collected), then a prediction model may incorrectly conclude that chevrons are associated with an increase in crashes. For this reason, Harwood et al., (2000) indicate that "regression models are very accurate tools for predicting the expected total accident experience for a location or class of locations, but they have not proved satisfactory in isolating the effects of individual geometric or traffic control features". However, there are situations where SPFs may be the only reasonable option available to determine the safety effect of an individual geometric or traffic control characteristics. For further discussion of the use of SPFs for determining the safety of individual geometric or traffic control characteristic, readers are referred to Carter et al. (2012).

Functional Form of the Model and the Explanatory Variables

Relationship between Traffic Volume and Crashes

Since traffic volume is usually the most important contributor to crashes, the relationship between traffic volume and crashes is discussed first. Let us assume that an SPF predicts the number of crashes for a roadway segment for the purpose of network screening and includes segment length and the annual average daily traffic (AADT). The most common form that has been used for this purpose is the following:

$$Y = L \times \exp[a + b \times \ln(AADT)] = L \times e^{a} \times (AADT)^{b} \quad (4.7)$$

where: Y is the expected number of crashes on a segment, L is the length of the segment, and a and b are regression coefficients to be estimated. This type of model is sometimes called a power function. This form is common because it is simple and also satisfies the boundary condition that if there is no traffic (i.e., if AADT is zero), then the expected number of crashes should be zero. In fact, the SPFs in *Safety Analyst* for roadway segments and ramps use this form.

In the power model, it is generally accepted that b is positive since the number of crashes are expected to increase with increase in traffic volume. Depending on whether b is less than 1, equal to 1, or greater than 1, Figure 4.1 shows the shape of the relationship between the expected number of crashes and traffic volume.

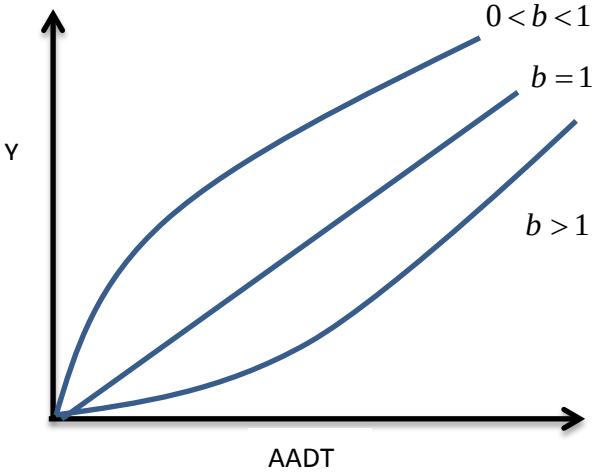

Figure 4.1: Shape of the Relationship between Crash Frequency and AADT as a Function of the Power, b

In discussing the underlying relationship between crash frequency and traffic volume, Kononov et al., (2011) state that the power model may not be the most appropriate functional form. As a starting point, Kononov et al. (2011) used neural networks (NN) to explore the underlying relationship between crash frequency and AADT using data from urban freeways from California and Colorado. The use of NN seemed to indicate that the relationship could be sigmoidal (Figure 4.2).

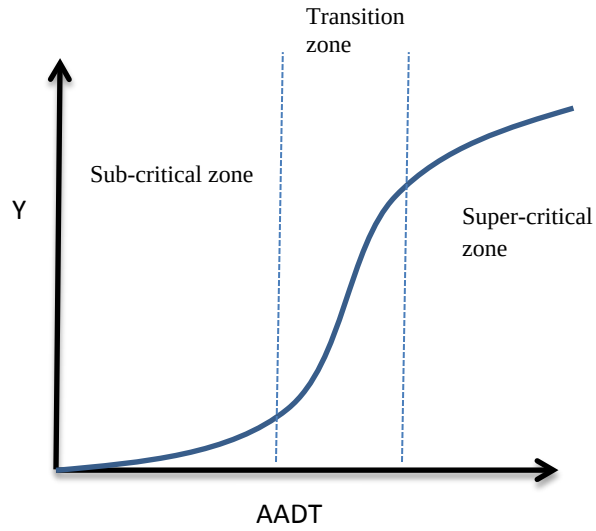

Figure 4.2: Sigmoidal Relationship between Crash Frequency and Traffic Volume (Kononov et al., 2011)

Kononov et al. (2011) then used the parameters from the Highway Capacity Manual (HCM) to explain the underlying relationship, by dividing the graph into three zones: sub-critical, transitional, and super critical zones. Kononov et al. (2011) maintain that in the sub-critical zone (i.e., when freeways are not congested and traffic density is low), crashes increase gradually with AADT. However, in the transition zone, the crashes start increasing rapidly. Finally, in the super-critical zone, crashes still increase with AADT, but only gradually. Based on the HCM, Kononov et al. (2011) found that when the AADT increases from the sub-critical to the transition zone, operating speeds remain almost the same, but with a significant increase in traffic density, and this may be a possible reason for the rapid increase in crashes with AADT in the transition zone. In the super-critical zone, speeds start dropping, possibly leading to only a gradual increase in crashes with AADT.

Hauer (2004) argued that the simplistic power function (discussed earlier) may not be appropriate, especially for single vehicle crashes. The probability of a single-vehicle crash has been shown to decrease with increasing traffic volume (e.g., see Qin et al., 2006). Hence a functional from that allows a peak/valley and a point of inflection may be more appropriate for modeling the relationship between single-vehicle crashes and traffic volume – based on Figure 4.1, the simple power model from equation 4.7 does not allow a peak/valley or a point of inflection.

Relationship between Segment Length and Crashes

In equations 4.7, segment length is included as a multiplier, i.e., if N crashes are expected to occur on 1 mi of a roadway, one would expect 2N crashes to occur on an identical roadway segment that is 2 mi long. However, if instead of assuming it as a simple multiplier, segment length could be included as another independent variable with its own coefficient that would be estimated along with the other coefficients. For example, equation 4.7 could be rewritten as:

$$Y = e^{a} \times L^{c} \times (AADT)^{b} \quad (4.8)$$

where c is a parameter to be estimated as part of the modeling process. If the estimate of c is close to 1, then equation 4.8 reduces to equation 4.7. However, in many situations, c may be significantly different from 1. There may be many reasons for this situation. For example, shorter segment lengths may be associated with corridors where there are more driveways per mile and intersections are closer together, and if these variables are not included in the model, then segment length may become a surrogate for variables that were omitted from the model.

Identifying the Relevant Individual Explanatory Variables and Determining the Most Appropriate Functional Form

There are two interrelated components to defining the relationship between crash counts and the explanatory variables: one is the choice of variables (and interactions, if necessary) from the many available and the other is the mathematical equation or functional form that relates the two. This aspect of modeling is undoubtedly the most challenging part in developing SPFs that are not merely a function of traffic volume and segment length. Before going into a discussion of approaches for identifying the explanatory variables and the most appropriate functional form, following is a discussion of interaction and its importance in the modeling process.

<u>What is Interaction?</u>

Interaction occurs when the effect of an independent variable on crash frequency depends on the values of another independent variable. To illustrate, first consider an SPF that includes segment length, AADT, lane width, and shoulder width, without interaction terms:

$$Y = L \times \exp[a + b \times \ln(AADT) + c \times LW + d \times SW] \quad (4.9)$$

where: LW is the lane width, SW is the shoulder width, and a, b, c, and d, are regression parameters to be estimated as part of the modeling process. Equation 4.9 can also be written as follows:

$$Y = L \times e^{a} \times (AADT)^{b} \times e^{cLW} \times e^{dSW} \quad (4.10)$$

If this SPF is used to determine the safety effect of changing from a lane width of LW_1 to a lane width of LW_2, then the CMF for changing a LW_1-ft lane to a LW_2-ft lane can be calculated as:

$$CMF = \frac{L \times \exp[a + b \times \ln(AADT) + c \times LW_2 + d \times SW]}{L \times \exp[a + b \times \ln(AADT) + c \times LW_1 + d \times SW]} \quad (4.11)$$

Equation 4.11 simplifies to:

$$CMF = \frac{\exp(c \times LW_2)}{\exp(c \times LW_1)} = \exp[c \times (LW_2 - LW_1)] \quad (4.12)$$

The CMF in this case is a function of just the parameter estimate c, LW_2, and LW_1. Alternatively, consider the following SPF that includes an interaction term between lane and shoulder width:

$$Y = L \times \exp[a + b \times \ln(AADT) + c \times LW + d \times SW + e \times LW \times SW] \quad (4.13)$$

where a, b, c, d, and e, are regression parameters to be estimated as part of the modeling process. Based on equation 4.13, the CMF of changing from LW_1 to LW_2 foot lanes will be the following:

$$CMF = \frac{L \times \exp[a + b \times \ln(AADT) + c \times LW_2 + d \times SW + e \times LW_2 \times SW]}{L \times \exp[a + b \times \ln(AADT) + c \times LW_1 + d \times SW + e \times LW_1 \times SW]} \quad (4.14)$$

Equation 4.14 simplifies to:

$$CMF = \frac{\exp(c \times LW_2 + e \times LW_2 \times SW)}{\exp(c \times LW_1 + e \times LW_1 \times SW)} = \exp[(LW_2 - LW_1) \times (c + e \times SW)] \quad (4.15)$$

It is clear that the CMF is not only a function of the parameter estimates c and e and LW_2 and LW_1, but also shoulder width (SW) because of the interaction between lane width and shoulder width shown in the SPF in equation 4.13. In this case, the CMF is a crash modification function, rather than a crash modification factor.

Interaction effects are not commonly found in SPFs probably because there is no easy way to identify which interactions are important and how they should be included in a model, unless there is some theoretical reason for including certain interactions. This does not imply that interactions do not exist or that they are not important. In fact, the following SPFs that were estimated for fatal and injury and PDO crashes using data from rural two-lane roads in Washington clearly indicate the interaction between curve radius and length of horizontal curves (Bauer and Harwood, 2012):

$$N_{FI} = \exp\left[b_0 + b_1 \ln(AADT) + b_2 G + b_3 \ln\left(2 \times \frac{5730}{R}\right) \times I_{HC} + b_4 \left(\frac{1}{R}\right)\left(\frac{1}{L_C}\right) \times I_{HC}\right]$$

$$N_{PDO} = \exp\left[b_0 + b_1 \ln(AADT) + b_2 G + b_3 \ln\left(2 \times \frac{5730}{R}\right) \times I_{HC} + b_4 \left(\frac{1}{R}\right)\left(\frac{1}{L_C}\right) \times I_{HC}\right]$$

where:

N_{FI}	=	fatal-and-injury crashes/mi/yr
N_{PDO}	=	PDO crashes/mi/yr
AADT	=	veh/day
G	=	absolute value of percent grade; 0 percent for level tangents; ≥ 1 percent otherwise

R	=	curve radius (ft); missing for tangents
I_{HC}	=	horizontal curve indicator: 1 for horizontal curves; 0 otherwise
L_C	=	horizontal curve length (mi); not applicable for tangents
ln	=	natural logarithm function
$b_0,\ldots,b_4$	=	regression coefficients

Identifying the Variables and Determining the Functional Form

In the type of model shown in equation 4.9 (or 4.10), a common approach for identifying significant variables from those available is a stepwise regression approach. The stepwise approach could be based on either a forward selection or a backward elimination procedure. Forward selection involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until no additional variable significantly improves the model at a predetermined significance level (e.g., 80 or 90 percent). Backward elimination involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, eliminating the variable (if any) that does not significantly degrade the model, and repeating this process until only variables that significantly contribute to the model remain. Examples of model comparison criteria include t-statistic, chi-square statistic, Akaike's information criterion (AIC), and Bayesian Information Criterion (BIC) (further discussion of some of the model comparison criteria can be found in the subsection entitled *Goodness of Fit* later in this section). The stepwise approach is popular because it can be readily implemented in many statistical software tools as long as the model is a generalized linear model (GLM).

When a simple model such as those in equations 4.9 and 4.10 is not satisfactory, then the user might want to investigate other forms of the explanatory variables to include in the model, such as the reciprocal of a variable, the variable to a certain power, or another transformation of a variable, but also two-way and perhaps three-way interactions between variables. Two example approaches are presented next.

Bauer and Harwood (2012) used a combination analysis of variance (ANOVA) and regression analysis approach to assess the functional form of the relationship between crash frequency and available parameters that can be summarized as follows:

1. except for AADT, categorize data into three groups, separately for each parameter; quantiles for continuous variables can typically be used
2. develop a crash prediction model including AADT and only the interaction of all categorized parameters
3. plot the safety effect of one parameter against the cell means of another parameter, encoding the data by the levels of the third parameter; if a four-way interaction was included, then multiple sets of plots will need to be generated
4. assess the shape of the relationships between safety effects and a given parameter across the levels of another parameter
5. assess whether these trends are consistent for a given model; if not, assess whether interactions exist

Based on the visual assessment of these relationships, if any, a final model can be developed using all parameters and relevant interactions on their original continuous scale. A stepwise approach is then used where first all parameters and interactions are included and then eliminating, one at a time, the least significant interaction(s) and then parameter(s).

Hauer (2004) argued that multiplicative models (such as in equation 4.9) may be appropriate to represent factors that apply to a stretch of road (e.g., lane width, shoulder width) but may not be appropriate to account for the influence of point hazards (e.g., such as driveways). Hauer (2004) suggested that a model that includes both multiplicative and additive forms may be more appropriate to study the influence of both point hazards and factors that apply to a stretch of road, such as:

Y = (scale parameter) × [multiplicative parameter + additive parameter] (4.16)

The multiplicative parameter could include variables such as traffic volume, lane width, and shoulder width. The additive portion may include traffic volume along with the number of driveways and number of short bridges (another possible approach to address the issue of point hazards such as driveways is to estimate separate SPFs for driveway-related crashes and non-driveway related crashes - this approach was used for estimating the SPFs for urban and suburban arterials that were included in Chapter 12 of the HSM).

In order to determine the appropriate functional form of the individual explanatory variables, Hauer and Bamfo (1997) introduced the Integrate-Differentiate (ID) method. This method was further discussed and developed in Hauer (2004). A summary of the steps involved in applying this method is provided next. For this illustration, it is assumed that the introduction of variable V is contemplated along with the appropriate functional form (some variables are already introduced in the model).

1. Divide V into groups (bins): V_1, V_2, etc. For example, if V represents lane width, V_1 may represent 10-ft lanes, V_2 may represent 11-ft lanes, etc.
2. For each group of sites within V, determine the total number of observed crashes ($N_o(V_i)$) and the total number of predicted crashes from the existing model ($N_p(V_i)$) that does not currently include the variable V (here, i represents the number of groups or bins in the variable V).
3. Calculate $R(V_i)$ and $\hat{\sigma}(R(V_i))$ (estimate of the standard deviation of $R(V_i)$). Depending on whether the variable is being introduced in the multiplicative part of the model or the additive part of the model (see equation 4.16), the formulas for $R(V_i)$ and its standard deviation are as follows:

   If V is introduced in the multiplicative part of the model:

$$R(V_i) = \frac{N_o(V_i)}{N_p(V_i)}$$

$$\hat{\sigma}(R(V_i)) = \frac{\sqrt{N_o(V_i)}}{N_p(V_i)}$$

If V is introduced in the additive part of the model:

$$R(V_i) = N_o(V_i) - N_p(V_i)$$

$$\hat{\sigma}(R(V_i)) = \sqrt{N_o(V_i)}$$

4. Plot R(V) versus V. If there is an orderly relationship between the two, then V can be introduced into the model with an appropriate functional form based on the shape of the relationship between R(V) with V. For guidance on selecting the appropriate mathematical function based on plots, readers can use websites, books, and articles on mathematical functions and curve fitting (e.g., http://functions.wolfram.com/).

This procedure can be applied as each variable is introduced into the model. Further discussion of alternative functional forms is available from Chapter 11 of Hauer (2013). The approach discussed in Hauer (2004) does not explicitly discuss the inclusion of interaction between the explanatory variables.

Cumulative Residual Plots

After a variable is included in the model and the parameters are estimated, Hauer (2004) recommends the use of cumulative residual (CURE) plots to obtain further insight into whether the selected appropriate functional form was reasonable. Following is a discussion of CURE plots and how they can be used. Suppose the goal is to develop a CURE plot to determine if the functional form used for AADT is appropriate. The first step is to create a data file that includes for each observation (i.e., segment or intersection) the AADT value and the residual from the SPF (the residual is the difference between the observed number of crashes and the predicted number of crashes from the SPF). Then this file is sorted in increasing order of AADT and the cumulative residuals are computed for each observation. The plot of the cumulative residual versus AADT is called a CURE plot. Following is an example of a CURE plot from Hauer and Bamfo (1997).

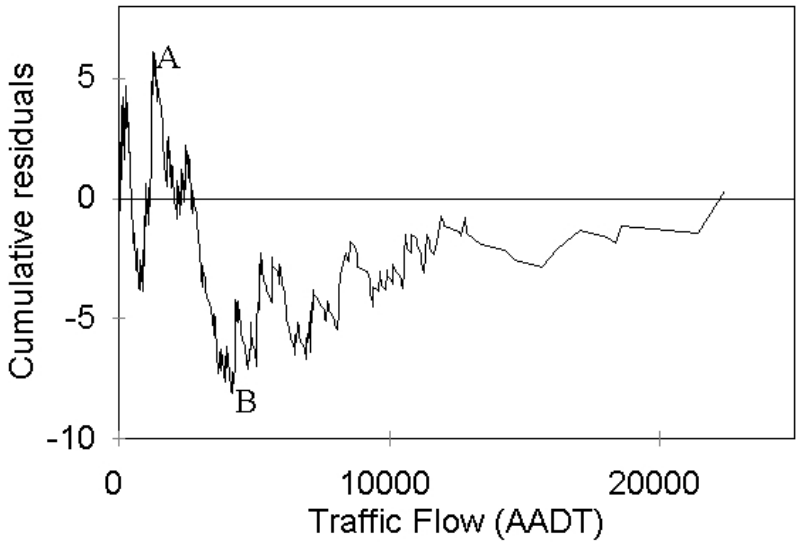

Figure 4.3: Example CURE Plot (Hauer and Bamfo, 1997)

The data in the CURE plot are expected to oscillate about 0. If the cumulative residuals are consistently drifting upward within a particular range of AADT, then it would imply that there were more observed than predicted crashes by the SPF. On the other hand, if the cumulative residuals are drifting downward within a particular range of AADT, then it would imply that there were fewer observed than predicted crashes by the SPF. Hauer and Bamfo (1997) also derived confidence limits for the plot ($\pm 2\sigma$) beyond which the plot should go only rarely. Following is a figure from Hauer and Bamfo (1997) that shows the CURE plot from figure 4.3 but with its confidence limits. This is an example of an acceptable CURE plot where the plot stays well within the confidence limits.

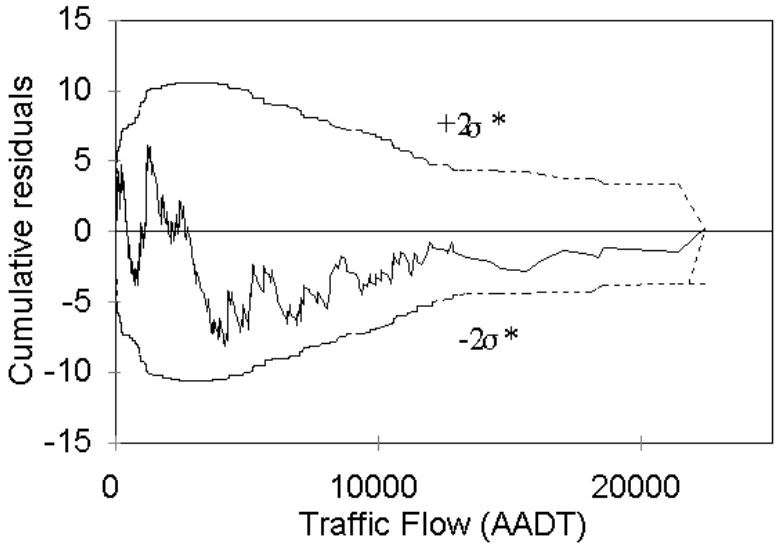

Figure 4.4: CURE Plot with Confidence Limits (Hauer and Bamfo, 1997)

In the context of CURE plots, it is important to recognize that the plot is not only a reflection of the functional form of the particular explanatory variable, but also whether other relevant explanatory factors have been included in the model in an appropriate form, i.e., the extent to which there is *omitted variable bias* (discussed earlier in subsection entitled *Selection of Explanatory Variables*). For example, Srinivasan and Carter (2011) estimated SPFs to predict the number of crashes in 1 year on rural two lane roads in North Carolina. The first SPF included AADT as the only explanatory variable – the resulting SPF was the following:

$$Y = L \times \exp[-4.0852 + 0.5830 \times \ln(AADT)]$$

For this SPF, the CURE plot for AADT is shown in Figure 4.5. Clearly, the plot is outside the confidence limits for a substantial range of AADT values, indicating that either the functional form of the SPF is not appropriate or that other important explanatory variables are not included in the model.

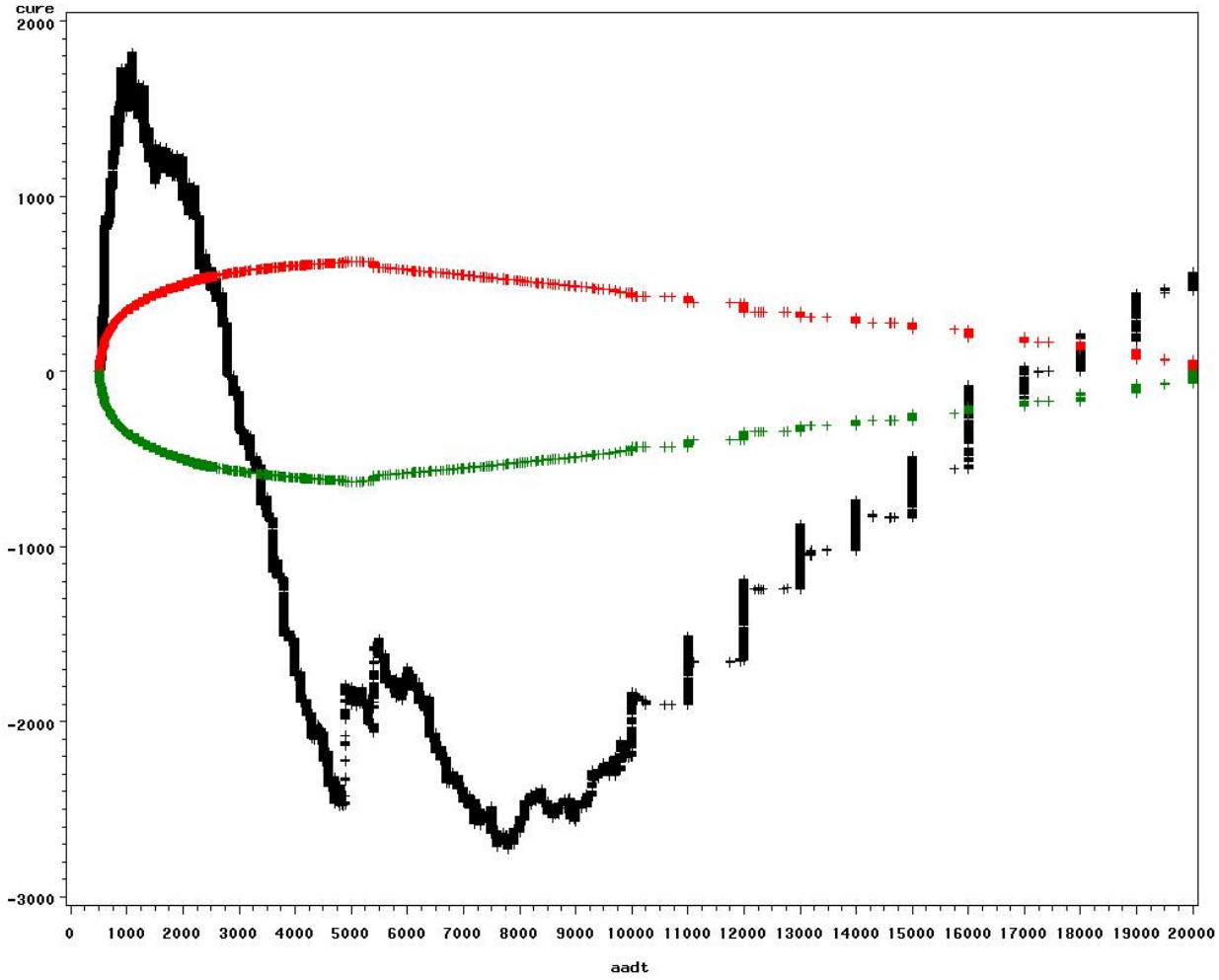

Figure 4.5: CURE plot for AADT based on AADT-only SPF

The SPF was modified by changing the functional form of AADT and including terrain, shoulder width, and shoulder type as other explanatory variables. Similar to the first SPF, the modified SPF was a log-linear model as well:

$$Y = L \times \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ..... \beta_n X_n)$$

Where β_0 is the intercept, X_1, X_2, X_3,…,X_n are the explanatory variables, and β_1, β_2, β_3,….., β_n, are the coefficient estimates. The various categories within terrain and shoulder type were included as individual specific indicator variables. Table 4.1 shows the coefficient estimates for the modified SPF:

Table 4.1: Coefficient Estimates for Modified SPF (Srinivasan and Carter, 2011)

Explanatory Variable		Coefficient estimate
Intercept		0.8727
ln(AADT/10,000)		0.4414
AADT/10,000		0.4293
Terrain	Flat	0.1264
	Rolling	0.1368
	Mountainous	0.0000
Shoulder Type	Unpaved	0.0354
	Paved	0.0000
Shoulder Width (in ft)		-0.0164

Based on this SPF, the average predicted total crashes in 1 year for a 1.5 mile rural two-lane road segment in rolling terrain with a 2-ft unpaved shoulder and an AADT of 1,500 vpd will be:

$$Y = 1.5 \times \exp\{0.8727 + 0.4414 \times ln(1{,}500/10{,}000) + 0.4293 \times (1{,}500/10{,}000) + 0.1368 + 0.0354 - 0.0164 \times 2\} = 1.905$$

Figure 4.6 shows the CURE plot for AADT for the improved SPF. A significant portion of the plot is now within the limits. It is clear that the alternate functional form for AADT along with the inclusion of the other independent variables provided an improved model. However, even with the improved model, there are portions of the CURE plot that are outside the limits (e.g., for AADT less than 5,000). Hence, it may be possible to further improve this SPF by modifying the functional form and including variables such as horizontal and vertical alignment.

Hauer (2004) also addresses the issue of the sequence in which variables could be added in a model. He suggests that traffic volume be introduced first since it is the dominant factor in terms of influence on crashes. This can be followed by including variables in the order in which they contribute to the increase in log-likelihood/parameter.

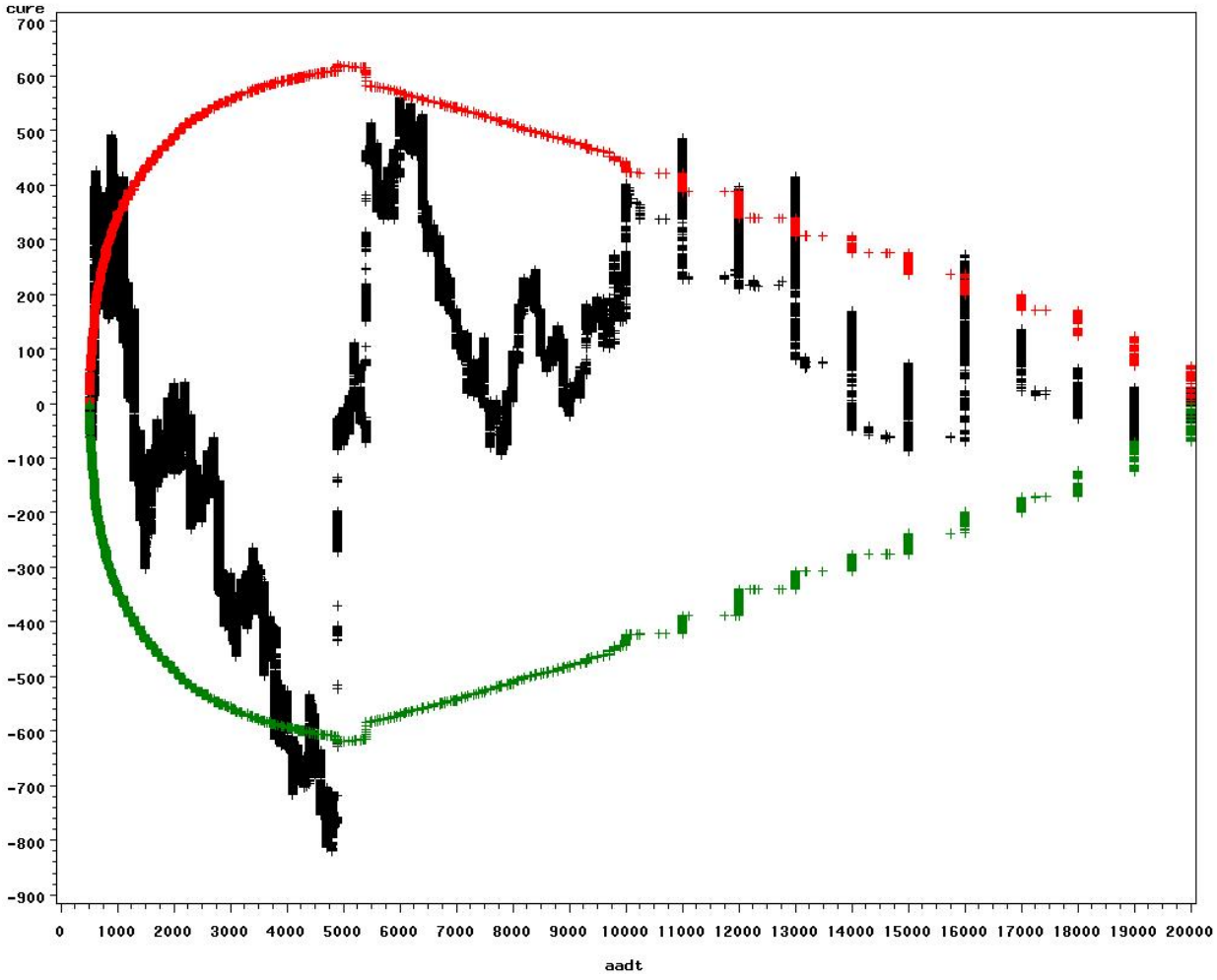

Figure 4.6: CURE plot for AADT for modified SPF

Overfitting of SPFs

Overfitting of SPFs may occur when too many parameters are included in the regression model. Even though they are found to be statistically significant (this is especially the case with large sample sizes), the inclusion of such parameters may not be of practical importance, and might even be counter-intuitive. Such complex models are often poor predictive models; in some cases, one has simply modeled "noise" in the data once the most relevant parameters were included in the model. The inclusion of too many parameters (i.e., roadway characteristics) may also lead to the introduction of correlation between different variables in the model. One way to address this problem is to assess the correlation between pairs of variables and only including one of the two, should they be highly correlated; typically one chooses the one that is easiest to obtain or makes the most engineering sense to include. However, one has to be careful when excluding variables since that could lead to omitted variable bias (discussed earlier under subsection entitled *Selection of Explanatory Variables*).

Another way to deal with overfitting is using cross-validation. When cross-validating, the data set is randomly divided into two parts, where one part is used for estimating the model parameters and the other part is used for validation. Examples of validation can be found in studies led by Simon Washington (Washington et al., 2001; Washington et al., 2005). Another approach is to use relative goodness-of-fit measures such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in selecting models; these measures penalize models with more estimated parameters than needed, and help reduce the possibility of overfitting.

Correlation among Explanatory Variables

A high degree of correlation among explanatory variables in the model (also called collinearity) makes it very difficult to determine a reliable estimate of the effects of particular variables. For example, if horizontal curvature is correlated with clear zone/roadside hazards, then it may be difficult to isolate the safety effect of horizontal curvature. There are no easy solutions to this problem. It may be tempting to remove one of the correlated variables, but this may lead to omitted variable bias, which was discussed earlier under the subsection entitled *Selection of Explanatory Variables*. Some statistical routines include tools to assess the extent of this problem. For example, one could examine the correlation matrix of the estimated parameters which will provide information about the extent of correlation between pairs of variables. A useful plotting tool is a scatterplot matrix which organizes N-choose-2 scatterplots where each individual plot shows the correlation between any 2 of N variables.

Homogenous Segments and Aggregation

The HSM advises that segments need to be divided into homogenous sections. For example, for rural two-lane roads, the HSM indicates that a segment should be created when anyone of the following variables change:

- Average daily traffic
- Lane or shoulder width
- Shoulder type
- Driveway density
- Roadside hazard rating
- Presence of an intersection
- Beginning or end of a horizontal curve
- Point of intersection of a vertical curve
- Beginning or end of a two-way left-turn lane (TWLTL)

Fitzpatrick et al. (2006) indicated that creating homogenous segments using this approach resulted in some very short segments (e.g., as short as 16 ft). Short sections can lead to a large number of sections with zero crashes which in turn leads to challenges in estimating a valid SPF (Lord et al. 2005). In addition, with short sections, there is a higher chance “that a feature of the road in one segment triggered a crash officially located on another segment” (Koorey, 2009). In fact, Hauer and Bamfo (1997) recommend that “road sections shorter than 0.1 mi should either be reassembled into longer road sections or removed from the database used for modeling”.

For these reasons, some researchers have suggested aggregating segments to create longer non-homogenous segments. A recent example is the work by Bonneson et al. (2012) that used longer non-homogenous segments, but included the proportion of segment lengths for which particular roadway characteristics are present, e.g., some of the SPFs included the proportion of segment lengths with a barrier present in the median, the proportion of segment lengths with rumble strips on the outside shoulder, and the proportion of segment lengths with rumble strips on the inside shoulder. This approach makes it more likely that all crashes attributable to severe features (or combination of features, for example short, sharp curves) fall within the appropriate analysis unit (segment) but at the cost that the safety effects of those features may be "watered down" by inclusion of adjacent features within the same analysis unit.

Presence of Outliers

In some cases, a single outlier or a few outliers can significantly influence the parameter estimates in an SPF and lead to misleading or incorrect findings. It is always recommended to perform basic quality checks of the data before attempting to model crash frequencies. This can be accomplished by plotting the data (e.g., X-Y plots, boxplots, and distribution plots) and calculating basic distributional statistics for each variable (dependent and independent). Values of predictor variables that are far outside the range of typical values for that variable could be considered leverage points in the regression analysis and should be investigated. Looking at crash rates (crashes per MVMT) for example across specific groups of segments will highlight unusual crash rates and crash counts. Extreme observations, unless they can be corrected, should be excluded from the data. During modeling, influential observations can be identified using Cook's D statistic (D stands for distance). This is a measure of the influence of a single observation on the model and is based on the comparison of the predictions with and without that observation. Rules of thumb are typically used to investigate further an observation with high influence.

Hauer (2004) proposed another approach for identifying outliers--he states that a vertical jump in the CURE plot indicates the presence of an outlier.

El-Basyouny and Sayed (2010) introduced alternative mixture models based on the multivariate Poisson lognormal (MVPLN) regression to deal with outliers. They proposed outlier resistance modeling techniques by down-weighting the outlying observations rather than excluding them unless the exclusion can be justified based on data related reasons (e.g., data collection errors).

Endogenous Explanatory Variables

Situations exist when some of the explanatory variables may depend on the dependent variable (frequency of crashes) themselves. This is known as endogeneity. Bias due to endogeneity can lead to incorrect conclusions from a model, e.g., a model may show that a treatment is associated with an increased number of crashes, when in reality the treatment may actually reduce crashes (Elvik, 2011). Obviously, this becomes a critical issue if the SPF is used to estimate the CMF associated with a particular treatment. Kim and Washington (2006) show an example as part of a study that examines the safety effectiveness of left-turn lanes. Since left-turn lanes are likely to be implemented at intersections

with large numbers of left-turn related crashes, a prediction model that includes the presence of left-turn lanes as an independent variable is likely to suffer due to endogeneity bias. Kim and Washington (2006) first estimated a model which predicted the number of angle crashes as a function of AADT, an indicator variable to represent the presence/absence of left turn lane on the major road, the number of driveways on the major road within 250 ft of the center of the intersection, and an indicator variable to represent the presence/absence of lighting on the major road. This model seemed to indicate that angle crashes would increase due to the presence of a left-turn lane. Next, to account for the possible bias due to endogeneity, the authors simultaneously estimated two models: in one model, the dependent variable was crash frequency, and in the second model, the dependent variable was a binary variable indicating the presence/absence of a left-turn lane – the two models were estimated simultaneously using limited information maximum likelihood (LIML). The models estimated using LIML indicated that the number of angle crashes would decrease with the presence of a left-turn lane, indicating that the original single equation approach did not adequately account for endogeneity bias.

Estimation of SPFs for Different Crash Types and Severities

There are many ways to estimate the predicted number of crashes by type and severity. One common approach is to use the observed proportion of the crash type or severity and apply it to the SPF that has been estimated for total crashes to estimate the predicted number of crashes for that crash type or severity. For example, if an SPF is available for total crashes for rural two-lane roads and rear-end crashes represent 20 percent of total crashes, then the SPF for rear-end crashes will simply be just the SPF for total crashes multiplied by 0.2. This approach assumes that the proportion of rear-end crashes is the same for all sites on a particular facility/roadway type regardless of the specific characteristics of the sites, e.g., this approach assumes that the proportion of rear-end crashes does not vary with AADT. Jonsson et al. (2009) found that using fixed proportions without consideration of site characteristics was a questionable practice and can lead to errors in estimation. Specifically, they found that the proportion of crashes may vary based on traffic volume and suggested that separate SPFs should be estimated for each crash type and/or severity if sufficient data are available. However, estimating separate SPFs by crash type and/or severity ignores the correlation between crash types and severities and a more complex multivariate model structure is needed to account for this correlation (e.g., see El-Basyouny and Sayed, 2009; Ma et al., 2008, Lan and Srinivasan, 2013).

An alternative approach is to start with an SPF for total crashes but explicitly model the proportion of the different crash types and severities as a function of site characteristics. This approach was proposed by Wang et al. (2011) who used a two-stage mixed multivariate model combining both accident frequency and severity. Wang et al (2011) showed that the two-stage mixed model approach was able to make use of detailed individual level accident data that the traditional approaches such as fixed proportion values for crash type/severity, separate SPFs by crash type and/or severity, or multivariate SPFs (to account for the correlation) are not able to do. Further discussion of model forms for estimating the proportion of crashes by severity can be found in Savolainen et al., (2011).

Goodness of Fit

For linear regression models, the R^2 statistic, the proportion of the total variation in the dependent variable explained by the model, is the goodness-of- fit (GOF) measure typically reported. For GLMs, a direct analogous R^2 statistic is not available. However, there are other fit statistics that have been suggested to assess how well a GLM fits the data. An overview and references are provided here.

The Scaled Deviance and the Pearson chi-square are two traditional GOF statistics calculated in GLMs (Wood, 2002). In addition, many pseudo R^2 statistics have been used by statisticians and other safety researchers. Examples include the pseudo R^2 based on the log-likelihood ratio, weighted (variance stabilizing) residuals, Freeman-Tukey transformation residuals, and the overdispersion parameter of the SPF. Further discussion of pseudo R^2 used in the highway safety field can be found in Fridstrom et al., (1995), Wood (2002), and Miaou (1996).

Another possible GOF was proposed by Liu and Cela (2008). Their approach involves the comparison of the empirical distribution of the observed counts to the negative binomial distribution with the mean estimated from the data. The probabilities from the two distributions are plotted. The extent of the overlap between the predicted and observed probabilities provides insight into the GOF of the model. The goal is to have nearly complete overlap between the predicted and the observed probabilities.

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are examples of measures that provide an assessment of the relative quality of specific models, for a given dataset. AIC and BIC penalize models based on the number of parameters (coefficient estimates). In other words, they deal with the trade-off between the GOF of the model and the complexity of the model. Therefore, AIC and BIC are statistics typically used as model selection criteria rather than for GOF assessment. Further discussion of AIC and BIC can be found in Burnham and Anderson (2004).

CURE plots, discussed earlier, also provide a means of assessing the GOF of a model. Unlike most of the other GOF statistics that look at overall model fit, the CURE plot is primarily aimed at assessing the adequacy of the functional form of a specific independent variable (conditional on other variables being in the model).

5. Steps Involved in Developing SPFs

This section presents a logical order of the steps users can follow for any specific situation for which they desire developing an SPF. These steps pertain to the intended use of the SPF and the facility type for which it is to be applied; the data collection and preparation; and the statistical modeling. Details for each step are provided next.

Step 1 – Determine use of SPF.

Based on the discussion in Section 3 of this document, SPFs may be used for the following purposes:

1. Network screening
2. Project-level analysis using the HSM prediction methodology. Two scenarios are possible here:
    - Scenario 1 – use only the data for the base conditions to estimate the SPF (if sufficient data are available with these conditions)
    - scenario 2 – use data for a broader set of conditions to estimate the SPF
3. Derive CMFs directly from the SPF
4. Before-after evaluation using the EB method

Step 2 – Identify facility type.

The user needs to select the facility type to which the SPF will be applied. The list of facility types (separately for roadway segments, intersections, and ramps) provided by *Safety Analyst* is a good starting point. For project level analysis using the HSM prediction methodology, Part C of the HSM provides a list of facility types. Both lists are shown in Appendix A of the ***SPF Decision Guide*** (Srinivasan et al., 2013).

Step 3 – Compile necessary data.

As discussed earlier, SPFs that are used for network screening (purpose 1), for each facility type, the number of crashes for each unit (intersection, segment, or ramp), along with the traffic volume (AADT) associated with that unit are required. For roadway segments and ramps, the segment length will be required as well. For intersections of any type, it is recommended that AADT for both major and minor roads be available. Since the SPFs in the HSM for the base conditions are also just a function of traffic volume, the explanatory variables in the SPFs for purpose 1 and scenario 1 of purpose 2 are the same. Here is an example.

Suppose the goal is to use the HSM prediction methodology to conduct project level analysis for rural divided multilane roadway segments, and the analyst wants to develop jurisdiction-specific SPFs instead of calibrating the default SPFs in Chapter 11 of the HSM. Based on page 11-17 of the HSM, following are the base conditions for the SPF for divided roadway segments on rural multilane roads

- Lane width = 12 feet
- Right shoulder width = 8 feet

- Median width = 30 feet
- Lighting = None
- Automated Speed Enforcement = None

If sufficient data are available corresponding to the base conditions (listed above) to estimate an SPF (i.e., scenario 1 of purpose 2), then the analyst can assemble the data for these sites, and the SPF can be estimated based on segment length and AADT as the explanatory variables . Table 1 of the ***SPF Decision Guide*** (Srinivasan et al., 2013) provides some guidance on the minimum sample size that is needed for estimating SPFs.

However, if sufficient data are not available to estimate an SPF based on sites with the base condition (i.e., scenario 2 of purpose 2), then data for a broader set of conditions than the base condition need to be assembled, so that a sufficient sample of sites are available for estimating the SPF. In this scenario, the SPF will not only include segment length and AADT, but the other explanatory variable such as lane width, shoulder width, median width, presence/absence of lighting, and the presence/absence of automated enforcement.

The data needs for purpose 3 are similar to those of scenario 2 of purpose 2. Suppose the goal is to estimate a CMF for shoulder width using an SPF, there is a need to assemble data from enough sites with a range of shoulder width. It is also necessary to compile data for other explanatory variables that are potentially related to safety (such as AADT, median width, lane width, curvature, and grade).

In purpose 4, the goal is to use the SPF as part of an EB before-after evaluation. Here the SPF needs to be estimated for a reference group that is similar to the treatment group, but without the treatment. For example, if the intent of the before-after evaluation is to determine the CMF for changing the shoulder width from 4 to 8 feet, then it is necessary to find a reference group of sites (similar to the treatment group) where the shoulder width remained at 4 feet during the entire study period (before and after).

For all the situations discussed above, if SPFs are needed for more than one crash type or severity, then crash counts by type and severity need to be obtained for each unit (i.e., segment or intersection).

Step 4 – Prepare and cleanup database.

At this stage, one or more databases (e.g., individual crash, traffic volume, and roadway characteristics databases) have been assembled. Basic quality checks and outlier checks are in order at this point. This is accomplished using the basic tools discussed in Section 4 (e.g., plotting tools, evaluation of basic descriptive statistics, checking for outliers and data entry errors).

Step 5 – Develop the SPF.

As discussed in Section 4, there are a variety of statistical issues that need to be considered in developing an SPF. This step might consist of a series of trials (i.e., this is typically an iterative process) that require both statistical modeling skills and engineering judgment. Using the available software, estimate the regression coefficients of the model for each desired crash type. Calculate model

diagnostics such as goodness-of-fit statistics and examining residual and CURE plots as discussed in Section 4; check that the regression coefficients make engineering sense (e.g., their sign is in the anticipated direction); and perhaps perform sensitivity analyses for a number of factors.

This step is easier to accomplish for purpose 1 and scenario 1 of purpose 2 since these SPFs typically only include segment length and AADT for roadway segments and ramps, and major and minor road AADT for intersections. The appropriate functional form of the SPF needs to be determined as well in this step. If the analyst decides to use *Safety Analyst* (for network screening), or *IHSDM* (for project level analysis), then the SPF needs to follow the functional form that these tools can accommodate. In *Safety Analyst*, for example, the SPFs for roadway segments and ramps are power functions of the form: $Y = L \times e^{a} \times (AADT)^{b}$, and the SPFs for intersections are of the form:

$Y = e^{a} \times (AADT_{major})^{b} \times (AADT_{\min or})^{c}$

Since the crash prediction model in the IHSDM implements the HSM prediction methodology from Part C, if the analyst decides to use IHSDM , the SPFs need to follow the functional form of the SPFs in Part C of the HSM.

On the other hand, if the analyst is planning to use their own customized tool for network screening or project level analysis (instead of *Safety Analyst* or IHSDM), then more complicated functional forms may be possible (e.g., Hauer, 2004; Kononov et al., 2011).

This step is more difficult to accomplish for scenario 2 of purpose 2, and purposes 3 and 4, since other explanatory variables in addition to segment length and AADT need to be considered for inclusion in the SPF. Readers are referred to Section 4 for further discussion of the statistical issues associated with this step.

Another decision that needs to be made is how the SPFs will be estimated for multiple crash types and severities. As discussed in Section 4, there are different options:

1. Multiply the total crash SPF with the proportion of a crash type and/or severity level to obtain the SPF for that particular crash type and/or severity
2. Estimate separate SPFs by crash type and/or severity level. An extension of this approach will be the use of multivariate models to account for the correlation between the crash types and severities
3. Model the proportion of the different crash types and severities as a function of site characteristics and then multiply the total crash SPF with these proportions

Step 6 – Develop the SPF for the base condition.

For scenario 2 of purpose 2, the SPF for given base conditions is obtained by substituting the value of the desired base conditions in the SPF.

Step 7 – Develop CMFs for specific treatments.

For purpose 3, follow the procedure described in Section 3 under *Evaluation the Effect of Engineering Treatments—Estimating CMFs directly from SPFs*. If the SPFs include interaction terms, the CMF may be a crash modification function instead of a crash modification factor (e.g., see equation 4.15).

Step 8 – Document the SPFs.

After the SPFs are estimated, it is important to document them so that they can be used by other analysts and researchers in the future. Here are some of the details that need to be included as part of the documentation:

- Crash type(s)/severity(s) for which the SPF was estimated
- Total number of crashes (by type and severity) used in the estimation
- Purpose of the SPF (e.g., network screening, project level analysis, CMF development, etc.)
- State(s)/county(s)/city(s) that were used
- Facility type (e.g., rural 2 lane, 3 leg stop-controlled intersection, freeway to freeway exit ramp)
- Number of years used in the estimation of SPF
- Number of units (segments, intersections, ramps)
- Minimum, maximum, and average length of segments
- Minimum, maximum, and average AADT
- Minimum, maximum, and average values for key explanatory variables
- Coefficient estimates of the SPF
- Standard errors of the coefficient estimates[2]
- Goodness of fit statistics
- Discussion of potential biases or pitfalls

[2] There has been some debate about the usefulness of this parameter (Hauer, 2013). Nevertheless, most statisticians are in favor of reporting this parameter.

6. Recent Advances in SPF Development and Estimation

This section provides a brief discussion of some recent advances in SPF development and estimation. It also directs the reader to other useful documents.

Variance of Crash Estimates Obtained from SPFs

Most of the discussion in previous research has focused on using SPFs to estimate the expected value of crashes under certain conditions. Wood (2005) illustrates how prediction intervals for the number of crashes at a new site can be calculated based on the coefficients of the independent variables and the covariance matrix of these coefficients. It is important to note that the procedure discussed in Wood (2005) is for GLMs with a log link. Lord (2008) used the method in Wood (2005) to develop a methodology for estimating the variance and 95-percent confidence intervals for the estimate of the product between baseline SPFs and CMFs (the method discussed in Part C of the HSM for estimating the expected number of crashes at a site).

Temporal and Spatial Correlation

Temporal correlation can lead to incorrect estimates of the standard errors of the coefficients. Temporal correlation may arise when multiple observations are used for the same roadway unit. This is often the case when multiple years/months of data are used in the modeling because "many of the unobserved effects associated with a specific roadway entity will remain the same over time" [Lord and Mannering (2010, p. 292)]. This type of data is also sometimes called panel data. A common approach to dealing with temporal correlation is to aggregate the data so that each roadway unit has one observation, e.g., if 3 years of data are available for SPF estimation, and crash counts and site characteristics (e.g., AADT) are available for each of these 3 years, then for each roadway unit, the total crash counts over the 3 years is computed and used as the dependent variable along with the average value of the site characteristics over the 3 year period as the explanatory variables (the number of years may then be used as an offset so that the SPF provides the prediction for the number of crashes per year). Kweon and Lim (2012) found that SPFs based on aggregated data could underestimate the overdispersion parameter of the SPFs.

Other methods for addressing temporal correlation include generalized estimating equations (GEE) (e.g., see Lord and Persaud, 2000), random effects models (e.g., see Shankar et al., 1998), and negative multinomial models (e.g., see Hauer, 2004). Ulfarrson and Shankar (2003), in their study of median crossover crashes, found that the negative multinomial model outperformed the random effects model in terms of fit, but that the negative multinomial model was more difficult to estimate because of convergence problems.

Similar to temporal correlation, spatial correlation may occur because "roadway entities that are in close proximity may share unobserved effects" (Lord and Mannering, 2010). Wang and Abdel-Aty (2006) used

GEE to account for spatial correlation in their examination of rear-end crashes at signalized intersections. Other examples include Aguero-Valverde and Jovanis (2006) and Guo et al. (2010).

Other Model Forms

Some recent studies have used model forms other than a negative binomial form. These include zero-inflated models, Poisson-lognormal models, and Conway-Maxwell-Poisson models. Zero-inflated (can be zero-inflated Poisson or zero-inflated negative binomial) are specifically used to handle data characterized by a significant number of zero crash sites. Zero-inflated models are based on the theory that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. Zero-inflated models have been used in a few studies. However, they have also been criticized because the "zero or safe state has a long-term mean equal to zero, this model cannot properly reflect the crash-data generating process" (Lord and Mannering, 2010). Further discussion of zero-inflated models can be found in Shankar et al., (2003) and Lord et al., (2007).

Some studies have used Poisson-lognormal models as an alternative to negative binomial models (e.g., Lan and Srinivasan, 2013; Aguero-Valverde and Jovanis, 2008). Poisson-lognormal models are similar to the negative binomial models except that the ε_i term in equation 4.3 is log-normally- instead of gamma-distributed. Lord and Mannering (2010) indicate that Poisson-lognormal may be more flexible, but also more difficult to estimate.

Unlike negative binomial models, Conway-Maxwell-Poisson model can handle both overdispersion and underdispersion. Recent research has found that for data that was overdispersed, the negative binomial and Conway-Maxwell-Poisson models were comparable (Lord and Mannering, 2010). However, the Conway-Maxwell-Poisson model is more difficult to estimate compared to the negative binomial model.

Generalized Additive Models

Generalized additive models (GAM) introduce smoothing functions for each explanatory variable in the model and hence provide a more flexible functional form. GAMs can include both parametric and non-parametric forms. However, GAMs do not have coefficients associated with the smoothing functions and hence are much more difficult to use and interpret as an SPF. For examples of the use of GAMs in highway safety, readers are referred to Xie and Zhang (2008).

Random-Parameters Models

Random-parameters models allow the estimated parameters (coefficients) to vary across the individual observations, but usually based on a pre-specified distribution. The goal of these models is to account for the unobserved heterogeneity among observations. Since these models allow for this flexibility, they provide a better statistical fit compared to models where the coefficients are assumed to be fixed. However, random-parameter negative binomial models are more difficult to estimate and they may not necessarily be more useful compared to fixed parameter models (Lord and Mannering, 2010).

Bayesian Estimation Methods

Bayesian models integrate Bayes' theorem with classical statistical models (Washington et al., 2011). Bayesian models allow the use of prior information about the parameters (i.e., regression coefficients) in addition to the data to obtain the "posterior estimate" of the parameter values. Bayesian models have become more common because of the accessibility of Markov Chain Monte Carlo (MCMC) methods. MCMC methods more easily allow the estimation of complex functional forms that are often difficult to estimate using traditional maximum likelihood methods. Bayesian methods are also more effective in modeling spatial correlation. Examples of the use of Bayesian estimation methods can be found in Lan and Srinivasan (2013), Guo et al., (2010), Ma et al., (2008), and El-Basyouny, K. and Sayed, T. (2009).

7. Software Tools for Estimating SPFs

A number of statistical software tools are available for estimating SPFs. Examples include SAS, SPSS, STATA, R, and LIMDEP. There are advantages and disadvantages to each software tool. However, the tools are constantly evolving as the developers strive to provide more features and improve the capability of the tools.

For estimating negative binomial models using maximum likelihood methods, most of the popular modules within these tools (e.g., PROC GENMOD and PROC GLIMMIX within SAS) are geared towards estimating a GLM (for sample SAS code to estimate SPFs using PROC GLIMMIX, calculate goodness of fit measures, and develop CURE plots, readers are referred to Appendix F of Srinivasan and Carter, 2011). The GLM allows the estimation of the negative binomial model without the need for starting estimates for the parameters. However, if more flexible functional forms (which may not be a GLM) need to be used (e.g., as in equation 4.16), then the analyst would need to code the log-likelihood function and provide starting estimates for the parameters before the estimation process can proceed. As an example, PROC NLMIXED in SAS provides this option (sample SAS code for estimating negative binomial models using PROC NLMIXED can be found in Liu and Cela, 2008). Another example is the Solver tool within Microsoft Excel. Hauer (2013) advocates the use of Microsoft Excel because of the convenience that it provides for the visualization and graphing of the data.

8. References

AASHTO (2010), *Highway Safety Manual*, AASHTO, Washington, D.C.

Aguero-Valverde, J., and Jovanis, P. (2006), Spatial Analysis of Fatal and Injury Crashes in Pennsylvania, *Accident Analysis and Prevention*, Vol. 38(3), pp. 618-625.

Aguero-Valverde, J., and Jovanis, P. (2008), Analysis of Road Crash Frequency with Spatial Models, *Transportation Research Record* 2061, pp. 55-63.

Banihashemi, M. (2012), Highway Safety Manual Calibration Dataset Sensitivity Analysis, *Presented at the 91st Annual Meeting of the Transportation Research Board*, Washington, D.C.

Bauer, K. and Harwood, D. (2012), *Safety Effects of Horizontal Curve and Grade Combinations*, Submitted to Federal Highway Administration.

Bonneson, J.A., Geedipally, S., Pratt, M.P., and Lord, D. (2012), *Safety Prediction Methodology and Analysis Tool for Freeways and Interchanges*, Final Report for NCHRP Project 17-45, Washington, D.C.

Burnham, K.P. and Anderson, D.R. (2004), Multimodel Inference: Understanding AIC and BIC in Model Selection, *Sociological Methods and Research*, Vol. 33, pp. 261-304.

Cameron, C.A. and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, UK.

Cafiso, S., Di Silvestro, G., Persaud, B., and Begum, M. (2010), Revisiting Variability of Dispersion Parameter of Safety Performance for Two-Lane Rural Roads, *Transportation Research Record* 2148, pp. 38-46.

Carter, D., R. Srinivasan, F. Gross, and F. Council (2012), *Recommended Protocols for Developing Crash Modification Factors*, Prepared as part of NCHRP Project 20-07 (Task 314), Washington, D.C. Available at http://www.cmfclearinghouse.org/resources_develop.cfm. Accessed January 2013.

CDOT (2009), *Safety Performance Functions for Intersections*, Report CDOT-2009-10, Developed for Colorado Department of Transportation, Developed by Persaud and Lyon, Inc. and Felsburg Holt & Ullevig, December 2009.

Dixon, K., C. Monsere, F. Xie, and K. Gladhill (2012), *Calibrating the Future Highway Safety Manual Predictive Methods for Oregon State Highways*, Final Report SPR 684, OTREC-RR-12-02, Oregon State University and Portland State University, Oregon.

El-Basyouny, K. and Sayed, T. (2009), Collision Prediction Models using Multivariate Poisson-Lognormal Regression, *Accident Analysis and Prevention*, Vol. 41(4), pp. 820-828.

El-Basyouny, K. and Sayed, T. (2010), A Method to Account for Outliers in the Development of Safety Performance Functions, *Accident Analysis and Prevention*, Vol. 42, pp. 1266-1272.

Elvik, R. (2011), Assessing Causality in Multivariate Accident Models, *Accident Analysis and Prevention*, Vol. 43, pp. 253-264.

Fitzpatrick, K., Schneider IV, W.H., and Carvell, J. (2006), Using the Rural Two-Lane Highway Draft Prototype Chapter, *Transportation Research Record* 1950, pp. 44-54.

Fridstrom, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., and Thomsen, L. K. (1995), Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts, *Accident Analysis and Prevention*, Vol. 27 (1), pp. 1-20.

Gross, F., B. Persaud, and C. Lyon (2010), *A Guide for Developing Quality Crash Modification Factors*, Report FHWA-SA-10-032, Federal Highway Administration, Washington, D.C. Available at http://www.cmfclearinghouse.org/resources_develop.cfm. Accessed January 2013.

Guo, F., Wang, X., and Abdel-Aty, M. (2010), Modeling Signalized Intersection Safety with Corridor-level Spatial Correlations, *Accident Analysis and Prevention*, Vol. 42(1), pp. 84-92.

Harwood, D., Council, F., Hauer, E., Hughes, W. and Vogt, A. (2000), *Prediction of the Safety Performance of Rural Two-Lane Highways*, Report FHWA-RD-99-207, Federal Highway Administration.

Harwood, D.W., D. J. Torbic, K. R. Richard, and M. M. Meyer (2010), *Safety Analyst: Software Tools for Safety Management of Specific Highway Sites*, Federal Highway Administration, Report Number FHWA-HRT-10-063, Available at http://www.safetyanalyst.org/docs.htm. Accessed January 2013.

Hauer, E. (1997), *Observational Before After Studies in Road Safety*, Elsevier Science, New York.

Hauer, E. and Bamfo, J. (1997), Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables, *Proceedings ICTCT (International Cooperation on Theories and Concepts in Traffic Safety)*, Lund, Sweden.

Hauer, E. (2001), Overdispersion in modeling accidents on road sections and in empirical Bayes estimation, *Accident Analysis and Prevention*, Vol. 33(6), pp. 799-808.

Hauer, E., Harwood, D., Council, F., and Griffith, M. (2002), Estimating Safety by the Empirical Bayes Method: A Tutorial, *Transportation Research Reco*rd 1784, pp. 126-131.

Hauer, E. (2004), Statistical Road Safety Modeling, *Transportation Research Record* 1897, pp. 81-87.

Hauer, E. (2010), Cause, Effect, and Regression in Road Safety: A Case Study, *Accident Analysis and Prevention*, Vol. 42, pp. 1128-1135.

Hauer, E. (2013), *Safety Performance Functions: A Workshop*, Baton Rouge, Louisiana, July 16-18, 2013.

Jonsson, T., Lyon, C., Ivan, J., Washington, S., Van Schalkwyk, I., and Lord, D. (2009), Investigating Differences in Safety Performance Functions Estimated for Total Crash Count and for Crash County by Collision Type, *Transportation Research Record* 2102, pp. 115-123.

Kim, D. and Washington, S. (2006), The Significance of Endogeneity Problems in Crash Models: An Examination of Left-Turn Lanes in Intersection Crash Models, *Accident Analysis and Prevention*, Vol. 38(6), pp. 1094-1100.

Kononov, J., C. Lyon, B.K. Allery (2011), Relationship of Flow, Speed, and Density of Urban Freeways to Functional Form of a Safety Performance Function, *Transportation Research Record* 2236, pp. 11-19.

Koorey, G. (2009), Road Data Aggregation and Section Considerations for Crash Analysis, *Transportation Research Record* 2103, pp. 61-68.

Kweon, Y.-J. and Lim, I.-K. (2012), Appropriate Regression Model Types for Intersections in Safety Analyst, *Journal of Transportation Engineering (ASCE)*, Vol. 138(10), pp. 1250-1258.

Lan, B. and Srinivasan, R. (2013), Safety Evaluation of Discontinuing Late Night Flash Operations at Signalized Intersections, *Presented at the 2013 Annual Meeting of the Transportation Research Board*, Washington, D.C.

Lord, D., Washington, S.P., and Ivan, J. (2005), Poisson, Poisson-Gamma, and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory, *Accident Analysis and Prevention*, Vol. 37(1), pp. 35.46.

Lord, D., Washington, S.P., and Ivan, J.N. (2007), Further Notes on the Application of Zero Inflated Models in Highway Safety, *Accident Analysis and Prevention*, Vol. 39(1), pp. 53-57.

Lord, D. (2008), Methodology for Estimating the Variance and Confidence Intervals for the Estimate of the Product of Baseline Models and AMFs, *Accident Analysis and Prevention*, Vol. 40, pp. 1013-1017.

Lord, D. and Mannering, F. (2010), The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives, *Transportation Research Part A*, Vol. 44, pp. 291-305.

Liu, W. and Cela, J. (2008), Count Data Models in SAS, *SAS Global Forum 2008: Statistics and Data Analysis* (Paper 371-2008).

Lu, J., A. Gan, K. Haleem, P. Alluri, and K. Liu (2012), Comparing Locally-Calibrated and Safety-Analyst Default Safety Performance Functions for Florida's Urban Freeways, *Presented at the 91st Annual Meeting of the Transportation Research Board*, Washington, D.C.

Ma, J., Kockelman, K.M., and Damien, P. (2008), A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, using Bayesian Methods, *Accident Analysis and Prevention*, Vol. 40(3), pp. 964-975.

Marinelli, F., F. la Torre, and P. Vadi (2009), Calibration of the Highway Safety Manual's Accident Prediction Model for Italian Secondary Road Network, *Transportation Research Record* 2103, pp. 1-9.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Second Edition, Chapman and Hall/CRC.

Miaou, S.P. (1996), *Measuring the Goodness-of-fit of Accident Prediction Models*. Federal Highway Administration, FHWA-RD-96-040, Oak Ridge, TN: Oak Ridge National Laboratory.

Miaou, S.P. and Lord, D. (2003), Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record* 1840, pp. 31–40.

Mitra, S. and Washington, S. (2007), On the Nature of the Over-Dispersion Parameter in Motor Vehicle Crash Prediction Models, *Accident Analysis and Prevention*, Vol. 39(3), pp. 459-468.

Persaud, B., D. Lord, and J. Palmisano (2002), Calibration and Transferability of Accident Prediction Models for Urban Intersections, *Transportation Research Record* 1784, pp. 57-64.

Persaud, B., Y. Chen, J. Sabbaghi, and C. Lyon (2012), Adoption of the Highway Safety Manual Methodologies for Safety Performance Assessment of Ontario Highway Segments, *Proceedings of the 22nd Canadian Multidisciplinary Road Safety Conference*, Alberta, June 2012.

Qin, X., Ivan, J. N., Ravishanker, N., Liu, J., and Tepas, D. (2006), Bayesian estimation of hourly exposure functions by crash type and time of day, *Accident Analysis and Prevention*, Vol. 38 (6), pp. 1071-1080.

Sacci, E., B. Persaud, and M. Bassani (2012), Assessing International Transferability of the Highway Safety Manual Crash Prediction Algorithm and its Components, *Presented at the 91st Annual Meeting of the Transportation Research Board*, Washington, D.C.

Savolainen, P.T., Mannering, F.L., Lord, D., and Quddus, M.A. (2011), The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives, *Accident Analysis and Prevention*, Vol. 43, pp. 1666-1676.

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., and Nebergall, M.B. (2003), Modeling Crashes Involving Pedestrians and Motorized Traffic, *Safety Science*, Vol. 41(7), pp. 627-640.

Srinivasan, R., Council, F., and Harkey, D. (2008), *Calibration Factors for HSM Part C Predictive Models*, Submitted to FHWA and HSM Task Force, October 2008.

Srinivasan, R., Lyon, C., Persaud, B., Martell, C., and Baek, J. (2011), *Methods for Identifying High Collision Concentration Locations (HCCL) for Potential Safety Improvements – Phase II: Evaluation of Alternative Methods for Identifying HCCL*, Submitted to California Department of Transportation (CFS Number 2078A DRI), January 2011.

Srinivasan, R. and D. Carter (2011), *Development of Safety Performance Functions for North Carolina*, Report FHWA/NC/2010-09, Submitted to NCDOT, December 2011.

Srinivasan, R., Carter, D., and Bauer, K. (2013), *How to Choose Between Calibrating SPFs from the HSM and Developing Jurisdiction-Specific SPFs*, Prepared for Federal Highway Administration.

Tegge, R.A., J. Jang-Hyeon, Y. Ouyang (2010), *Development and Application of Safety Performance Functions for Illinois*, Research Report ICT-10-066, Illinois Center for Transportation, Civil Engineering Studies, March 2010.

Vogt, A. and Bared, J.G. (1998), *Accident Models for Two-lane Rural Roads: Segments and Intersections*, Federal Highway Administration, Report # FHWA-RD-98-133.

Wang, C., Quddus, M.A., and Ison, S.G. (2011), Predicting Accident Frequency at their Severity Levels and its Application in site ranking using a Two-Stage Mixed Multivariate Model*, Accident Analysis and Prevention*, Vol. 43, pp. 1979-1990.

Washington, S., J. Leonard, D.G. Manning, C. Roberts, B. Williams, A.R. Bacchus, A. Devanhalli, J. Ogle, and D. Melcher (2001), *Scientific Approaches to Transportation Research Volumes 1 and 2*. NCHRP Online Report 20-45, Transportation Research Board, National Cooperative Highway Research Program, Washington, DC. Available online at: http://onlinepubs.trb.org/Onlinepubs/nchrp/cd-22/start.htm

Washington S., B. Persaud, C. Lyon, and J. Oh. (2005), *Validation of Accident Models for Intersections*. Federal Highway Administration, FHWA-RD-03-037, Washington, DC.

Washington, S.P., Karlaftis, M.G., and Mannering, F.L. (2011), *Statistical and Econometric Methods for Transportation Data Analysis*, Second Edition, Chapman and Hall/CRC, Boca Raton, FL.

Wood, G.R. (2002), Generalized linear accident models and goodness of fit testing, *Accident Analysis and Prevention*, Vol. 34, pp. 417-427.

Wood, G.R. (2005), Confidence and Prediction Intervals for Generalized Linear Accident Models, *Accident Analysis and Prevention*, Vol. 37, pp. 267-273.

Xie, Y. and Zhang, Y. (2008), Crash Frequency Analysis with Generalized Additive Models, *Transportation Research Record* 2061, pp. 39-45.