2018

SEGMENTATION STRATEGIES FOR ROAD SAFETY ANALYSIS

Eric R. Green
University of Kentucky, eric.green@uky.edu
Author ORCID Identifier:
https://orcid.org/0000-0002-6587-2606
Digital Object Identifier: https://doi.org/10.13023/ETD.2018.056

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Eric R. Green, Student

Dr. Reginald R. Souleyrette, Major Professor

Dr. Yi-Tin Wang, Director of Graduate Studies

SEGMENTATION STRATEGIES FOR ROAD SAFETY ANALYSIS

DISSERTAION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Engineering at the University of Kentucky

By

Eric Green

Lexington, Kentucky

Co Directors: Dr. Reginald R. Souleyrette Commonwealth Chair, Professor of Civil Engineering
And Dr. Nikiforos Stamatiadis Professor of Civil Engineering
Lexington, Kentucky

ABSTRACT OF DISSERTATION

SEGMENTATION STRATEGIES FOR ROAD SAFETY ANALYSIS

This dissertation addresses the relationship between roadway segment length and roadway attributes and their relationship to the efficacy of Safety Performance Function (SPF) models. This research focuses on three aspects of segmentation: segment length, roadway attributes, and combinations of the two. First, it is shown that choice of average roadway segment length can result in markedly different priority lists. This leads to an investigation of the effect of segment length on the development of SPFs and identifies average lengths that produce the best-fitting SPF. Secondly, roadway attributes are filtered to test the effect that homogeneity has on SPF development. Lastly, a combination of segment length and attributes are examined in the same context.

In the process of conducting this research a tool was developed that provides objective goodness-of-fit measures as well as visual depictions of the model. This information can be used to avoid things like omitted variable bias by allowing the user to include other variables or filter the database. This dissertation also discusses and offers examples of ways to improve the models by employing alternate model forms.

This research revealed that SPF development is sensitive to a variety of factors related to segment length and attributes. It is clear that strict base condition filters based on the most predominant roadway attributes provide the best models. The preferred functional form was shown to be dependent on the segmentation approach

(fixed versus variable length). Overall, an important step in SPF development process is evaluation and comparison to determine the ideal length and attributes for the network being analyzed (about 2 miles or 3.2 km for Kentucky parkways). As such, a framework is provided to help safety professionals employ the findings from this research.

KEYWORDS: Road Safety, Segmentation, Safety Performance Functions, Highway Safety Manual, Network Screening

Eric R. Green

March 7, 2018

Date

SEGMENTATION STRATEGIES FOR ROAD SAFETY ANALYSIS

By

Eric Randolph Green

Dr. Reginald R. Souleyrette
Co-Director of Dissertation

Dr. Nikiforos Stamatiadis
Co-Director of Dissertation

Dr. Yi-Tin Wang
Director of Graduate Studies

March 7, 2018

To my family and friends,

My children,

My colleagues,

My Advisors,

And especially my wife for the strength,

I dedicate this work.

Acknowledgements

I would like to thank my committee: Dr. Souleyrette, Dr. Stamatiadis, Dr. Chen, and Dr. Wendroth. Their guidance made this research possible. As co-chairs, Dr. Souleyrette and Dr. Stamatiadis provided a perfect combination of mentoring styles, each providing a unique perspective.

I would also like to thank my supervisor, Jerry Pigman. His guidance started while I was a young student in 1998 and continued through my career at the Kentucky Transportation Center. While pursuing this advanced degree, Mr. Pigman provided me with the flexibility to continue my education while I was still growing as a professional engineer. My colleagues, Mr. Fields and Dr. Blackden also played large roll in this research including countless brainstorming sessions in front of a dry erase board.

I would like to thank my safety counterparts at the Kentucky Transportation Cabinet. Mr. Lovell, Mr. Vaughn, Mr. Stanley, and Mr. Durman. Their work for the Highway Safety Improvement Program helped to inspire this research. Their passion for safety helped us innovate Kentucky's safety prioritization methodologies.

Finally, I would like to thank my family. My children for understanding nights and weekends in front of my computer. My mother and father for encouraging me to pursue my career in civil engineer. And, of course, my wife for taking on more responsibilities to cover my absence, encouraging me to never stop, and helping me see my full potential.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Problem and Background

Data-driven approaches to highway safety have been widely used to identify high-risk road segments and intersections through the Highway Safety Improvement Program (HSIP) in order to improve highway safety. Wu et al. (2012) found that national traffic fatalities declined approximately 7.5 percent following the introduction of the HSIP. Interventions based on data driven-prioritization methods are responsible for much of this reduction. Still, according to the National Highway Traffic Safety Administration, motor vehicle crashes resulted in 37,461 deaths in the United States in 2016 (a rate of 1.18 fatalities per 100 million vehicle miles travelled) (NHTSA, 2017).

The *Highway Safety Manual* (HSM) outlines methodologically sophisticated techniques to predict the number of crashes for specific facility types. Transportation agencies can implement these to predict the potential number of crashes and use their findings to develop cost-benefit estimates in order to better allocate funding and maximize the benefits of safety improvements. Techniques that had been applied before the introduction of these methods generally relied on crash frequencies or crash rates. Despite their widespread use, the randomness of crash data could often result in inappropriate selections for safety improvements (AASHTO, 2010, Srinivasan et al., 2011).

Methods described in the HSM, particularly the Empirical Bayes (EB) method, have proven extremely effective. States which have prioritized hazardous sites through the use of detailed roadway inventory data and the EB method have experienced the most significant crash reductions (Wu et al., 2012). Elvik (2008) demonstrated that an EB technique performs better at identifying hazardous locations as compared to four other methods; including counts, crash rates, and critical counts.

The HSM describes a network screening approach for prioritizing roadway segments for safety analysis (AASHTO, 2010). Network screening is a technique that

analyzes homogeneous roadway segments (i.e., segments with similar geometric and traffic characteristics). Crashes are assigned to each segment and a Safety Performance Function (SPF), Crash Modification Factors (CMFs), and calibration identify the number of crashes expected for the section. An SPF is a negative binomial regression model that is used to predict crash frequency typically using traffic volume and segment length as predictors. EB adjusts the expected number of crashes based on historical data for a better estimate. Research has shown that segment length can affect the outcome of safety prioritization using methodologies predating the HSM (Cook et al., 2011, Green et al. 2017). Research based on the HSM methodologies has demonstrated this effect using SPFs (Srinivasan et al., 2011) but there is currently no guidance as to what segment length to use for network screening to identify and prioritize hazardous locations. The research presented here investigates the effect of segment length on safety analysis in the context of network screening. Transportation professionals can benefit from this research with guidance as to what segment length is most appropriate and beneficial for particular safety analyses.

The development of SPFs requires a data set of roadway segments or intersections that are homogeneous; that is, with similar roadway characteristics. A common way to create a dataset of homogeneous roadway segments is to begin with roadway inventory data. The HSM offers guidance as to what roadway characteristics could be used for creating homogeneous segments (AASHTO, 2010). In the U.S., state transportation agencies benefit from a uniform set of roadway elements developed by the Federal Highway Administrations (FHWA) known as the Model Inventory of Roadway Elements (MIRE) (FHWA, 2010). Many of these inventories were created at different times, by different groups within an agency, and, most importantly, using a variety of segmentation techniques. In the context of roadway segments, segmentation is usually defined by beginning and ending milepoint. This facilitates the use of a linear reference system – encouraging the use of a Geographic Information System (GIS). The decision of where to start and stop a given segment depends on the presence of inventory attributes. For instance, traffic volumes will change at major intersections,

whereas, the width of a right shoulder might change due to terrain or the availability of right of way. Segments may also be defined at the beginning and end locations of vertical or horizontal curves. Consider the following roadway segment in Figure 1 that depicts the changes of seven attributes and the resulting segmentation that would be required for homogeneity (at the bottom).

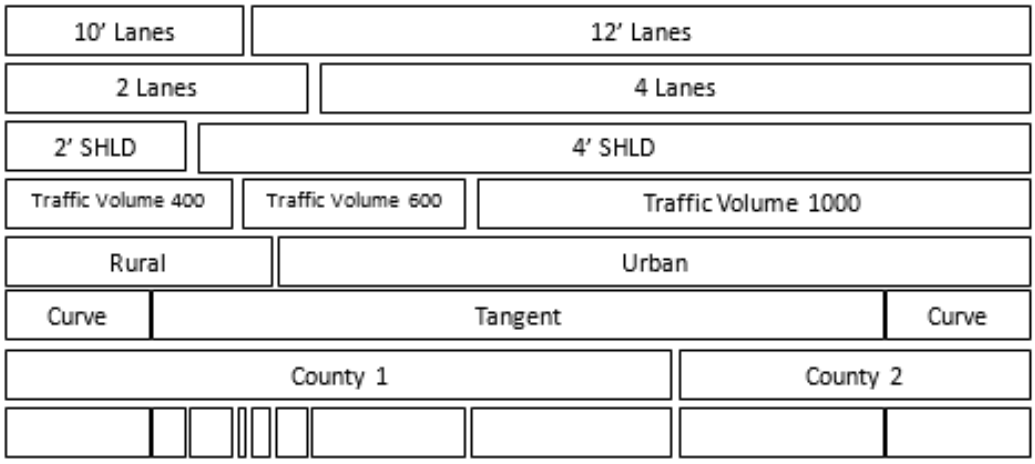


Figure 1. Typical Segmentation Resulting from Varying Roadway Attributes

The combination of these seven attributes results in 10 homogeneous segments, some shorter in length compared to others. This network segmentation method results in the creation of segments of varying lengths and, in some cases, based on arbitrary break points (such as county boundaries[1]). This type of segmentation is based solely on the roadway attributes.

While it is necessary to use the attributes to create a roadway network for safety analysis, it is important to consider the length of the segment. In Kentucky, a network was segmented using a fixed length, a variable length, and a modified variable length to produce three distinct segmentation schemes (described in detail later). A network screening approach was used to analyze each network and each produced remarkably different ranking lists based on the safety performance of each segment. It is important

[1] While these break points are necessary for political or for funding reasons, sometimes the breaks are meaningless with respect to the safety of the roadway.

to realize that one of the segmentation approaches is the most likely to produce a priority list that, when improved, will lead to a greater reduction in crashes.

Whether choosing attributes or changing segment length, the start- and end-points of segments are likely to change. Therefore, the results of an analysis can be affected simply by changing the spatial domain of the network. This concept is well known in other disciplines. Geographers refer to this phenomenon as the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). The concept is also exemplified in political boundary modification (gerrymandering). The same concept is also referred to as the *scale effect* by GIS software such as ESRI's ArcGIS which describes *"The scale effect exhibits different results when the same analysis is applied to the same data, but changes the scale of the aggregation units."* (ESRI, 2017).

This concept is discussed by transportation engineers in recent research that examines macro-level safety level analysis (Lee et al., 2014). At the macro-level, Traffic Analysis Zones (TAZs) are used as the spatial unit for analysis. These zones are formed from census blocks and are therefore typically smaller than county boundaries. Census blocks are the smallest geographical unit collected by the US Census Bureau.

In contrast, when performing network screening, it is beneficial to analyze data at the segment level as opposed to points in space as in TAZs (consider the difference between a line and a polygon). At the segment level, it is helpful to employ a linear reference system to integrate roadway and crash data. The FHWA has identified the need for increased use of GIS for safety analysis as many agencies still rely on legacy, non-spatial data storage or face administrative or technical obstacles (FHWA, 2013). Due to the nature of linear networks, this research relied on experience in both highway safety and spatial analysis.

The network screening approach in the HSM requires that a roadway should be divided into homogeneous segments based on engineering judgment and using certain roadway attributes. The HSM suggests a minimum length of 0.10 miles (0.16 km), but the manual does not offer further guidance or statistical techniques to help researchers

decide what length to use (AASHTO, 2010). This work explores the problem of identifying the most important variables when considering roadway segmentation for safety analysis. The research also provides guidance on selecting a segment size and what attributes should be used to create segments.

1.2. Research Objectives

Traffic safety professionals should be given guidance as to how roadways should be segmented to maximize the quality of safety performance functions and the network screening process. A primary objective of this dissertation is to develop guidance for specification of roadway segmentation in safety analysis. A key aspect is to explore the trade-offs between homogeneity and segment length.

This research seeks to explore three main aspects related to roadway homogeneity, segment length, and safety modeling when performing safety analysis:

- What are the statistical implications of segment *length* when performing safety analysis?
- What are the implications of roadway *homogeneity* on safety analysis?
- What are the trade-offs between *homogeneity* and segment *length* on safety analysis?

The outcome of this research offers a better understanding of how the segmentation and homogeneity of a network affect highway safety. This information provides guidance to safety practitioners as to which segmentation should be used in safety analysis depending on user perspective. The resulting methodology offers safety practitioners a set of guidelines and tools to help improve network screening techniques. These methods can be extended to other states' data and needs.

1.3. Paper Organization

This research is organized to address the three main aspects discussed in the previous section. Following this introduction is a literature review (0) with two primary focuses: segment length and roadway attributes. The next four chapters are described below.

0 explores the impact that changing segment length has on the quality of safety performance functions. A network of rural parkways was used in an effort to isolate the effect of segment length without introducing the effect of changing attributes. Parkways in Kentucky are functionally similar to interstates as they tend to be flat and straight with consistent roadway geometrics making them mostly homogeneous.

Chapter 4 tests changes in roadway attributes on Kentucky roadway data. In contrast to 0, segment length is not specified; instead, the length is defined by the selected roadway attributes (recall the resulting segmentation from Figure 1). A by-product of this exploration was the creation of a tool that automates the development of SPFs. A key advantage to this method of SPF development is the near-immediate feedback. The geometric attributes of a roadway network can be adjusted and the resulting SPFs can be quickly evaluated using a variety of metrics. These metrics also help identify data errors that can easily go unnoticed using more passive techniques.

Chapter 5 combines the efforts of two previous chapters by changing both length and roadway attributes while evaluating the resulting SPFs. The automation tool provides a visualization technique for this analysis allowing SPFs to be evaluated along two dimensions: length and attributes.

Finally, Chapter 6 summarizes the findings related to length and attributes in the context of highway safety. Recommendations are provided along with a framework for helping to develop an ideal SPF.

Chapter 2. Literature Review

The following sections describe the current state of the art related to roadway segmentation. Two primary areas of research for this analysis are segment length and roadway attributes. The first section pertains to the length of a segment and its impact on safety – generally these are fixed length segmentation techniques. The other section discusses how the selection of attributes relates to safety. These segmentation techniques are mostly variable length where the attributes of the road (or crashes) control the start and endpoints (and therefore the length) of the segment.

2.1. Segment Length

Previous work on roadways in the state of Iowa has demonstrated that the choice of segment length significantly influences the identification of high-crash locations (Cook et al., 2011). Geyer et al. (2008), summarizing California's data, found that segment length could affect the consistency of high crash identification. Segment length can also affect the outcomes of safety analysis for both extreme long and short roadway segments (Lu et al., 2013). For example, if segment lengths are chosen based on roadway attributes, on limited access roadways this may result in very long segments because there is little variation in attributes over long distances. Yet, using long segments for analysis may be inappropriate for two reasons: it would be economically impractical to improve them due to their long lengths, and only a small portion of the segment may, in fact, require improvements.

On the other hand, using shorter segments can result in higher crash variations, and these fluctuations can introduce more uncertainty into SPF development. Srinivasan et al. (2011) showed that the EB method performs better with longer segments. Previous work has indicated that segment length can affect SPF development, the identification of high-crash locations, and feasibility of improvements, however, there is little guidance on setting optimal segment lengths, or if there should be statistical methods to define segment lengths. The *Guide for Producing usRAP Star Ratings and Safer Roads Investment Plans* suggests a minimum length of 2 miles (3.2 Km) for rural

areas, 1 mile (1.6 Km) for semi-urban, and 0.5 miles (805 m) for urban areas. However, the guide specifies no upper limit for length (usRAP, 2012).

The accuracy of recorded crash location is also a factor to consider when identifying segment length. Green and Agent (2011) found that up to 8 percent of crashes may be incorrectly located by over 500 feet (152 m). Further, safety analysis based on data coded to very short segments will be more sensitive to errors in location (Ogle et al., 2011, Qin and Wellner, 2012).

When developing homogeneous segment lengths, other important considerations are roadway attributes and factors relevant to the safety study (e.g., traffic volume, shoulder width, number of lanes). As the number of roadway attributes increases, the length of homogeneous segments declines. This reduction can be quite significant. Shorter segments typically reduce the statistical robustness of SPFs (Souleyrette et al., 2007). Due to the costs associated with constructability and mobilization, shorter segment length also diminishes the practicality of applying a treatment.

Engineering countermeasures are applied to a specific roadway type based on roadway attributes and factors. Some countermeasures might only be applied over a short distance, such as the installation of a guardrail to prevent run-off-road crashes or shield a roadside from hazards. Other countermeasures, such as centerline rumble stripes, may be applied over much longer sections of roadways (Qin and Wellner, 2012). Crash analysis or pre-selected countermeasures can dictate the roadway data necessary to build a homogeneous network. As Koorey (2009) explained, the segmentation approach is often based on data availability.

In addition to potential countermeasures, another critical factor for determining what roadway data are required is a user's application. For instance, a state highway authority may segment a network based on highway district boundaries to more equitably allocate funding to each district. A list of hazardous locations, therefore, might

need to be stratified by district despite the fact that there may be more hazardous locations in one district compared to another.

While a particular segmentation scheme may be appropriate for highway agency use, it could be inappropriate to disseminate that information to the public, which may view state transportation agency segmentations as arbitrary. This is a primary consideration of risk mapping, which is one of the protocols the Unites States Road Assessment Program (usRAP) uses to create thematic maps that inform motorists about the crash risk associated with different roadway segments (Harwood et al., 2015).

2.2. Segment Attributes

Segmentation of the roadway is often dictated by the attributes chosen based on the analysis performed (Cafiso et al., 2008; Borso et al., 2014). The pattern revealed here is that the attributes of the roadway often control the segmentation used. Ideally, safety professionals could be offered guidance as to the segmentation length and technique based on the safety analysis to be performed. For example, a specific segmentation technique and length might be recommended for the implementation of cable barriers whereas another technique and segment length might be more appropriate for rumble strip installations. For each recommendation, the segment length, roadway characteristics needed, and crash type could be clearly defined.

Network screening requires segmentation of a road network so that each segment can be analyzed. The roadway geometrics and traffic characteristics are typically defined as line events along a route. These events are typically divided when more than one roadway attribute is used. For example, a two-lane roadway segment with constant roadway geometrics (e.g. shoulder width, presence of a median) but with a change in the traffic volume somewhere along the segment would be treated as two segments separated at the point of the traffic volume change.

There are many network screening techniques described in the literature that are used to identify roadway segments. Sliding Moving Window, Peak Searching, Continuous Risk Profile, and Latent Class Clustering are among the most referenced

techniques. In each technique, a quantitative comparison is made to determine the start and end points to be used in the safety analysis. Much of the research determines an ideal segmentation technique based on the roadway data used. For instance, Borsos et al. (2014) based their segmentation technique on data from AADT, road width, shoulder width, horizontal curves, and speed limit. Cafiso et al. (2008) determined that a fixed length segmentation technique having two curves and two tangents in each segment provided the best results. A data-intensive collection process was used to obtain horizontal and vertical curvature of the roadway and a review process to assess roadside hazard. The authors underscore that there are a variety of methods to create segmentation yet there is no widely used method. Table 1 summarizes recent research related to segmentation techniques as well as the data used to create segments.

Table 1. Summary of Network Screening Techniques Including Year and Author

Segmentation Technique	Data Used to Create Segment Endpoints	Reference	Year
Continuous Risk Profile	traffic volume, collision data, safety performance function	Kwon et al.	2013
Sliding Moving Window	traffic volume, collision data, safety performance function	Kwon et al.	2013
Peak Searching	traffic volume, collision data, safety performance function	Kwon et al.	2013
Fisher's Clustering	Crash data	Lu et al.	2014
Change in roadway data	Road width, radius of curvature, shoulder width, number of lanes, traffic volume, posted speed limit	Borsos et al.	2014
5 different segmentation techniques	Volume, radius of curvature, vertical gradient, type of section, roadside attributes	Cafiso et al.	2012
Latent class clustering	Crash data	Depaire et al.	2008
Variable length	Volume, roadway geometrics, driveway density, roadside hazard, curves, etc.	Koorey	2009
Fixed length	Volume, roadway geometrics, driveway density, roadside hazard, curves, etc.	Koorey	2009
Sliding window	Window size, crash data	Qin and Wellner	2012

The segmentation technique used has also been shown to impact safety analysis. Five segmentation approaches were compared and the goodness of fit of the SPF was used to evaluate each approach (Cafiso et al., 2008). Their evaluation determined that a fixed length segment with two tangents and two curves resulted in the best fitting SPF. Consistent with the studies described in the previous section, the segmentation technique resulting in the shortest segments performed the worst. In contrast, a New Zealand study showed that a variable length segmentation is preferred over fixed length (Koorey, 2009). The author contends that such an approach is computationally simpler when dealing with raw attribute data as compared to fixed length segmentation, which requires weighting of the attributes to fit predetermined lengths. Koorey (2009) also points out that despite the need for such a step in network screening, the guidance on segmentation is very limited. Qin and Wellner (2012) agree that a sliding window (variable) method works better than fixed length, and adds that with the prevalence of the EB method and the use of the HSM there is a need to understand the effect that segmentation has on safety analysis. Moreover, Qin and Wellner (2012) caution that segments based on changes in roadway data could introduce bias into the safety analysis.

Kwon et al. (2013) compared three network screening procedures and two segment sizes to determine which method performed the best at hot spot identification. The performance of each was rated based on metrics that revealed how well the method identified previously known hot spot locations. A method scored higher, for instance, if it was able to identify more of the previously known hot spots in the same number of ranked sites. Other metrics included how many miles of roadway needed to be reviewed to identify the previously known hotspots and a measure of hotspot detection efficiency. These metrics also translate into more effective use of a safety engineer's time as they would have fewer sites to review. The study also points out that the different guidelines (as defined by a state transportation agency) used to create the segmentation can result in different SPF models. The study found that the Continuous Risk Profile (CRP) screening method out-performed the Sliding Moving Window and

Peak Searching methods. The CRP method uses a weighted moving average to filter out noise in the data and then a corresponding SPF to determine which segments have excess crashes. Additionally, it was found that a smaller segment size increased the number of sites that required review in order to identify previously known hotspots.

Crash data have also been exclusively used to determine segmentation. Clustering techniques have been applied to crash data to identify roadway segments that reduce the heterogeneity of the crash distribution (Lu et al., 2013; Depaire et al., 2008). In these studies, the segments are defined by the locations of crashes based on similarities in the crash data. Lu et al. (2013) compare the goodness of fit of SPFs developed by three screening methods: fixed length, variable length, and Fisher's clustering. Fisher's clustering is a technique that creates a segmentation based on sections with similar crash distributions, and it produces the model with the highest predictive performance of the three. The authors indicate that Fisher's clustering may also identify roadway segments where changes in geometry could be the cause for a high crash location. For example, a change in speed limit or in the number of lanes may contribute to the safety performance of a segment such as in a transition zone. Using traditional screening techniques, such changes would result in dividing the roadway segment in favor of roadway homogeneity; however, a clustering technique could identify these segments (Lu et al., 2013). Admittedly, they explain that this technique was only applied to freeways and should be further studied for other highway facilities and should be expanded to include multiple variables during the calibration process.

Another advantage to clustering techniques is that using a specific crash type for analysis may mask an underlying contributing factor. For instance, an increase in injury risk may exist for truck crashes on Sundays and holidays, however, research has shown that an analysis of all crashes (not just truck crashes) can hide the injury risk observed in truck crashes (Valent et al., 2002).

2.3. Summary

There are various methods applied to segmentation, yet there is no apparent preferred one (Cafiso et al., 2013). The significant amount of work which has been completed recently trying to identify the ideal segmentation shows the level of interest in improving this aspect of safety analysis. The research reviewed does not indicate consensus regarding the best way to create a network screening that allows for economic and precise roadway crash data analysis. Cafiso et al. (2008) and Borsos et al. (2014) agree that there are various methods available yet there is currently no consensus on the best method to utilize. Researchers have looked at factors such as segmentation technique and length, but all recognize that these variables have some uncertainty when applied to safety analysis. Qin et al. (2012) demonstrate that while segment size has influence over safety analysis, it is not the only factor. The authors continue that segment length is a complex subject and other factors can influence segment length (e.g. the countermeasure or geographic extent). Koorey (2009) points out that the advantages of variable over fixed segmentation length diminish when segment sizes are small, but it is not clear what the minimum length should be.

Previous research has conclusively demonstrated that segment length can significantly affect both SPF development and network prioritization screening. Research that uses segment lengths that are inappropriately or casually selected without proper justification may generate inaccurate models — just as models based on poorly chosen statistical techniques may produce dubious results. This research addresses this issue and identifies potential segment lengths, and attributes which should inform the establishment of segment lengths, in order to improve SPF prediction and network screening procedures.

Chapter 3. Optimizing Segment Length

This chapter explores the effect that segment length has in the context of highway safety using network screening. The chapter begins with a primer on safety performance functions, which is the basis for the HSM's network screening approach. Next, it is demonstrated that a road network can be split using three different segment lengths to produce three separate network screening analyses. This is followed by the methodology and results of two scenarios, each testing various segment lengths. Lastly, the effect that segment length has on network screening is discussed.

3.1. Safety Performance Function: A Primer

The HSM has facilitated the adoption of new approaches by safety professionals to address highway safety since its release (AASHTO, 2010). Highway safety has traditionally been measured using number of crashes, crash rates, crash costs, or a combination of those metrics. High-crash locations are selected based on somewhat arbitrary ranking or by comparison of crash rates to a critical rate factor. All methods have demonstrable disadvantages, particularly in network screening (Wu et al., 2012). Most notably, none of these methods account for regression-to-the-mean or selection bias (AASHTO, 2010; Persaud, 1984). When observed in crash data, these biases can produce misleading results when not corrected for. Traditional crash analysis relies on crashes normalized by exposure—typically traffic volume—to create a rate. However, the use of rates erroneously assumes a linear relationship between crashes and volume (Srinivasan et al., 2011). Most SPFs exhibit an exponential relationship between crashes and exposure (only when the exponent equals 1, a constant rate is observed across the volume spectrum). In general, segment length is treated as an offset in that it is directly proportional to the crash prediction. Equation 1 describes the relationship between crash prediction, traffic volume, and segment length.

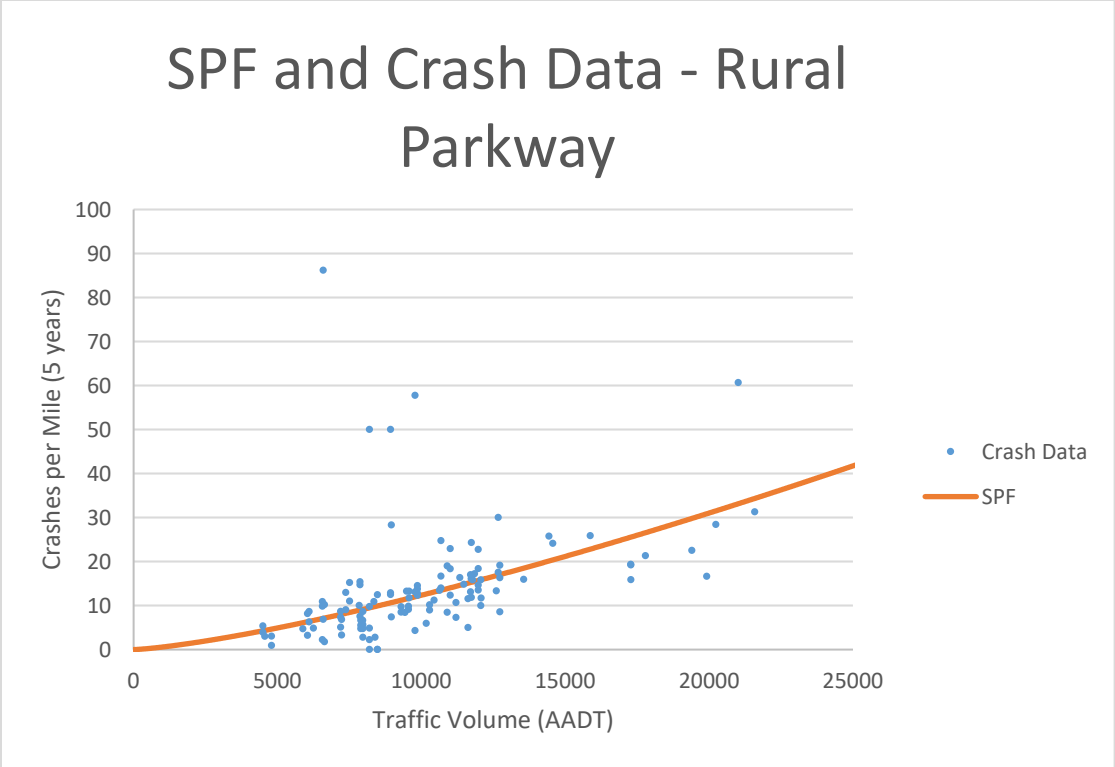
$$y = L * e^{a} ADT^{b} \qquad (1)$$

where:

y=estimated crashes
L=segment length (miles)
ADT=traffic volume (AADT)
And a and b are coefficients that describe the interaction among length, AADT, and the estimated number of crashes.

SPFs are models used to predict crashes based on traffic volume and other factors. A common modeling technique is to fit a statistical distribution to crash data (Zhang et al., 2007). A Poisson distribution is an ideal description for a specific roadway segment. In this case, the variance is equal to the mean. However, at the network level (i.e., across several of roadway segments) crashes exhibit a large variance and a small mean (i.e., the variance is greater than the mean). This is known as overdispersion. A more appropriate distribution is the Poisson Gamma or negative binomial distribution, which produces two parameters: the mean and the overdispersion (or shape) parameter. In this research, overdispersion is referred to as either *theta* or the inverse dispersion parameter *k*, where *k*=1/*theta*.

Figure 2 compares two SPF scatter plots: one with an SPF based on rural parkways (top) and one based on rural 4-lane divided (non-interstate and parkway) roads (bottom).



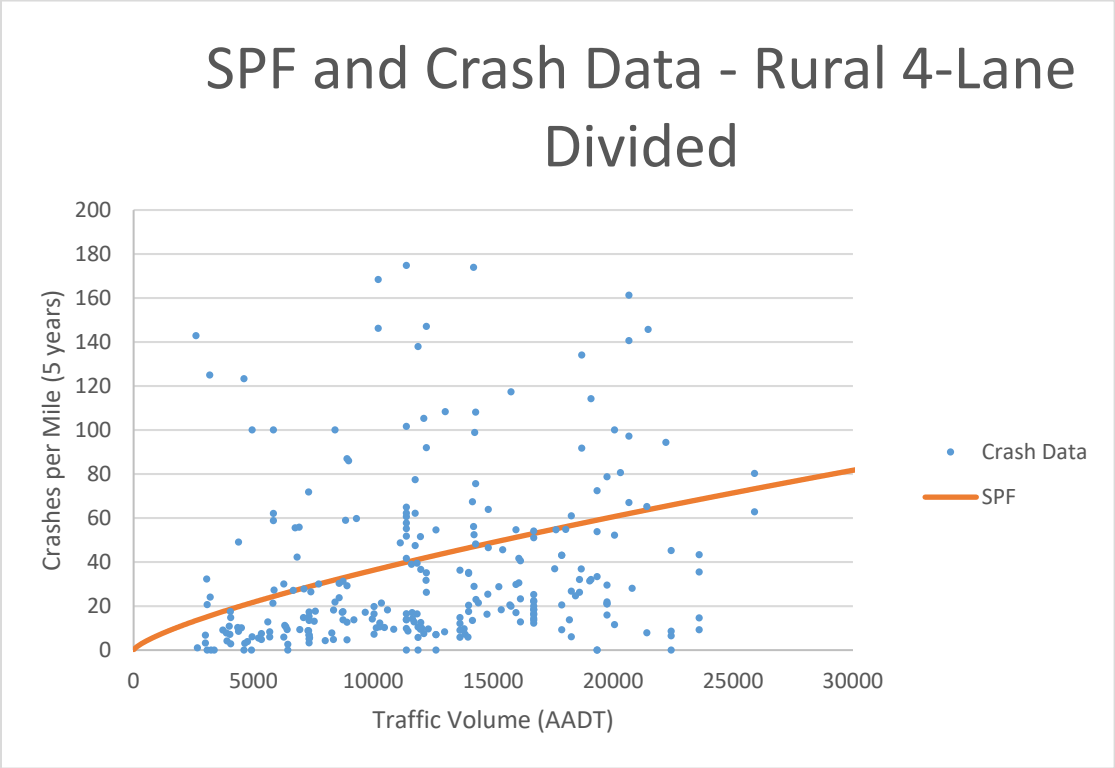


Figure 2. Comparison of Overdispersion for Two SPFs

The rural parkway model has an overdispersion (*theta)* about 8 times as high as overdispersion for the rural 4-lane divided model. In this context, a higher *theta* indicates less overdispersion and, hence, a better model fit.[2] This is expected, as rural parkways are generally homogeneous with respect to roadway geometrics. In contrast, other 4-lane divided roadways vary in design attributes and lack the homogeneity of the parkways. This design heterogeneity contributes to the overdispersion, as these changes in geometry are excluded from the model. This omission is typically detectable using cumulative residual (CURE) plots.

A CURE Plot is graph of the cumulative residuals versus an independent variable (typically traffic volume) (Srinivasan and Bauer, 2013a). Residuals are the difference between actual crashes and the SPF prediction at a given site. Plotting the residuals (not cumulative) versus a variable such as traffic volume produces a graph as shown Figure 3.

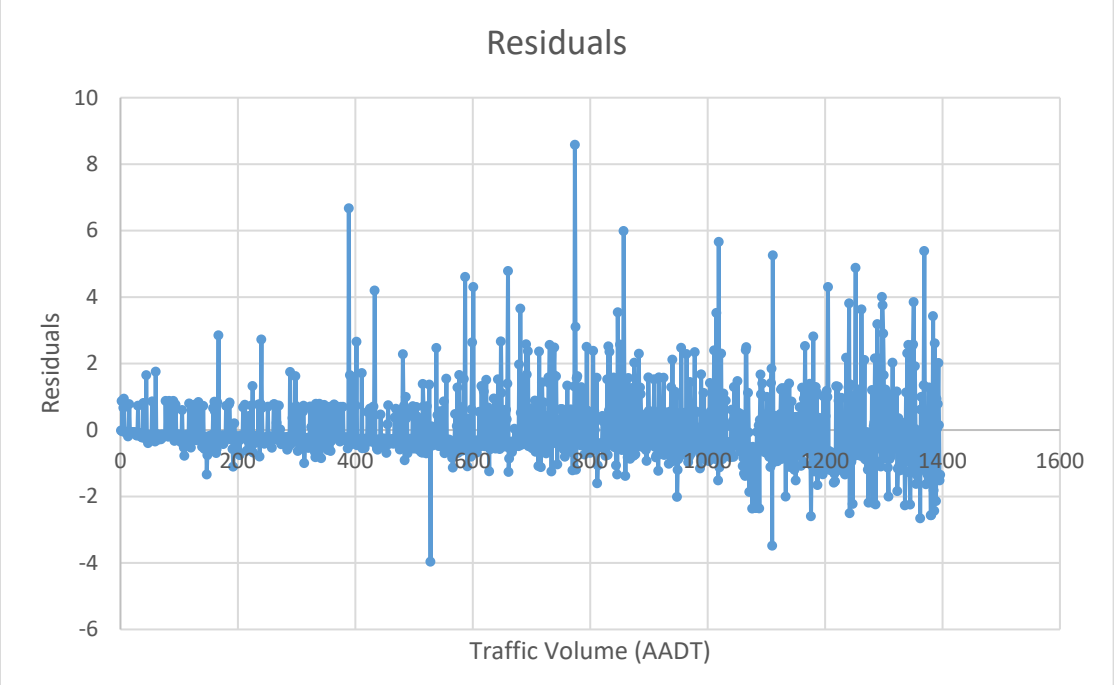


Figure 3. Residuals Versus Traffic Volume (AADT)

[2] It is likely that this is the reason some references prefer the use of *k*, the inverse dispersion parameter. It is perhaps more intuitive to relate an increase in overdispersion with an increase in the parameter.

The farther the point is from the x-axis, the greater the residual (i.e., the worse the model's prediction). In some cases, the actual crashes are more than the SPF predicted (positive residual) and sometimes below (negative residual). There can be more than one residual with the same AADT (but this is not easily shown in Figure 3).

The cumulative residuals, however, offer a better indication of when several residuals are stacked at the same traffic volume. Using typical network screening techniques, it is very common to have a long stretch of road with the same traffic volume, which could result in several segments with identical AADT. The cumulative residuals are computed by adding the residuals from a roadway segment to that of the previous site's cumulative residual. This cumulative summation is computed with the segments ordered by traffic volume (or in some cases segment length). Plotting the cumulative residuals versus traffic volume results in a CURE plot as shown in Figure 4.

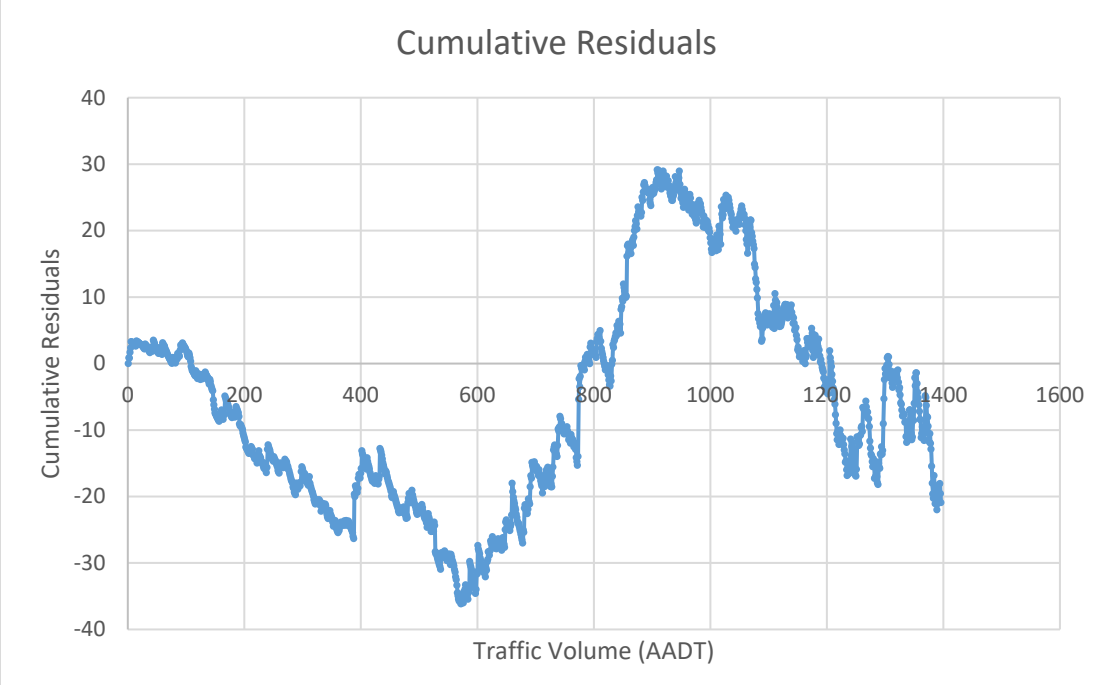


Figure 4. Cumulative Residuals Versus Traffic Volume (AADT)

Statistically, oscillation about the x-axis is expected due to random error – approximately following a normal distribution[3] (Hauer, 2015). Anything that is not random error will deviate from the oscillation and can indicate a bad model fit or omitted variable bias (discussed in more detail in Section 4.2.1). The overdispersion parameter is useful in CURE plots too as it helps define confidence boundaries (Hauer and Bamfo, 1997). The boundaries are defined by two standard deviations (positive and negative). The data points in the CURE plot within these boundaries are more likely to be explained by random walk.

The assessment of CURE plots, while somewhat subjective, can provide high-level screening to the SPF development process. When evaluating CURE plots, there are several aspects that indicate a good model (each demonstrated below).

- Oscillating around the x-axis indicate; ending near zero.
- Free of outliers as they can adversely affect the model parameters.
- The cumulative residuals should rarely transgress the confidence bands.
- Minimal drifting; either upward or downward.

Despite the subjectivity of these metrics, there are a few key advantages to this method of assessment. This evaluation is graphical and therefore can be performed quickly, especially when comparing several CURE plots at once. Figure 5 shows an example of a comparison of several CURE plots using Windows Explorer's thumbnail images.

[3] It should be noted that the approximately normal distribution is applied to the residuals and not the actual crash data. It is well known that a normally-distributed error term is typically not observed in crash count data (Zhang et al, 2009).

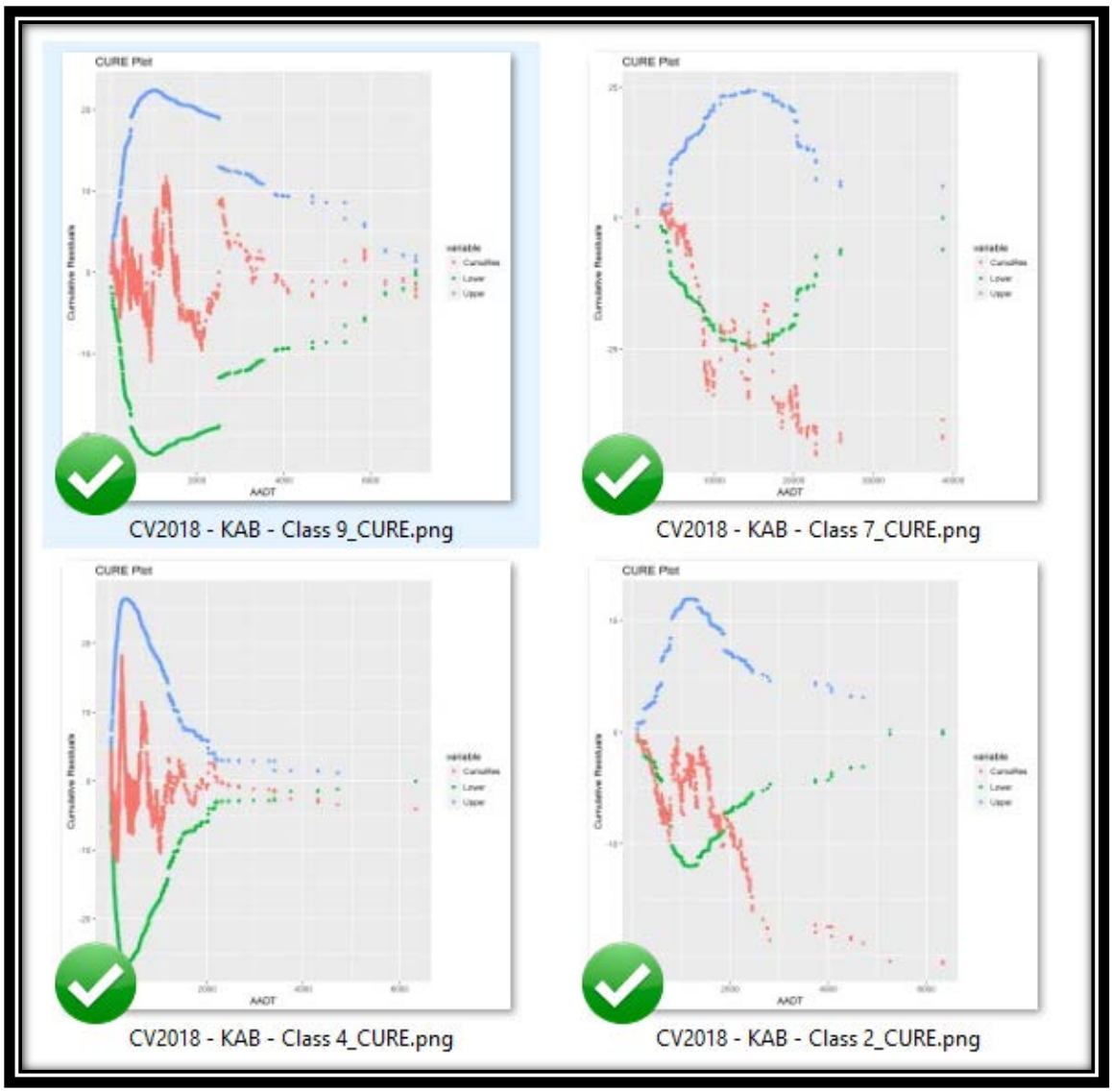


Figure 5. Comparing CURE Plots using Thumbnail Images

Another advantage to this assessment is that most of the aspects in a CURE plot that lead to a good model are mutually beneficial. That is, oscillation around the x-axis tends to produce a CURE plot without drifting. Likewise, the lack of large outliers tends to produce CURE plots with residuals within the confidence bands. Similarly, other combinations of these aspects lead to the same relationships.

The following figures provide examples of CURE plots that exhibit indicators of both desirable and undesirable models. In each figure, the red dots represent the

cumulative residuals with the blue and green dots representing the upper and lower confidence boundaries, respectively.

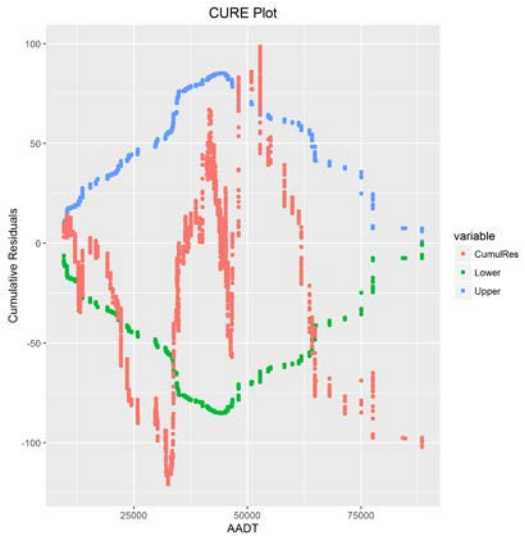


Figure 6. A CURE Plot with Good Oscillation and Outside of the Confidence Bands

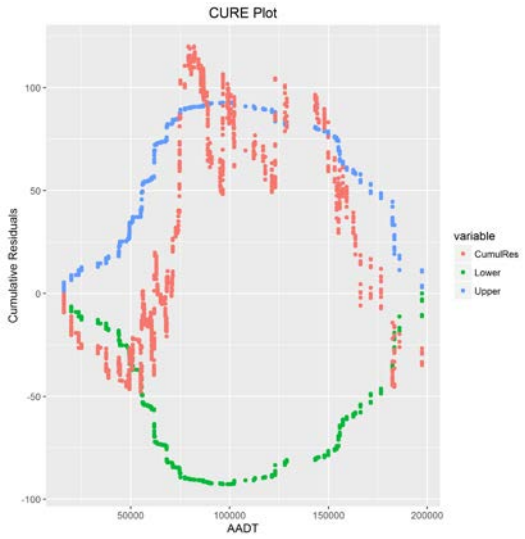


Figure 7. A CURE Plot with Poor Oscillation and Outside of the Confidence Bands

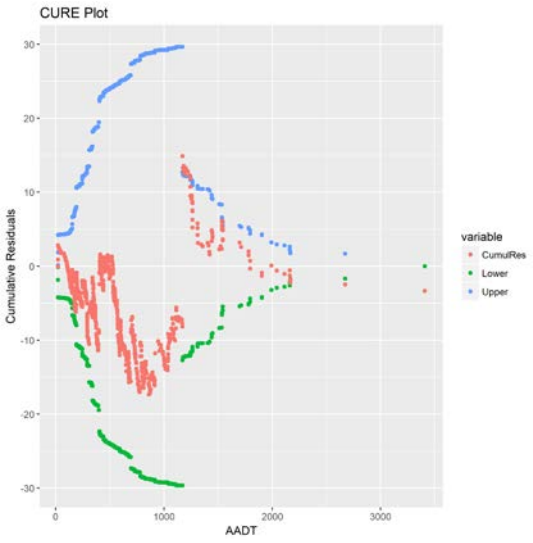


Figure 8. A CURE Plot with a Likely Outlier and Inside of the Confidence Bands

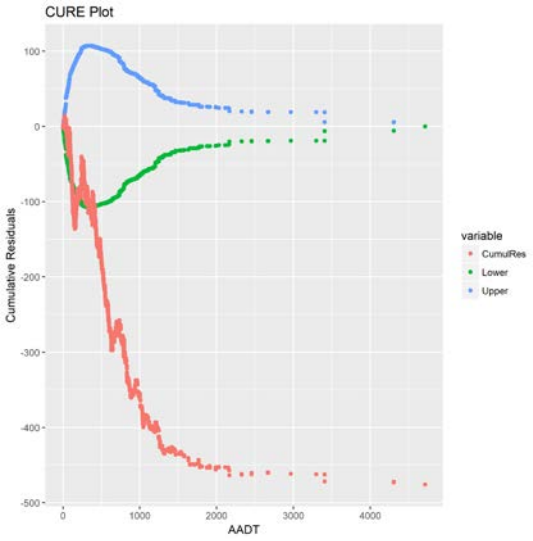


Figure 9. A CURE Plot with Significant Drift, no Oscillation, and Outside of the Confidence Bands

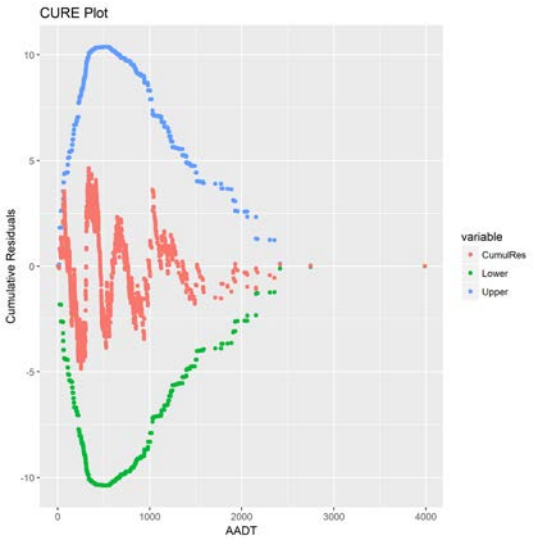


Figure 10. A CURE Plot with All Desirable Aspects

In addition to improving the models, CURE plots can also helpful in detecting data errors. An unexpected result was observed when comparing two models. The exclusion of very short segments had a dramatic effect on model performance – specifically with regard to omitted variable bias. In this application, this effect was counterintuitive. These segments varied in length between near zero and 0.7 miles. Consider the two CURE plots in Figure 11, with and without short segments.

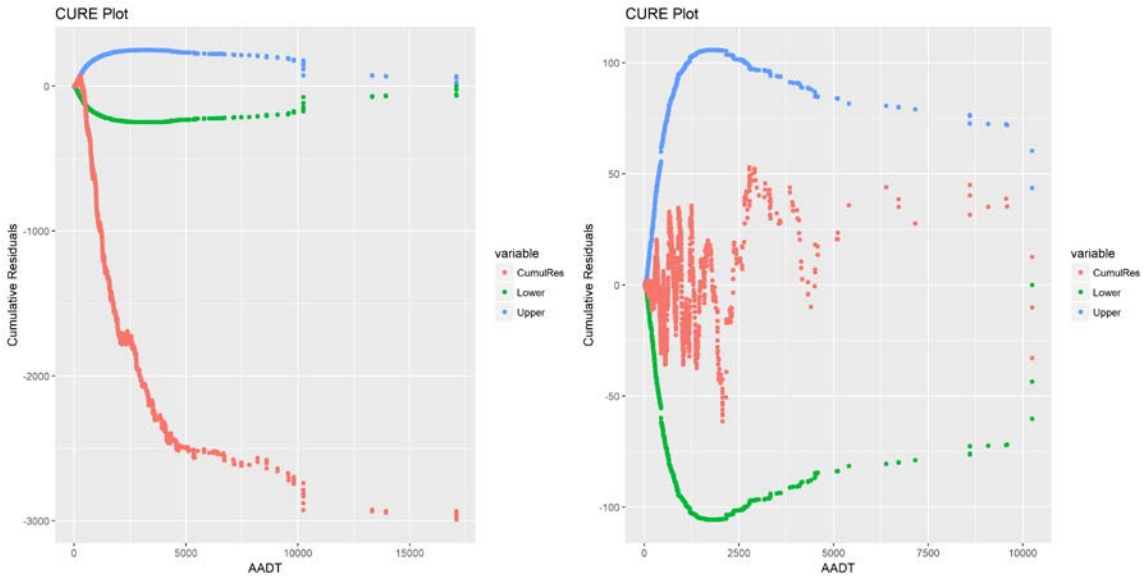


Figure 11. CURE Plots for a Rural 2-Lane with (left) and without Short Segments (right)

Further inspection revealed that short segments are not actually contributing to this bias but rather were suggestive of a data error. The segments were plotted on a map in an effort to better understand why the short segments (referred to as remainders in Figure 12 below) were sensitive to omitted variable bias.

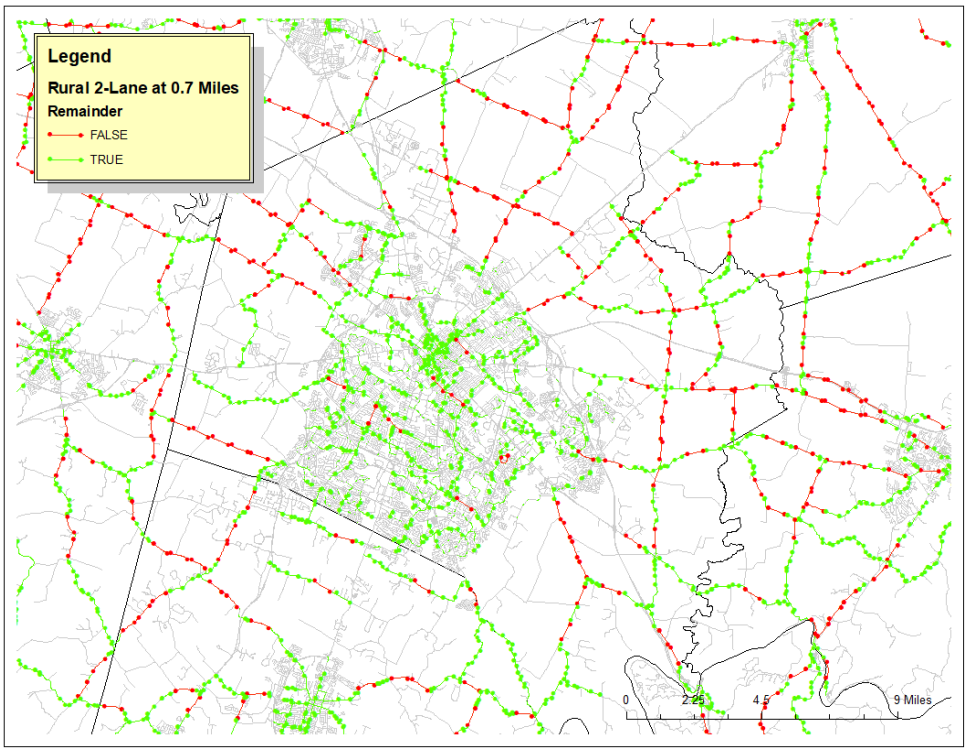


Figure 12. Segments with and without Short Segments Around Fayette County Kentucky

The plot quickly revealed that urban segments were erroneously included in the model. A high concentration of short segments (green segments) were clustered in downtown Lexington. These segments should have been filtered out as they were not rural. The inclusion of urban segments introduced significant heterogeneity in the network. Urban segments are also typically shorter than rural segments (city streets have more breakpoints with changes in volume or changes in geometrics being more frequent). Because of this, the length filter likely excluded more urban than rural segments resulting in model improvement.

As stated, the assessment of CURE plots can provide a high-level screening when comparing or improving models. Further refinement is achieved through comparing other goodness-of-fit metrics (discussed in 3.3.3). Once a model is selected, the parameters can be used to predict crashes for similar roadway types in the network screening process. Network screening using the HSM's methodologies addresses many of the disadvantages of the traditional methods. SPFs are developed to better characterize the relationship between crashes and traffic volumes as well as other variables. Empirical Bayes (EB) addresses regression-to-the-mean bias by using actual crash data and the overdispersion parameter to adjust the expected crash experience at a site. This adjusted value is a more realistic measure of a site's safety performance. More importantly, it describes the magnitude of crash reduction that could potentially be achieved. In Kentucky, this is referred to as "Potential for Crash Reduction" (PCR). Other states use the term "potential for safety improvement" synonymously.

3.2. Demonstration of the Problem

A case study was conducted to demonstrate how segment length influences safety planning and to investigate appropriate procedures for defining segment length. Data from Kentucky's HSIP were used to underscore the critical role that segment length plays in network screening. Each year, as part of the HSIP, a priority list of candidate locations for High Friction Surface Treatment (HFST) is prepared. HFST is typically used on horizontal curves to enhance vehicle grip and traction during wet conditions and therefore reduce roadway departure crashes.

Using a single dataset of rural parkways (4-lane divided highways), the HSM's network screening process was used to divide roadways into three distinct segment lengths. The following segmentation models were used (depicted in Figure 13):

- Model 1: A fixed length of 0.3 miles (480 m)
- Model 2: Variable segment length —adjacent segments with the same AADT were combined
- Model 3: Segments from Model 2 bisected at their respective midpoints.

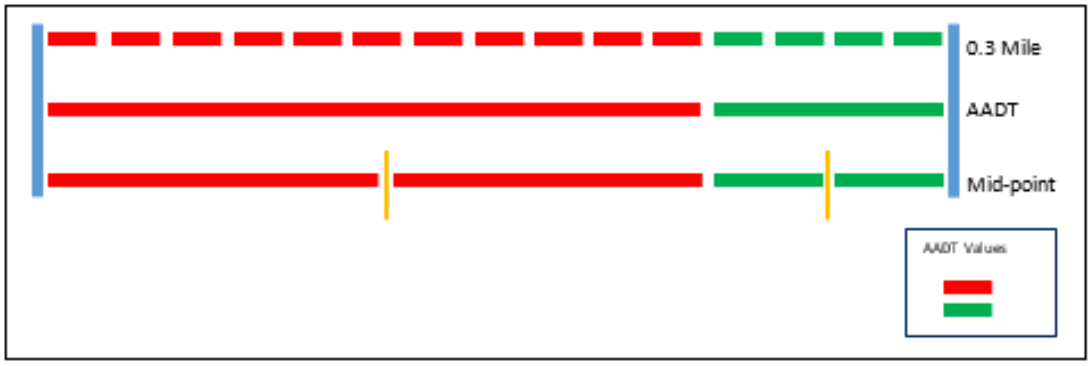


Figure 13. Segmentation Models Compared

SPFs were developed for each of the models using lane departure crashes (all severity levels). The results were used to conduct network screening and develop network prioritizations based on PCR as described above. In practice, the resulting priority lists would be used to identify candidate locations for HFST installations. Preliminary analysis generated three priority lists — one for each model specified above. Table 2 lists SPF and overdispersion parameters for each model (details on these parameters are given in the results section).

Table 2. Comparison of SPF Parameters and Overdispersion for All Three Models

Analysis	SPF Parameter[4]		Overdispersion Parameter (*k*)
	a	b	
Model 1 (0.3 miles)	-4.6***	0.6***	1.82
Model 2 (Combined)	-5.2*	0.7**	0.66
Model 3 (Midpoint)	-4.8**	0.6**	1.01

*95% significance level
**99% significance level
***99.9% significance level

The SPFs are plotted against a range of traffic volume values for each model (Figure 14).

[4] Based on Equation 1for a 5-year period

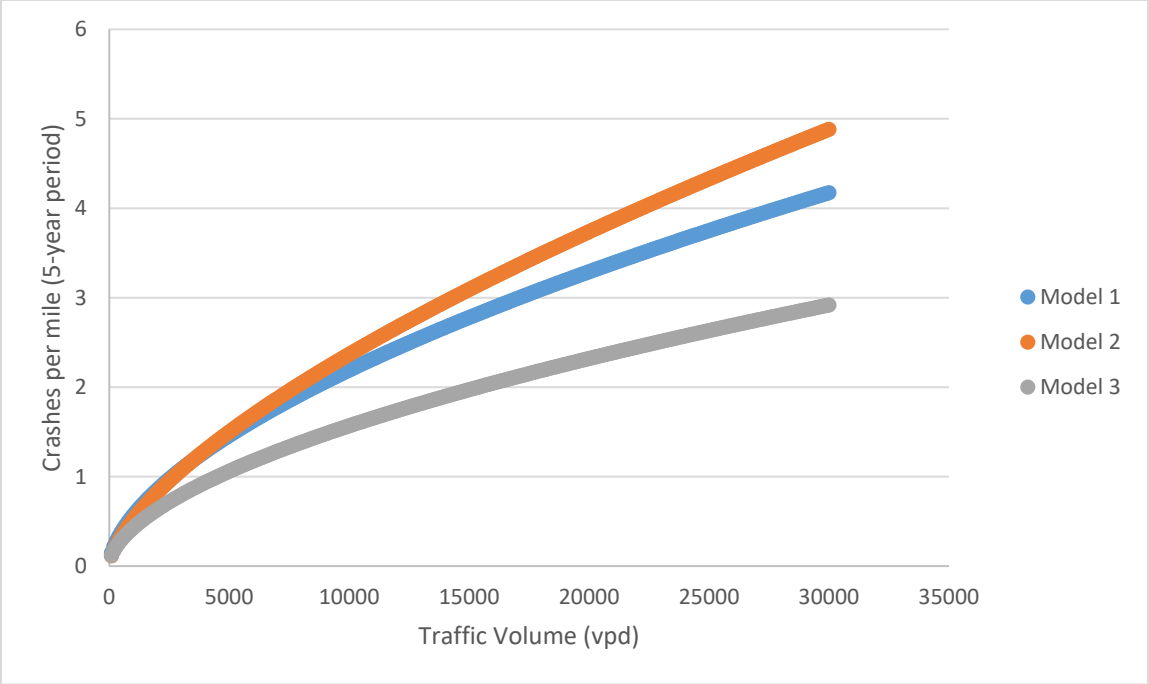


Figure 14. Safety Performance Functions

The top 100 miles (161 km) of the priority lists from each model were compared. The segments identified by each model varied. In some cases, two or all three of the models identified parts or all of the same roadway segments. In other cases, the models identified nearly exclusive segments. Figure 15 shows a map of the roadway segments identified by each model. An offset was used to plot the segments so that viewers can identify where overlap is present along the same routes.

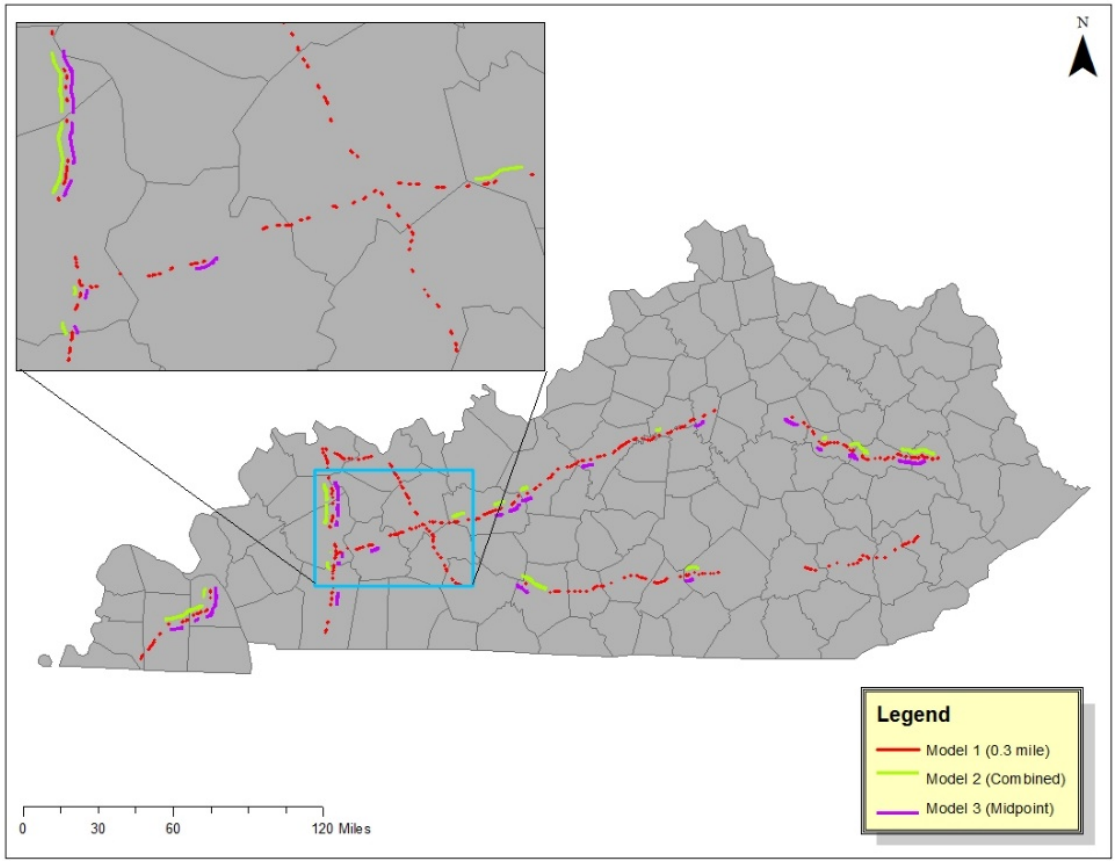


Figure 15. Comparison of the locations of the highest PCRs for all three models (offset used for clarity).

Segment length influenced the priority lists created by the HSM-based network screening process (Figure 15). While each method generated models that overlap with one another to some extent, each produced discrete networks. The overlap (areas where both models identify the same segments) between Model 1 and Model 2 was 18 percent. For Model 1 and Model 3 this was 17 percent. There was significantly more agreement between Model 2 and Model 3 — approximately 85 percent overlap, which is expected as they were based on similar segments. The real implication here is that because all three models produce different results there is a need to evaluate the arbitrary nature of segmentation. Another aspect is that this analysis only considers the first 100 miles of the network screening list. In most cases, states deploy systemic improvements across a much larger number of roadway miles. A key point here is that this analysis was performed on a specific crash type, for a specific countermeasure

application. It is suspected that other specifications would result in even more differing priority lists.

3.3. Methodology

A database of Kentucky's rural parkways was developed for the work to be completed in this research by combining Highway Information System (HIS) layers. Parkways are similar to the interstate system in Kentucky. Representative street images obtained from Kentucky's Photolog[5] are shown in Appendix A. The layers used were Traffic Flow (TF), Functional Classification (FS), and Median Type (MD). Along with the route ID, these layers were used to filter out segments that lacked traffic flow data, included ramp segments, large urbanized areas, and undivided parkway segments (there are very few miles of undivided parkways in Kentucky). The resulting network contained 961 segments representing 480 miles (772 Km) of parkways. For this analysis, other geometric attributes were not included such as lane and shoulder widths. These attributes are similar for rural parkways in Kentucky and therefore result in a homogeneous network that is ideal for this analysis – changes in roadway attributes can adversely affect model development.

The following sections describe the two segmentation scenarios that were used.

3.3.1. Scenario 1 – Rural Parkways with Fixed Length

The parkway network was matched to the crash database file. This analysis used all crash types and crash severities. A program was developed that produced a new segmentation of the network. Roadways were segmented using 16 predefined length categories. The segments were created starting at the beginning of a route and continued until either the route ended, AADT changed, or the length category was achieved. The following length categories were used:

[5] Images obtained from http://maps.kytc.ky.gov/photolog/

- 0.10 miles (161 meters)
- 0.20 miles (322 meters)
- 0.30 miles (483 meters)
- 0.40 miles (644 meters)
- 0.50 miles (805 meters)
- 0.60 miles (966 meters)
- 0.70 miles (1127 meters)
- 0.80 miles (1287 meters)
- 0.90 miles (1448 meters)
- 1.00 miles (1609 meters)
- 1.50 miles (2414 meters)
- 2.00 miles (3219 meters)
- 2.50 miles (4023 meters)
- 3.00 miles (4828 meters)
- 3.50 miles (5633 meters)
- 5.00 miles (8047 meters)

Each record included the route, start and end milepoints, and total number of crashes. The segment was discarded if the resulting length was less than the target length. This was typically the case at the end of route or where a change in AADT occurred. These segments were discarded as they were less than the length category and therefore would affect the segment length. The result was 16 new road networks that represented the same roadway and crash data but each with different lengths.

An SPF was developed for each dataset for use in network screening. An SPF was used to predict crashes based on segment length and AADT for each segment following the form in Equation 1. The SPF predicts crashes over a five-year period while using a single year's AADT. While this did not affect regression results, it impacted the scale of the regression parameters, which is important to recall when comparing the results to other SPFs. Kentucky does not collect AADT every year for all roadway segments. Therefore, a single AADT value was used to represent the five-year period. Previous research has demonstrated that AADT values in Kentucky for one year vary by under a half a percentage point when compared to the previous four years (Green et al., 2015). This very minor change is insufficient to justify using different AADT values for each year, especially considering that this might complicate the segmentation process.

Regression parameters were derived using the statistical program R, which fits the model using negative binomial regression. The resulting SPF and overdispersion parameter were used to conduct a network screening process on the roadway network. The overdispersion parameter measures the degree to which the variance exceeds the mean (AASHTO, 2010). PCR was calculated for each segment using the EB Method, as

recommended by the HSM. The Empirical Bayes Estimate (EBE) was calculated with Equation 2:

$$EBE_i = Weight_i * SPF_i + (1 - Weight_i) * OC_i \qquad (2)$$

where:

$$Weight_i = \frac{1}{1 + k * \frac{SPF_i}{L_i}}$$

SPF_i = predicted crashes at site *i* using SPF (for 5-year period)[6]
k = overdispersion parameter (or 1/*theta*)
L_i = Length of site *i* in miles
OC_i = Observed crashes at site *i*

The PCR at site *i* was calculated by subtracting the predicted crashes (from SPF) at site *i* from the EBE at site *i*:

$$PCR_i = EBE_i - SPF_i \qquad (3)$$

This is represented, graphically, in Figure 16. The green line represents an SPF with E[N] representing SPF_i at site *i*. Similarly, N represents the observed crashes at site *i* (OC_i) and EB[N] represents EBE_i.

[6] The *year* term is omitted from this equation since the data are for a 5-year period and is justifiable for the reasons discussed above.

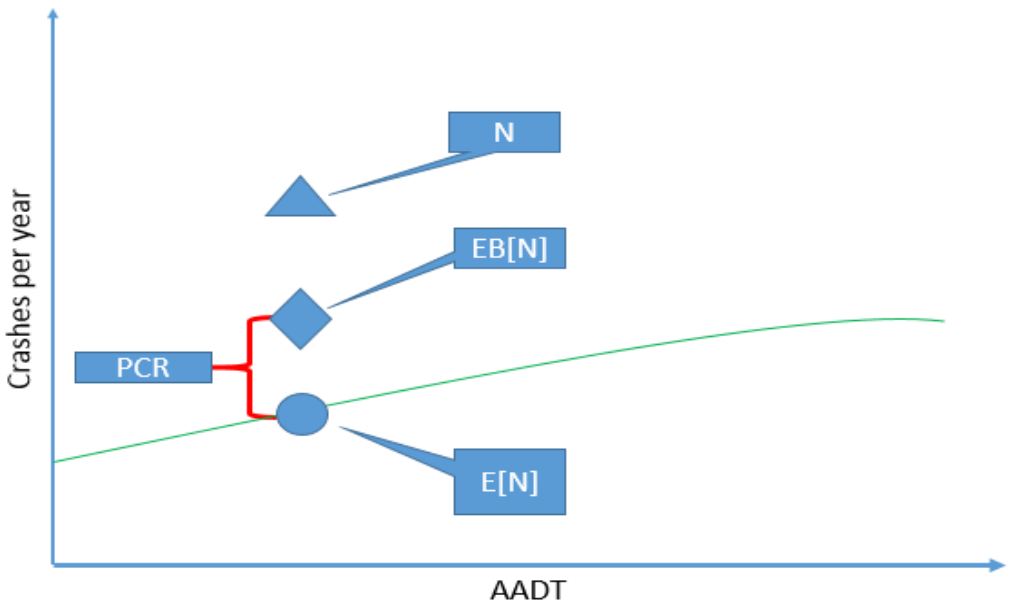


Figure 16. Graphical representation of potential for crash reduction.

The PCR represents the likely number of crashes that could be eliminated with appropriate improvements. Each site can be prioritized by its PCR value. Typically, this list is sorted in descending order, with the top sites having the most potential for safety improvements. In addition to the SPFs, several metrics and descriptive statistics were calculated to evaluate the models as well as CURE Plots (discussed earlier) and scatter plots. It should be noted that the cumulative residuals are plotted versus traffic volume and not length since, in this analysis, length is constant.

3.3.2. Scenario 2 – Rural Parkways with Lower AADT

This scenario used the same procedure to establish segment lengths. However, any segment with an AADT over 15,000 was omitted from further analysis. This decision was motivated by an examination of the CURE Plots from Scenario 1. The CURE Plots tended to stop oscillating about the x-axis above an AADT of 15,000, which is indicative of model bias when the AADT approaches that range. This is discussed in more detail in the results section.

3.3.3. Safety Performance Function Metrics

Each scenario resulted in 16 SPFs. Formulas from an Excel-based SPF analysis tool — FHWA's The Calibrator — were used to generate metrics and compare them. The Calibrator User Guide was referenced for the following metrics in an effort to evaluate the SPFs (Lyon et al., 2016).

- Modified R^2
  - Measures the amount of variation explained by the SPF. Higher values are optimal. Values over one indicate overfitting, which is not optimal.
  - This is a pseudo R^2 — negative binomial regression does not generate a metric strictly analogous to R^2.
- Mean Absolute Deviation (MAD)
  - Measures the average absolute variation between the predicted and observed crashes at each site. Lower values are optimal.
- Akaike Information Criterion (AIC)
  - A measure that considers both goodness-of-fit and model complexity. Lower values are optimal.[7]
- CURE Plot
  - A unique assessment tool for SPF; unlike the other metrics, they provide a measure of the SPF's functional form (Srinivasan and Bauer, 2013a).
  - CURE plots that oscillate around the x-axis indicate the absence of model bias, which is ideal.
  - Outliers can be identified as large vertical jumps.
  - The cumulative residuals should rarely transgress the confidence bands
- Percentage CURE Deviation (PCD)
  - A more objective measure of bias in the SPF model. Values under 5% are statistically significant at the 95% confidence level.
- Maximum Absolute CURE Deviation
  - A measure that represents the largest — positive or negative — deviation (cumulative residual) from the CURE Plot. Lower values are optimal.

3.3.4. Evaluation of Potential for Crash Reduction

In addition to comparing goodness-of-fit metrics for the various length categories, it was also meaningful to compare the resulting segments with the highest potential for crash reduction. A comparison was performed similar to the analysis in

[7] AIC is generally best for comparing different model forms from the same dataset with the same sample size (Geedipally et al., 2008). This measure is therefore inappropriate for comparing different length categories as the sample sizes change. AIC will be meaningful in the next chapter.

section 3.2 to demonstrate how changing the length can affect the resulting priority lists. The PCR was calculated for all segments and the top ten PCRs were mapped and compared to the top ten lists from all other length categories.

3.4. Results

In this section SPF metrics and CURE plots are used to evaluate ideal segmentation lengths with the goal of providing guidance to practitioners on roadway network segmentation.

3.4.1. Scenario 1 Results

Results from Scenario 1 were used as a starting point to evaluate segmentation length categories. The SPF parameters for the 16 SPFs ranged from -5.84 to -5.06 for *a* and around 0.86 for *b*. Values for the metrics discussed in the previous section informed this assessment. Table 17 displays these values for each length category. Total crashes, overdispersion, and sample size are included as well. The least optimal values are in dark grey with more optimal values indicated by lighter shading. The CURE Plots were examined and interpreted to derive information about outliers and oscillation.

Table 3. SPF Metrics and Descriptive Statics for Scenario 1 by Length Category

	Length Category (Miles)															
	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	**1.00**	**1.50**	**2.00**	**2.50**	**3.00**	**3.50**	**5.00**
Segments	4652	2318	1532	1138	895	738	629	543	472	472	263	189	144	107	91	52
k	0.6	0.4	0.4	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1
Crashes	5488	5453	5408	5328	5257	5213	5183	5042	4946	4946	4485	4214	4074	3627	3559	2800
AIC	13716	9152	6991	5692	4769	4155	3690	3262	2944	2944	1783	1349	1080	844	730	442
Mod. R2	0.07	0.09	0.09	0.11	0.11	0.12	0.13	0.15	0.13	0.15	0.25	0.33	0.31	0.31	0.34	0.32
PCD	9.20%	9.58%	8.16%	8.44%	8.60%	7.99%	7.15%	10.13%	8.26%	10.40%	20.91%	5.29%	11.81%	7.48%	1.10%	1.92%
MACD	100.7	101.0	101.7	96.4	94.3	97.0	93.6	92.7	97.6	114.7	110.4	87.8	112.8	91.3	82.0	78.9
MAD	1	1.6	2.2	2.6	3.1	3.5	4	4.1	4.6	4.9	5.9	7	8.1	9.4	10.3	13

The length categories which stand out are 0.7 and 2.00 miles. Among all the categories, these offer the best trade-offs among all metrics. A desirable length should have a high Modified R^2, a PCD ideally under 5 percent, a low MAD, a low MACD. The 2.00 length has better Modified R^2 but a less optimal MAD when compared to other lengths. The data show a few general trends: 1) MAD improves as segment length decreases and, 2) Modified R^2 improves as segment length increases. The overdispersion parameter increases when segment length declines. The HSM suggests that models with a lower overdispersion parameter, *k*, are more statistically reliable (AASHTO, 2010). This suggests that longer segment lengths produce better models; however, this runs counter to the results of the MAD metric. Length-based overdispersion will be discussed later which can help explain this discrepancy. Also, recall that AIC comparisons are better suited when the sample size is constant (e.g. when comparing model forms). Example CURE Plots are shown in Figure 17Appendix B and Figure 18 as representatives of Scenario 1.

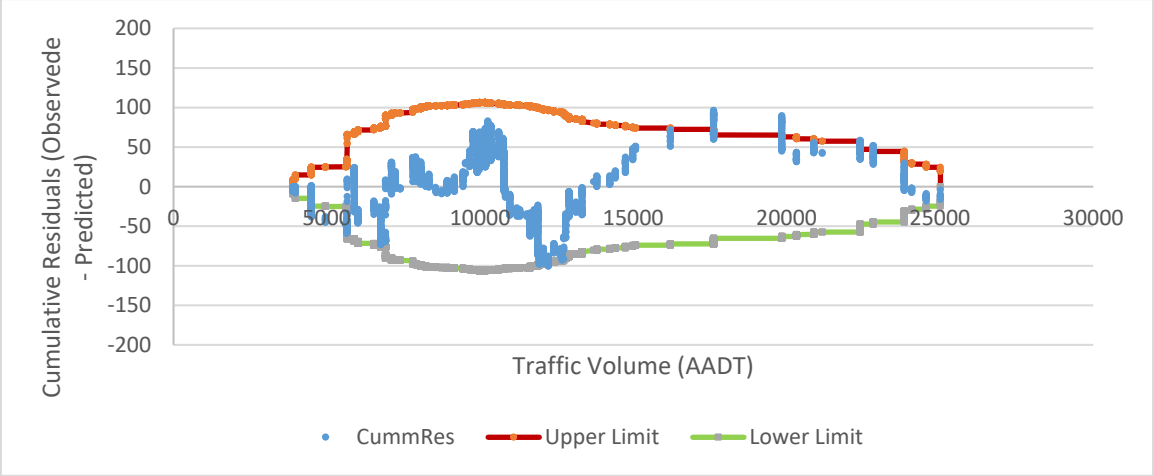


Figure 17. CURE Plot for Scenario 1 at 1.0 mile.

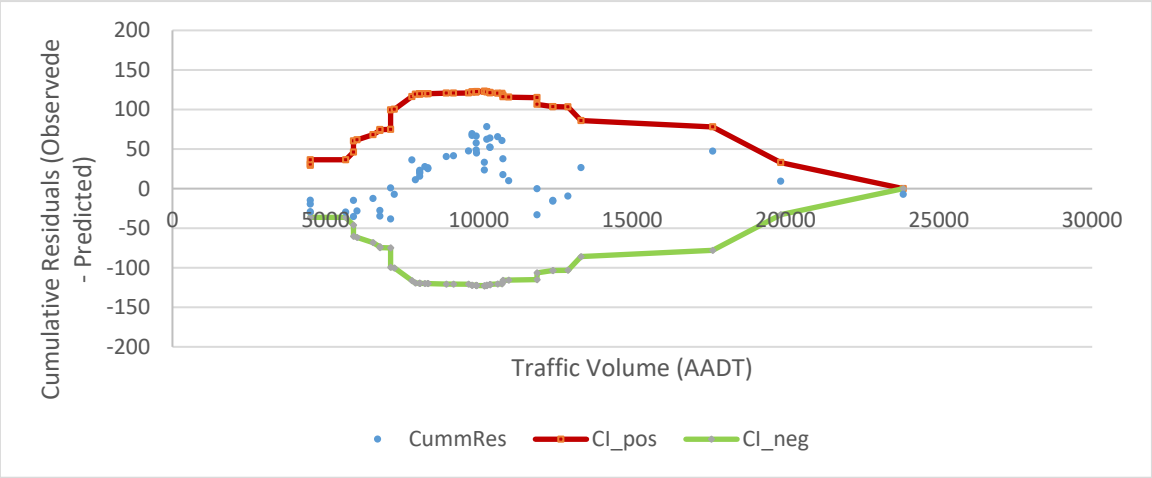


Figure 18. CURE Plot for Scenario 1 at 5.0 miles.

Both CURE plots exhibit the indicators of a good model — values oscillate about the x-axis while staying within the confidence bands. Figure 18 has fewer data points due to the longer segment length. The lack of drift on both plots also suggests little model bias – confirmation of roadway homogeneity for network. All CURE plots for Scenario 1 are shown in Appendix B. As a rule, PCD should be under 5% and it is suspected that it could be lowered by filtering out heterogeneous segments (e.g. curvature, exit/entrance ramps); however, PCD is fairly consistent among the length categories with the exception of the two longest categories. For these categories, it is likely that the segments are so long and many of the small, heterogeneous segments that have short changes in geometry or traffic volume (such as near interchanges) are excluded. Such filtering will be explored in the next chapter.

3.4.2. Scenario 2 Results

The same analysis was repeated for Scenario 2. Recall, this scenario excluded segments with AADT over 15,000. This threshold was based on some CURE plot deviation at higher traffic volumes observed in Scenario 1. This resulted in the omission of about 25 miles of segments. Table 4 summarizes the results of this analysis, and includes the same greyscale shading scheme as in Table 1 (lighter values indicate more optimal results).

Table 4. SPF Metrics and Descriptive Statics for Scenario 2 by Length Category

	Length Category (Miles)															
	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	**1.00**	**1.50**	**2.00**	**2.50**	**3.00**	**3.50**	**5.00**
Segments	4400	2193	1449	1076	847	698	596	514	447	400	249	178	136	101	87	49
k	0.6	0.4	0.4	0.3	0.3	0.3	0.3	0.2	0.3	0.2	0.1	0.1	0.1	0.1	0.1	0.1
Crashes	5362	5329	5285	5204	5136	5097	5074	4932	4835	4849	4389	4105	3982	3545	3492	2731
AIC	13226	8802	6718	5458	4574	3985	3546	3131	2824	2588	1709	1288	1030	805	703	419
Mod. R2	0.05	0.06	0.07	0.08	0.08	0.09	0.10	0.11	0.10	0.11	0.19	0.28	0.25	0.25	0.29	0.24
PCD	9.5%	3.3%	3.4%	6.0%	3.9%	2.7%	0.8%	1.8%	5.1%	5.8%	8.0%	1.7%	2.9%	1.0%	1.1%	2.0%
MACD	133.7	112.2	126.6	133.0	132.6	135.8	125.7	121.8	137.9	139.9	100.5	75.3	78.3	95.2	94.8	58.9
MAD	1.1	1.7	2.2	2.7	3.1	3.6	4.1	4.3	4.8	5.1	6.0	7.3	8.2	9.7	10.4	13.0

The patterns among MAD and Modified R^2 are comparable to Scenario 1. However, the PCD has improved for nearly all length categories; many below the 5% threshold. This is likely due to removing heterogeneous segments with high traffic volumes. The CURE plots were similar to those from Scenario 1, although less deviation was observed. As before, CURE Plots for Scenario 2 are presented in Appendix C.

3.4.3. Evaluation of the Top 10 Segments from All Length Categories

A network screening process was performed on the segments used to develop the SPF for Scenario 1. A PCR was calculated for each segment using EB as described earlier. The segments with the top ten highest PCR were mapped for each length category. These segments are compared using maps in Appendix D. As observed in section 3.2, changing length had a dramatic effect on the locations of segments with the highest PCR. In practical terms, this means that changing the size of a fixed length analysis zone can directly affect the appropriation of safety funds.

The same top ten lists were reviewed and the resulting segments were examined using Kentucky's Photolog. The idea here was to develop a subjective rating of the roadway character throughout the segment. In some cases, a segment was only identified in one or two of the priority lists, while in other case a segment appeared in all priority lists. This inconsistency suggests that roadway attributes may play a larger role in PCR even on Kentucky parkways (which are generally homogeneous). Segments that repeatedly appeared on top ten lists were reviewed and representative images are presented in Appendix E. The key takeaway from these segments is the existence of roadway attributes that likely adversely affect safety. These attributes are not controlled for when considering segment length alone.

3.5. Conclusions and Discussion

Analysis indicates a clear relationship between well-established measures of goodness-of-fit and segment length. While there was not an optimal segment length that included best values across all of the metrics, some patterns clearly emerge.

Increasing segment length improves Modified R^2 while MAD values become less optimal.

In all likelihood, there is redundancy between what these metrics evaluate. These patterns are consistent across both scenarios. The most likely explanation for the decrease in MAD with decreasing length is the increase in sample size. Results also showed a clear pattern of decreasing overdispersion as segment length increased. The HSM states that as the overdispersion parameter approaches zero a model's statistical reliability increases (AASHTO, 2010). However, as *k* values declined in the models described here, the values of other metrics indicated the SPFs performed less well (as noted by MAD and MACD in Scenario 1). Overdispersion and Modified R^2 all seem to follow the same trend of improving as roadway segments lengthen.

The HSM does address the need for a length-based overdispersion for specific highway types (chapters 10 and 11) (AASHTO, 2010). Research has shown that by assuming a constant overdispersion for a set of data can lead to inconstancies in the way that safety is estimated when short and long segments are in the same dataset (Hauer, 2001). The data in this chapter suggest that increasing length leads to an improved model when considering overdispersion (*k* closer to zero). In the context of this chapter, length is varied to examine the goodness-of-fit of the models. The effect of length-based overdispersion will be considered when length is not controlled for in the next chapters.

Based on the values of each metric and evaluations of the CURE plots, the ideal segment length for Kentucky rural parkways is 2.0 miles (3.2 Km). While 2 miles is not likely to be the optimal length for all analyses, the process demonstrated here could be duplicated to identify appropriate lengths for other road categories and allow for the determination of the optimal segment length. It is possible that a different segment length could be identified for each roadway category and this could also vary from state to state. As noted above, in setting segment length, one needs to strike a balance between ability to discern changes and countermeasure implementation. As such, some

engineering judgment is needed when evaluating data similar to those shown in Table 1 and Table 4 in order to determine the optimal segment length. The removal of low traffic volumes in Scenario 2 also shows AADT's impact on SPF development. Much less PCD was observed when low volume segments were removed. This is an indication that some systematic error was removed likely due to the fact that AADT is a proxy for omitted variables contributing to heterogeneity of the segments. This is explored in the next chapter.

Additional work may be needed to further refine the segment used here. For example, removing curves and interchanges could improve the Scenario 1 model because doing so would increase the road network's homogeneity. Another option is to filter the road network to exclude small urban areas. However, this may reduce sample sizes to below the minimum thresholds the HSM recommends for SPF development. If this were to occur, curvature and urban area could be introduced as additional variables in the SPF model to address omitted variable bias. These improvements are the subject of the next chapter.

Finally, it should be underscored that the optimal segment length is sensitive to a variety of variables. For instance, in Section 3.2 the priority lists changed based on the segmentation techniques (each with different lengths). Furthermore, goodness-of-fit measures from Scenarios 1 and 2 suggest different optimal lengths. In this case, the traffic volume range was the only distinction between the two. The conclusion drawn here is that there is **no globally uniform length** that leads to the best SPFs, but rather analysis tools should be used to evaluate model development. Moreover, roadway homogeneity plays a large role in model development even for roadways designed fairly consistently like Kentucky parkways.

Chapter 4. Optimizing Attribute Specification and Aggregation

4.1. Introduction

The next step in this research was the investigation of the effect that changes in roadway attributes during the segmentation process have on highway safety. For example, is the effect on highway safety more sensitive to a change in shoulder width or to the change in number of lanes? A network can be filtered based on attributes and the effect on the resulting SPFs (in terms of goodness-of-fit and predictive power) can be observed. Guidance is provided based on the sensitivity each attribute has on the SPF's goodness of fit. This will give safety practitioners a better idea of what attributes help define homogeneity.

Despite the fact that filtering by attributes makes the roadway network more homogeneous, there are disadvantages as well. An obvious downside is that filtering reduces the sample size (segments or intersections) used to develop a model. Depending on the extent of the filter, this can reduce the network to such a small size that model development is not feasible. For example, the HSM recommends 100-200 intersections or miles for SPF development (AASHTO, 2010). It is demonstrated however that a careful assessment during the development process can help improve SPF development, even below these limits. Another trade-off is that the filtering process alters the base conditions of the SPF and therefore introduces the need for Crash Modification Factors or functions for segments that are dissimilar to the base conditions.

In contrast to the segment length analysis performed in the previous chapter, a more interactive SPF development process was required. Previously, SPFs were developed for the same network using different length categories. In this analysis, filters are applied to the network to explore the effect of attribute range specification and aggregation on SPF quality. As such, a less cumbersome SPF development process was desired.

4.2. Model Assessment

The objective of this section is to describe how the use of analytical tools can improve the SPF development process. In the previous chapter, a relatively homogeneous network was used to isolate the effect of segment length alone on SPFs. In this chapter, the effect of roadway attribute specification is analyzed, therefore, homogeneity is required only at the segment level.

The specification of attributes complicates the modeling process as it introduces the potential for omitted variable bias. This bias occurs when a regression is used to predict a dependent variable while influential independent variable(s) is (are) not included. However, analytical tools and metrics can be used to detect and minimize such biases leading to improved models.

Another complication of this methodology is the number of steps required to produce SPFs based on attributes. In the previous chapter, the same network was used to produce networks at various segment length. This process was easily accomplished in Excel and using a simple R script. In order to test the effect that attributes have on SPFs, various filters were needed and much larger databases were required (including a variety of roadway attributes for a very heterogeneous network of roads). It became apparent that the previous methodology was too time-consuming to reliably produce and compare outputs. Moreover, in some cases the datasets were too large to work in Excel. As such, there was a need to automate the SPF development process. This would enable several SPFs to be compared quickly and the effect of small changes could be examined. For instance, an agency can develop a statewide SPF for a specific set of geometric conditions that mirror the HSM's base conditions for two-lane rural roads. However, the agency may not have a preponderance of shoulders that are 6 feet wide (as recommended in the HSM for rural two-lane roads). Instead, SPFs can be developed for shoulder widths of two and three feet. The corresponding SPFs can be compared and evaluated to determine the best model for the agency. Previously, the development and

comparison of SPFs has been a lengthy and laborious task requiring the use of several software packages (Excel, R, and SQL Server).

The development of SPFs at the state level is growing in the United States (FHWA, 2016). According to the CMF Clearinghouse website's resources page, 12 states have developed their own SPFs and seven states have calibrated existing SPFs. In Kentucky, SPFs have been developed with state-specific data since 2013. The Federal Highway Administration (FHWA) has produced tools and documents to facilitate SPF development (such as The Calibrator and Safety Performance Function Development Guide: Developing Jurisdiction Specific SPFs). Helpful resources are listed at the end of Appendix F. These resources offer insights on how to evaluate SPF models. Tools like The Calibrator provide goodness-of-fit measures such as modified R2 and CURE Plots. When developing state-specific SPFs, these measures can be used to identify ways of improving SPF model development. One way to improve these models is to detect and avoid omitted variable bias (Srinivasan et al., 2013b).

4.2.1. Omitted Variable Bias

Omitted variable bias occurs when a variable that contributes to crash prediction is not included in a regression model[8]. Model development based on heterogeneous roadway geometry can be one cause of this bias. For example, a dataset may include roadways with varying geometrics such as wide lanes and shoulders in some segments and narrow lanes and no shoulders in others. Heterogeneity contributes to omitted variable bias if the variation in geometrics is not part of the model. Adding more independent variables to a model can minimize omitted variable bias; however, depending on number of varying geometrics, this can lead to overfitting (Srinivasan and Bauer, 2013a). Overfitting can result in goodness-of-fit measures that improve when adding variables, but these improved measures may result from modeling "noise" or correlation of different variables (e.g., more than one variable is modeling the same effect) (Srinivasan and Bauer, 2013a, Hauer and Bamfo, 1997). Another way to address omitted variable bias is to filter the dataset to a more homogeneous network (i.e., base conditions) provided the sample is large enough. The HSM and the Safety Performance Function Decision Guide offer sample size guidance for SPF development (AASHTO, 2010, Srinivasan et al., 2013b). For example, the SPF Decision Guide suggests 100-200 sites and 300 crashes per year for SPF development for network screening.

Recall that CURE Plots provide a visual method of detecting omitted variable bias, and, as discussed below, model form and outliers. These plots graph cumulative residuals against another variable (such as traffic volume or length) in a scatter plot. Residuals are computed by subtracting the crash prediction at a site (based on the SPF) from the number of crashes recorded for that site. Residuals are sorted by the variable being compared (often AADT) and the residuals are cumulated (the residuals from site i are added to site i+1 and so on). Residuals are positive if the model predicts fewer crashes than were recorded. Ideally, the magnitude of residuals should balance out. This

[8] In fact, a strength of the negative binomial regression is that it allows for some variation by variables not included in the model (Tegge et al, 2010).

manifests in a CURE plot by steady oscillation around the x-axis. Large jumps in the CURE plot are indicative of outliers, poor modeling, or data errors (large residuals).

Steadily increasing or decreasing residuals, however, can indicate omitted variable bias. Upper and lower limits are typically plotted along with the residuals to identify if the residuals stay within two standard deviations (Hauer and Bamfo, 1997). These confidence limits are plotted along with the residuals, and the residuals should only rarely go outside of the limits. In fact, the CURE plot should end near zero indicating that the model does not over- or under-predict crashes. Confidence limits are used to discern the difference between the expected random error and undesired systemic bias (Hauer and Bamfo, 1997). Hauer and Bamfo derived an equation for confidence bands based on the probability density of the random walk (oscillation) of the CURE plot. This drift can be demonstrated easily using a roadway network filtered in the three following ways:

- Scenario 1 – All Rural two-lane roads in Kentucky with nine-foot lanes
- Scenario 2 – Scenario 1 with no median, shoulder width of two feet, and no curvature
- Scenario 3 – Scenario 2 and traffic volume less than 500

The following tables list segment lengths (in miles) by the parameters from Scenario 2 (Table 5) and then filtered by AADT for Scenario 3 (Table 6). Notice that the column and row total of 9855.8 miles (15861.4 km) represents the total length of Scenario 1, and the underlined total of 1712.6 miles (2756.2 km) is the total for Scenario 2.

Table 5. Segment Lengths for Scenarios 1 and 2

Length	**Curve**		
Median	**No**	**Yes**	**Grand Total**
No			
Other Shoulder	4,525.4	2,315.0	6,840.4
Shoulder=2ft	1,712.6	1,301.0	3,013.6
Yes			
Other Shoulder	0.7	0.6	1.4
Shoulder=2ft	0.2	0.2	0.4
Grand Total	**6,239.0**	**3,616.8**	**9,855.8**

Table 6. Length of Segments for Scenario 3

Length	**Curve**		
Median	**No**	**Yes**	**Grand Total**
No			
Other Shoulder	2,483.2	1,028.8	3,512.0
Shoulder=2ft	935.7	624.7	1,560.4
Yes			
Other Shoulder		0.1	0.1
Grand Total	**3,418.9**	**1,653.6**	**5,072.5**

CURE Plots are used to compare the SPFs from the three Scenarios. Figure 19 shows three CURE Plots, one for each scenario.

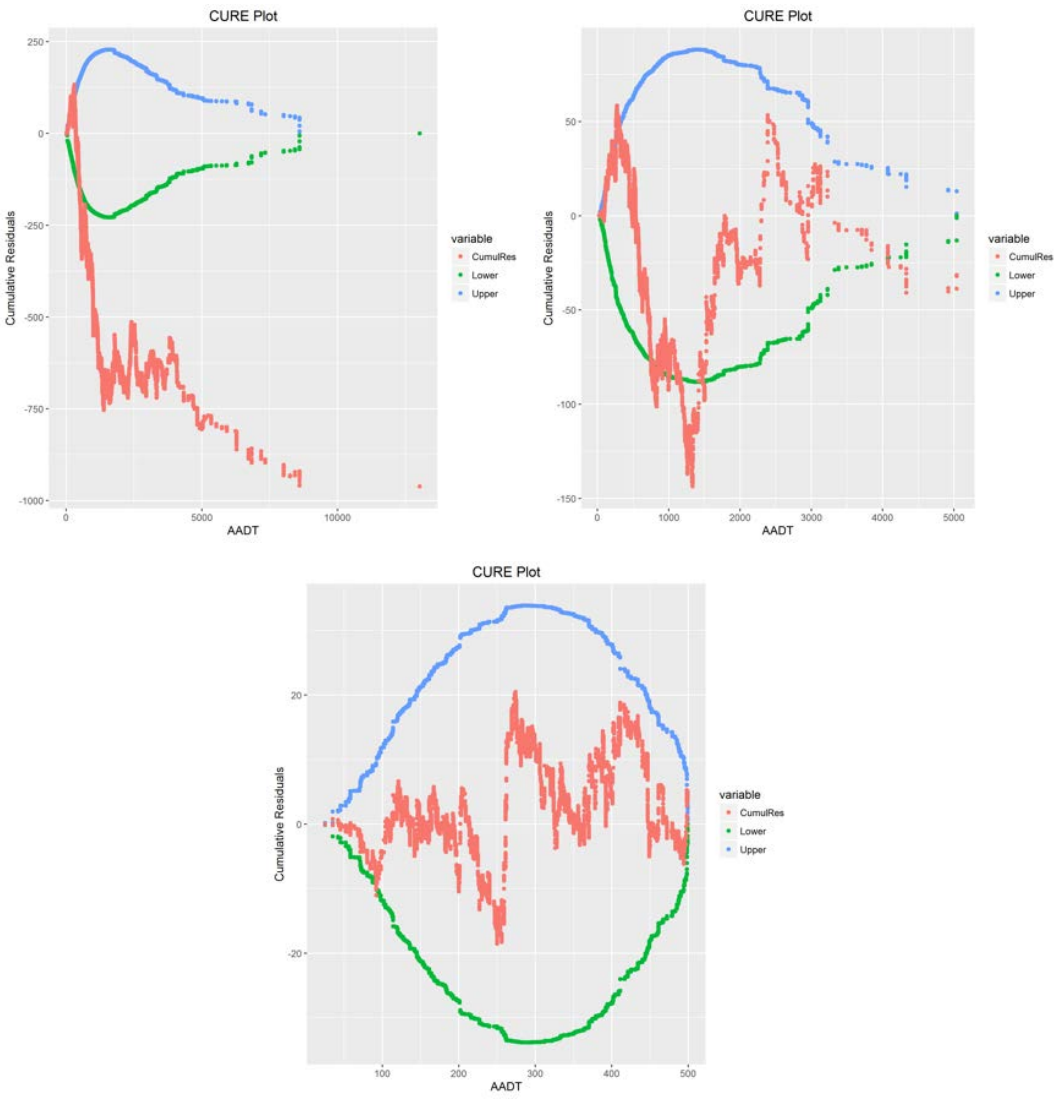


Figure 19. CURE Plots for Rural 2-Lane Roads in Kentucky for Scenarios 1, 2, and 3 (top-left to bottom).

There is a clear downward drift in the residuals in Scenario 1, which is an indication of omitted variable bias. Scenario 2 partially addresses this bias by limiting the roadway geometry, resulting in a more homogeneous network. Residuals for Scenario 2 move outside the confidence bands for AADT values between approximately 750 and 1,500. For this scenario, the residuals have a larger absolute value than what is expected due to random error. Notice, however, that the large residuals occur at higher AADT. Scenario 3 corrects for the large residual by limiting the network to sites with AADT under 500. This

final plot indicates good oscillation, with the residuals remaining within the confidence bands and approach zero at the end of the plot.

The CURE plot for Scenario 3, along with other goodness-of-fit measures, suggests that of the three scenarios it is the most desirable model. However, without the aid of the CURE plots or goodness-of-fit measures, there is little evidence to suggest that Scenarios 1 and 2 are undesirable SPFs. Figure 20 compares the scatter plots for the same three scenarios. The number of crashes at each site is plotted against the site's traffic volume.

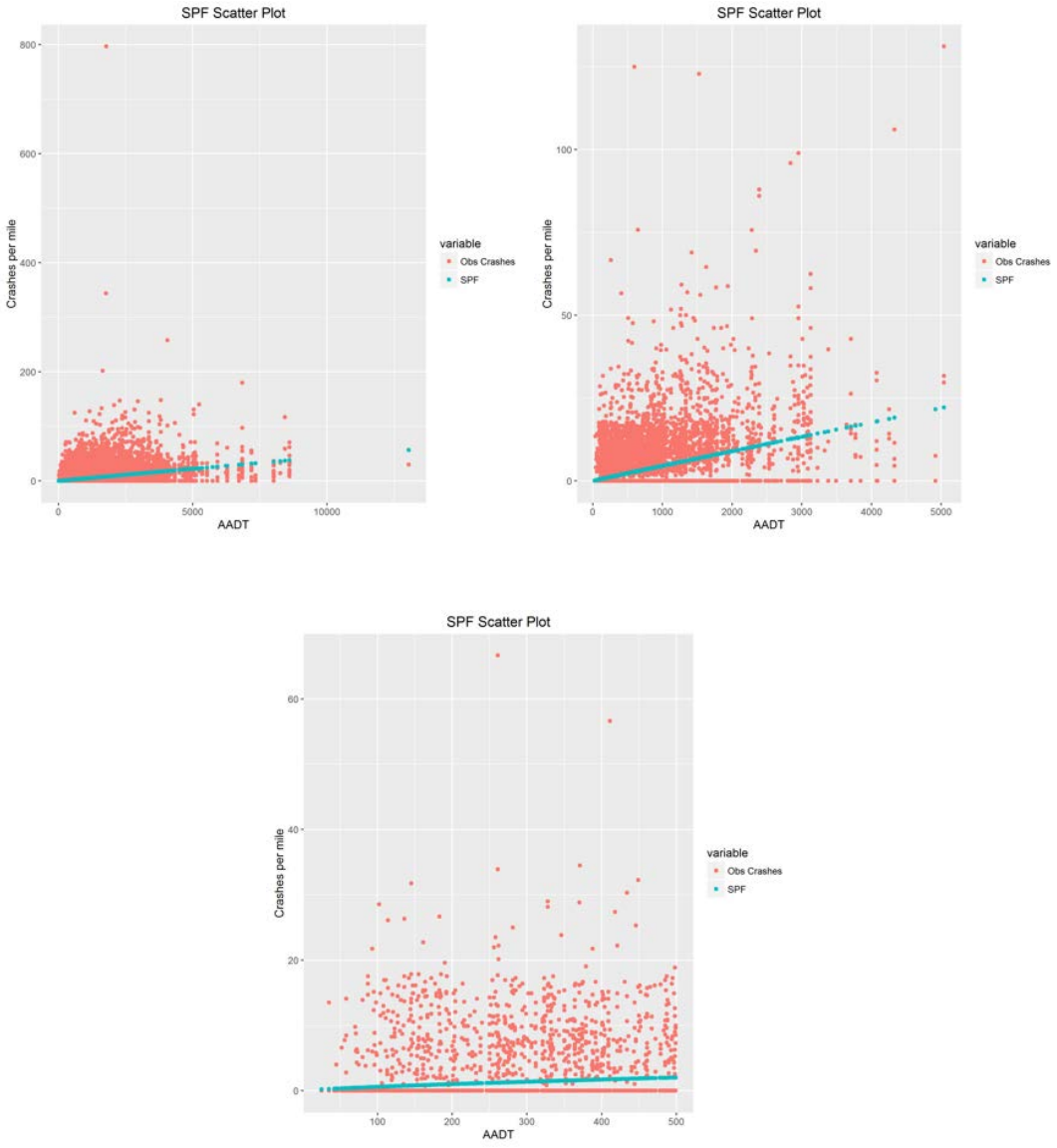


Figure 20. Scatter plots for Scenarios 1, 2, and 3 (top-left to bottom).

The scatter plots offer little insight into which scenario offers the best SPF. Table 7 shows varying regression parameters (which the next section discusses in detail) and overdispersion for each scenario.

Table 7. Regression Parameters and Overdispersion for Three Scenarios

	Scenario 1	Scenario 2	Scenario 3
Theta*	1.313776	1.556977	1.50734
Alpha	-5.23151	-5.24279	-4.01983
Beta	0.97871	0.97832	0.760655

*RStudio reports the overdispersion parameter as theta, which is the reciprocal of k.

The overdispersion parameter (*theta*[9], defined as *1/k*) is directly proportional to EB estimate as outlined in the HSM (AASHTO, 2010). A larger *theta* suggests a better SPF model when accounting for the EB estimate. The parameters listed in Table 7, if taken alone, would provide misleading evidence in favor of Scenario 2 because they do not detect omitted variable bias. Even employing a length-based overdispersion would be unlikely to change this outcome as all three scenarios have an average length of about 0.14 miles (225 m). Interpolation based on this table alone stands in contrast to results derived from comparing CURE Plots. CURE Plots along with goodness-of-fit measures are critical when comparing SPFs. For comparison, Alpha and Beta are the regression coefficients associated with each scenario (as defined in Equation 1).

While the point here is to illustrate the usefulness of CURE Plots in detecting omitted variables, this data demonstrates the well-known relationship between homogeneity and crash prediction (AASHTO, 2010). The sensitivity of this effect is tested in this chapter by filtering the network by roadway attributes (a measure of homogeneity) and comparing the SPFs. In the next chapter, both length and homogeneity will be tested simultaneously and in the same manner.

It is worth noting that, while the CURE plots described above are compared to AADT, other variables can be used. Site specific variables can be used to make

[9] Many documents, including the HSM, refer to this parameter as *k*, which is the reciprocal of *theta*. In this case, the relationship of the parameter and the model will be inverted.

improvements to an SPF by plotting them with cumulative residuals as discussed in this chapter. Ranges of AADT can be isolated and used to improve the model. Segment length can also be used in the same manner. In the previous chapter, this was unnecessary as length was, in general, held constant. In this and the next chapter, this comparison becomes more meaningful.

4.2.2. Outliers And Data Errors

CURE Plots and residuals can also be very helpful in identifying data errors in the form of outliers. Hauer (2004) has proposed that large jumps in these plots can indicate the presence of an outlier. While an outlier may be a data point an unusually high or low value, it also might be indicative of a data error. Depending on the magnitude of error, the removal of data errors can greatly improve the CURE plot and have a dramatic effect on the model parameters. Figure 21 shows examples CURE plots for rural, 2-lane roadway before and after the removal of two data errors (very large residuals).

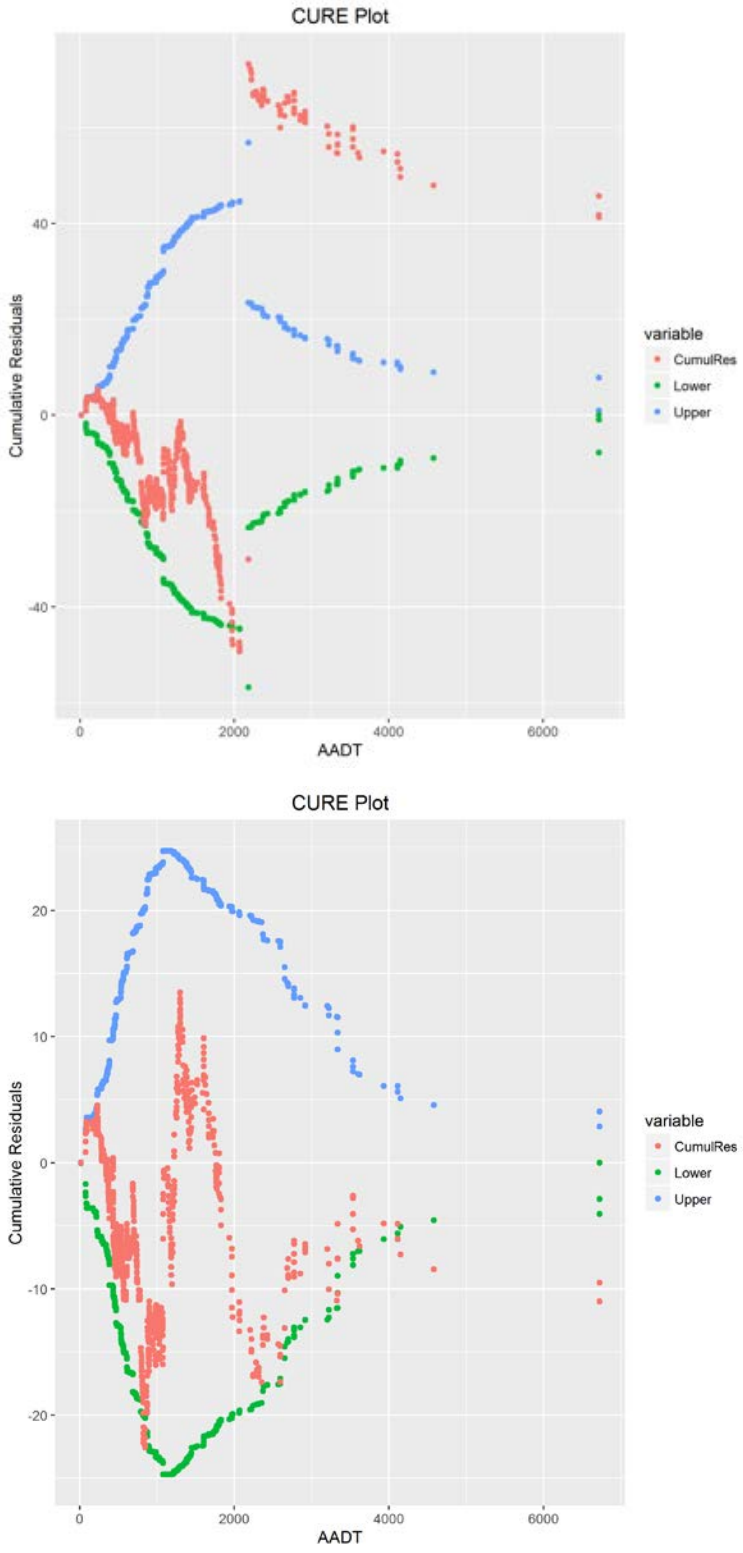


Figure 21. CURE plot before (top) and after (bottom) the removal of data errors.

The large vertical discontinuity in the plot on the top of Figure 21 is an indication of a data quality issue. The discontinuity is a result of a relatively large residual at one site (as compared to neighboring residuals). That is, the model is either largely over- or under-predicting when compared to the observed number of crashes at a site. Further inspection reveals that the abrupt shift in this example is the result of over 100 crashes incorrectly assigned to a rural, two-lane segment located in the far western portion of Daviess County, Kentucky. In this case, the incorrect assignment of crashes was due to the re-designation of routes in downtown Owensboro. A section of US 60 was re-routed to bypass the city but the base map in the crash collection tool has not been updated, resulting in the assignment of an incorrect milepoint. Removal of this segment dramatically improved the CURE Plot, as shown on the lower plot in Figure 21.

Many agencies use county, route, and milepoint (CRMP) for crash analysis as it can be more unambiguously matched to roadway information as compared to coordinate data – especially in urban areas or at intersections. This method is more useful to data users as the location is easier to communicate without the aid of a map. In Kentucky, the CRMP data is dynamically assigned when a police officer codes the location using a GIS-based map called MapIt (Green and Agent, 2011). While most location errors in Kentucky have been mitigated by use of the MapIt system, basemap errors are still possible and in many cases they are systematically detectable. An error such as the one described above may likely result in a site with a very high PCR value (the data quality error results in a large number of observed crashes while the low AADT would generate a much lower model prediction). These errors may go unnoticed until further study is conducted. CURE plots can be used to easily identify such errors before time is wasted studying locations that are ultimately not of interest.

4.2.3. Safety Performance Function Development Process

A generalized linear model using negative binomial regression is typically used to create an equation that relates predicted crashes to traffic volume and length (as well as other independent variables, if desired). As described earlier in Equation 1, a commonly used variation is:

$$SPF = L * e^{a} AADT^{b}$$

Where,

L = Length of segment in miles
AADT = traffic volume of the segment
a = regression parameter for intercept
b = regression parameter for AADT

It should be noted that the model form may be adjusted and the values of the regression parameters will change based on the highway type used in the regression. For example, the HSM recommends this model form for rural multilane and for urban and suburban arterials[10].

Statistical packages such as SPSS, Stata, SAS, and RStudio perform this regression easily with built-in tools. SPFs can also be developed in Microsoft Excel using solver or custom functions. The above-mentioned tools are simple enough to generate an SPF manually, but attempting to improve model development manually can be cumbersome. Model improvement requires several iterations and the filtering of the roadway dataset. Moreover, creating CURE Plots requires several steps and can be time-consuming, particularly for a large database. FHWA's Calibrator tool readily generates CURE Plots but is separate from the SPF development. This separation necessitates several intermediate and repetitive steps.

In an effort to aid in the assessment of several models an automated process was developed. A tool was produced that consolidates SPF development and assessment, including the generation of CURE plots, into one streamlined process. Work previously done using a combination of R, Excel, and the Calibrator tool is now accomplished with a single source code run in the program RStudio and accordingly named "SPF-R." The use of other software, such as Excel, has been reduced to organizing the input and output.

---

[10] The HSM defines equations 11-7 and 12-13 using a slightly different notation but they are mathematically equivalent to the form used in

$$y = L * e^{a} ADT^{b} \qquad (1.$$

This tool was used in this and the subsequent chapter to produce fast and consistent results so that SPFs can be compared and analyzed. A detailed user's guide is included in Appendix F with examples. The source code is available on GitHub[11]. GitHub is an online, collaborative tool that allows anyone to download the source code and contribute to its improvement. The code can be modified as needed and meaningful changes may be committed to the GitHub repository so that other safety professionals will benefit from the enhancements. The code is also presented in Appendix G.

4.3. Methodology

The Kentucky Transportation Cabinet maintains roadway planning (e.g. rural/urban, traffic volume) and geometric data (e.g. shoulder width, curvature) in individual shapefiles, each segmented based on a particular asset. Generally, segments are split when an attribute changes. For example, the Lane asset (LN) describes segments in Kentucky with the same number of lanes and the same lane width. A new segment is created when either the number of lanes or lane width changes. All of the attributes are tied to Kentucky's roadway centerline shapefile using a unique route identifier (RT_Unique) and the starting and ending milepoints (typically stored to the nearest thousandth of a mile).

These attributes can be combined using a GIS tool called Route Overlay. The overlay process creates a new segmentation that splits at every breakpoint from all combined attributes. Consider a section of roadway from mile marker zero to mile marker two where the number of lanes changes from two to four lanes at mile marker one. Further, consider that the route changes from rural to urban at mile marker 1.5. The resulting segmentation would create three segments:

- From 0 to 1 – Rural, 2-lane
- From 1 to 1.5 – Rural, 4-lane
- From 1.5 to 2 – Urban, 4-lane

[11] http://github.com/irkgreen/SPF-R

Recall this was described earlier in Figure 1.

Route Overlay was used to combine nine attributes that are most likely related to highway safety performance:

- Horizontal curvature – the direction and degree of curvature (CU).
- Functional classification – the functional classification of the road including whether it is rural or urban (FS).
- Vertical curvature – the direction and percent grade (GR). This asset has limited coverage.
- Lanes – the number of lanes and lane width (LN).
- Median – the presence, type, and width of a median (MD).
- Shoulder – the presence, type, and width of a shoulder (SH). These attributes are reported in both directions however only the cardinal direction was used for simplicity. There is rarely a difference between the cardinal and non-cardinal shoulder.
- Traffic flow – the traffic volume of a segment (TF).
- Speed limit – the posted speed limit for the segment (SL).
- Intersection – Kentucky maintains an intersection database that was used to flag segments that were near intersections (Green et al., 2016).

The resulting segmentation was then linked to Kentucky's crash database. The number of crashes was obtained for each segment. A crash was included if the crash occurred between the start and end mile points of the segment. If a crash occurred exactly at one of the segment's edges, then the crash was assigned to the segment with the lower endpoint.

Each segment included the number of crashes, traffic volume, and the segment length – as well as all of the other attributes resulting from the Route Overlay. Negative binomial regression was used to develop SPFs that relate crashes to the length and traffic volume using the model form described in Equation 1. It is unrealistic to develop an SPF for all of the segments in the database as the segments change from rural to

urban, divided to undivided, and vary in number of lanes – in addition to other changes in roadway homogeneity. Therefore, the following analysis was limited to rural, 2-lane roadways in an effort to explore the effect of roadway attributes on the quality of SPFs. The RStudio code, SPF-R, described earlier, was used to develop a variety of SPFs. Two approaches were used to examine the effect of attributes on SPF development: 1) filtering the database based on attributes, and 2) adding additional variables to the SPF model. CURE plots served as the primary means for SPF assessment. A more complete assessment included comparing other goodness-of-fit measures.

4.3.1. Database Filters

The first approach used filters to exclude segments from the SPF development process. A base filter was applied that limited the database to rural, 2-lane segments. This filter was used for the remainder of the analysis in addition to other filters. The following attributes were used in the filtering process:

- Horizontal curvature
- Vertical curvature
- Presence of a median
- Presence of an intersections
- Segments with known data errors[12]
- Lane width
- Shoulder width
- AADT ranges
- Speed limit

For the first five filters listed above, segments were excluded based on the presence of a curve, median, intersection, or known data errors. Very flat curves were not excluded, as this would adversely affect the sample size of segments. The last four filters were used to include a specific number or a range of values. For instance, lane

[12] Known data errors were included in the dataset to both illustrate and to test the effect on the SPF development process.

width could be limited to 9 feet or shoulder widths could be limited to between 1 and 3 feet. Table 8 was used to help guide the selection of filters for lane and shoulder widths.

Table 8. Total Length (miles) of Rural, 2-Lane Roads by Shoulder and Lane Width in Kentucky[13]

Lane Width (feet)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
6																
7	0	31	68	17	4						0					121
8	0	26	362	261	26	4	3									681
9	2	69	1687	1454	197	35	7	0	2		0					3452
10	5	140	3089	4581	1597	303	118	3	13	1	6	0	0			9857
11	21	42	1638	2251	993	196	173	11	87	1	30	1	0	1		5444
12	18	11	216	295	440	99	110	3	93	13	134	22	4	2	1	1461
13	24	4	83	59	129	41	114	7	150	45	718	160	137	0	14	1684
14	6	1	8	1	5	1	7		2		5		3			38
15	5	2	1	1	0	0	0		6	0	2		1			18
16	4	1	1	1	1		0		1		1		0			9
17	1	0	2	1			0	0	0		2					6
18	2	1	1	0	0		0									4
19	1	0	7	0	0											9
20	0		0								0					1
21	1		2	0			0				0					2
Total											0					0
	89	328	7166	8921	3391	680	533	25	352	61	899	183	145	3	14	22790

[13] Length of zero indicates that the segments are under 0.05 miles (0.08 km) likely a result of the Route Overlay process.

The HSM recommends 100 to 200 miles of segments for SPF development (AASHTO, 2010). As such, many of the shoulder-lane combinations are likely unsuitable for SPF development. That is not to say that a filter producing less than 100 miles would result in a poor model, but rather other filters are likely to produce better models. It is clear, however, that many of the combinations are unrealistic for model development (e.g. combinations resulting in under 100 miles). Table 8 does provide an indication that lane widths between 8 and 10 feet and shoulder widths between 2 and 4 feet are among the most prominent configurations on rural, 2-lane roads in Kentucky.

A filtering process was used starting with the base filter of rural, 2-lane roads and progressing through a variety of attribute filter combinations. The automation tool was used to evaluate the 9 attributes (listed above) and CURE plots were compared in an effort to identify which attributes had the strongest effect on model improvement. Keep in mind that CURE plots were used as a screening tool and other goodness-of-fit measures are still considered. It was unnecessary to compare every conceivable combination of the 9 attributes. Some attributes had little effect while other attributes showed an effect when in combination with other attributes. This process resulted in 18 database filters made up of various attribute filter combinations. These filters were compared with respect to the goodness-of-fit measures and CURE plots.

4.3.2. Additional Model Parameters

In another comparison, additional variables were added to the model. The addition of model parameters increases the sample size of sites. That is, instead of filtering the database to only include segments with a lane width of 9 feet, lane width could be added to SPF as a variable. As argued earlier, this makes the network more heterogeneous as some segments have narrow lanes and others have wide lanes. This heterogeneity can be accounted for by adding width to the model.

Models were developed including lane and shoulder widths as parameters under a variety of configurations. The models were compared to a base model from the previous section as a basis for improvement. The following models were developed.

- *Base Model*
- $L * e^{a} AADT^{b}$
- *Model 1 – including shoulder and lane width variables*
- $L * e^{a} AADT^{b1} e^{SW*b2} e^{LW*b3}$
- *Model 2 – including shoulder width as a variable and filtering lane width to 9 feet*
- $L * e^{a} AADT^{b1} e^{SW*b2}$
- *Model 3 – including lane width as a variable and filtering shoulder width to 3 feet*
- $L * e^{a} AADT^{b1} e^{LW*b2}$
- *Model 4 – including roadway width as a variable (shoulder plus lane width)*[14]
- $L * e^{a} AADT^{b1} e^{(SW+LW)*b2}$
- *Model 5 – including shoulder and lane widths as variables and adding an interaction term*
- $L * e^{a} AADT^{b1} e^{SW*b2} e^{LW*b3} e^{SW*LW*b4}$

Model 5 includes an interaction term that describes any dependence shoulder width might have on lane width. In SPF development, interaction among independent variables can be difficult to detect unless there is an intuition for the interaction (Srinivasan and Bauer, 2013). The code used for the models above is shown in Table F-1 in Appendix F.

All five models were used to develop SPFs using the rural, 2-lane database with the following filters: no Vertical Curves, no Horizontal Curves, no intersections, speed limit of 50 miles per hour or more, no median, and no known data errors (in addition to any filters defined above). This filter provided a homogeneous network to help isolate model form improvements. Homogeneity lessens the potential for unexpected omitted variable bias. It should be noted that not all independent variables were tested as it was not feasible. Lane and shoulder widths were included based on their influence as discussed below.

[14] While this may seem to duplicate the regression in model 1, there are two key distinctions: model 1 is able to independently adjust the weighting of shoulder and lane widths, and model 4 makes the assumption that a 9 foot lane with a 3 foot shoulder is comparable to a 10 foot lane with a 2 foot shoulder.

4.4. Results

The data used in the following sections was comprised of a database that contained over 407,600 roadway segments. Each segment included roadway attribute data, traffic volumes, length, and the associated crash data. The segments total to approximately 22,790 miles (36,677 km) of rural, 2-lane roadways. This analysis was limited to rural, 2-lane roadways to simplify the variety of attributes as urban roadways have a more complex array of geometrics.

4.4.1. Database Filters

The SPF development tool, SPF-R, was used to perform the comparisons in this section. A great advantage to this approach is the efficiency in which attributes can be changed and the results compared. As a baseline, an SPF was developed for all rural, 2-lane roads with no other filters applied. As expected, the CURE plot demonstrated significant omitted variable bias (shown below in Figure 22).

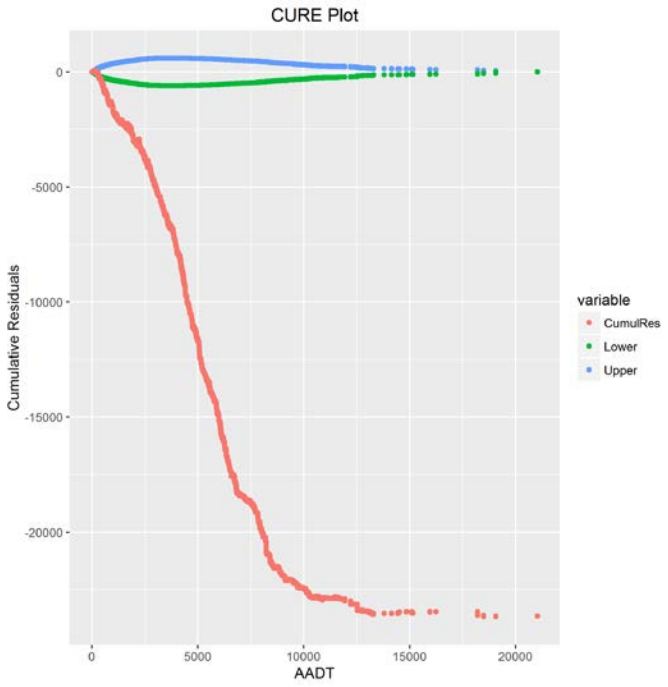


Figure 22. CURE Plot for All Rural, 2-Lane Roads in Kentucky (no other filters)

As filters were introduced, the bias reduced. Initially, the following filters were individually applied:

Filter 1. No filter

Filter 2. No horizontal curvature

Filter 3. No vertical curvature

Filter 4. Lane width of 9 feet

Filter 5. Shoulder width of 3 feet

Based on the CURE plots, none of these filters significantly addressed omitted variable bias, however lane width (Filter 4) exhibited the lowest magnitude of drift followed closely by shoulder width (Filter 5). This is expected as lane and shoulder widths are likely proxies for other variables that explain homogeneity (such as land use or topography). Median and speed limit filters were observed to have no meaningful effect. This is likely due to the few number of records excluded by these filters.

Next, combinations of filters were applied. These filters are listed below (following the same numbering scheme).

Filter 6. Shoulder Width=2, Lane Width=9

Filter 7. No Vertical Curves, no Horizontal Curves

Filter 8. Shoulder Width=2, Lane Width=9, no Vertical Curves, no Horizontal Curve

Filter 9. Shoulder Width=2, Lane Width=9, no Vertical Curves, no Horizontal Curve, no known data errors

Filter 10. Shoulder Width=2, Lane Width=9, no Vertical Curves, no Horizontal Curve, no known data errors, no intersections

CURE plots for all ten filters are shown in Appendix H. The goodness-of-fit metrics are shown in Table 9 (shading is used as before to indicate preference).

Table 9. SPF Metrics for all Filters

	1	2	3	4	5	6	7	8	9	10
Segments	407608	283707	39778	163675	149717	71612	31760	4112	4106	2911
k	1.5	1.3	1.6	1.2	1.1	1.1	1.5	2.0	0.9	0.6
Total Crashes	111002	88776	16916	31742	35929	13702	14695	1057	866	721
Modified R2	-0.08	-0.04	-0.08	0.27	0.28	0.23	-0.06	0.09	0.39	0.48
PCD	76.4%	62.5%	75.4%	80.6%	55.4%	51.0%	60.3%	23.9%	17.7%	6.9%
MACD	23687.1	19146.9	4688.8	2318.1	3242.2	896.2	4090.2	128.8	45.5	37.6
MAD	0.4	0.4	0.6	0.3	0.3	0.3	0.6	0.4	0.3	0.3

Filter 10 has a clear improvement among all metrics, which is consistent with comparing the CURE plots. This is logical as this filter produced the most homogeneous network.

In addition to the filters above, ranges of attributes were also considered. The idea here is that small changes in an attribute such as lane or shoulder widths might not impact safety significantly differently, yet including ranges would increase the sample size used to develop the SPF. Attributes that are binary (i.e. cannot be used in ranges) that seemed to improve the model based on the previous filtering process were used as a starting point (no vertical curves, no horizontal curves, no intersections, and no known data errors). Using Table 8 as a guide, ranges of widths were modeled as described below.

Filter 11. Lane Width=9, Shoulder Width 2-3

Filter 12. Lane Width=9, Shoulder Width 2-4

Filter 13. Lane Width 9-10, Shoulder Width=3

Filter 14. Lane Width 8-10, Shoulder Width=3

Filter 15. Lane Width 9-10, Shoulder Width 2-3

Filter 16. Traffic volume below 500

Filter 17. Traffic volume below 2000

Filter 18. Traffic volume below 2500

Once again, the CURE plots were used to identify ranges of traffic volumes where the model performed best. The resulting CURE plots from all ranged filters are shown in Appendix H. None of the CURE plots suggested an improvement from Filter 10 using attribute ranges, however, a 500 AADT filter did show improvement. These results are consistent with the HSM's based condition methodology where single values, not ranges, are listed for most attributes. The HSM has a worksheet for rural, 2-lane SPFs with the following base conditions.

Input Data	Base Conditions
Length of segment, L (mi)	—
AADT (veh/day)	—
Lane width (ft)	12
Shoulder width (ft)	6
Shoulder type	paved
Length of horizontal curve (mi)	0
Radius of curvature (ft)	0
Spiral transition curve (present/not present)	not present
Superelevation variance (ft/ft)	<0.01
Grade (%)	0
Driveway density (driveways/mi)	5
Centerline rumble strips (present/not present)	not present
Passing lanes (present/not present)	not present
Two-way left-turn lane (present/not present)	not present
Roadside hazard rating (1–7 scale)	3
Segment lighting (present/not present)	not present
Auto speed enforcement (present/not present)	not present
Calibration factor, C_r	1.0

Figure 23. Worksheet 10A from the *Highway Safety Manual* For Rural 2-Lane Roads

It should be noted that these base conditions are not universally ideal for all agencies. For instance, in Kentucky, there are only about 114 miles (183 km) for 12 foot lanes/6 foot shoulders on the rural, 2-lane system. The CURE plot from such a low sample size is shown below (left) along with the CURE plot including filters for curvature and intersections as well (right).

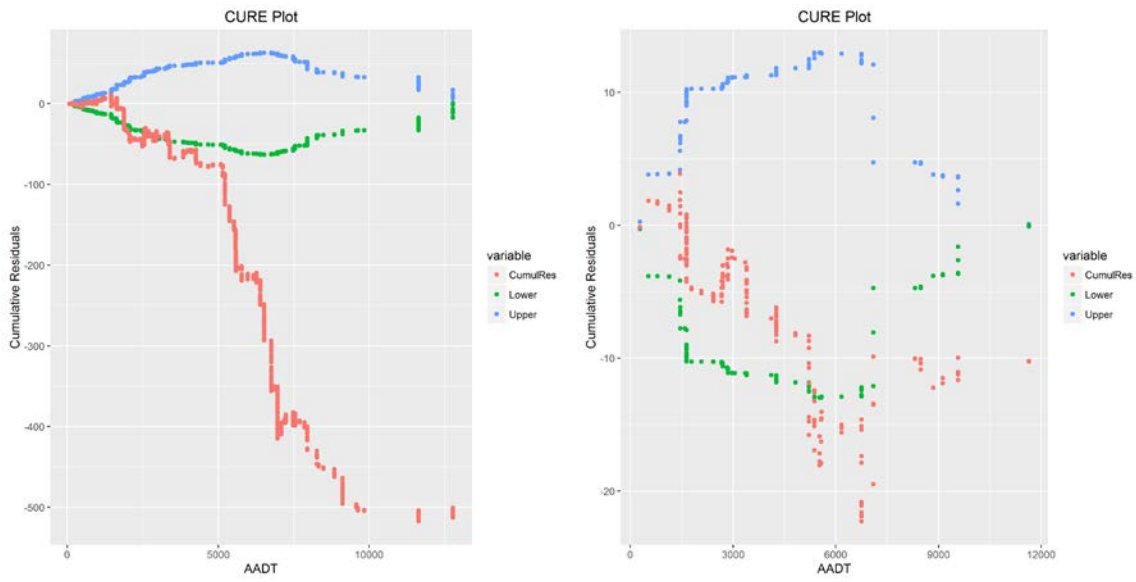


Figure 24. CURE Plots for 12 foot Lanes and 6 Foot Shoulders (Left) and Including Other Filters (Right)

Neither model would be adequate for SPF development. Moreover, alternative lane and shoulder widths such as those used in Filter 10 produce a much better model.

Other attributes shown in the HSM's base conditions are dependent on data availability. Kentucky does not maintain some of the base conditions suggested. It is suspected that if some of these attributes were very influential there would be more omitted variable bias observed. Additionally, it is possible that some other variables are performing as a proxy for some of the missing variables. For instance, the low volume filter is likely also filtering out sections with two-way left turn lanes. Table 10 below compares all eight filters plus Filter 10 using the same shading scheme as before.

Table 10. SPF Metrics for all Ranged Filters

	10	11	12	13	14	15	16	17	18
Segments	2911	4551	5347	9472	6055	5775	809	2541	2635
k	0.6	0.7	0.7	0.6	0.6	0.6	0.9	0.8	0.8
Total Crashes	721	1156	1513	3659	2485	2461	68	520	567
Modified R2	0.48	0.43	0.46	0.47	0.47	0.46	0.42	0.28	0.30
PCD	6.9%	7.4%	9.9%	14.4%	15.4%	16.9%	2.5%	10.6%	9.4%
MACD	37.6	48.7	56.2	160.7	120.1	121.5	8.3	37.6	36.6
MAD	0.3	0.3	0.4	0.4	0.5	0.5	0.1	0.3	0.3

The metrics shown above are consistent with the CURE plots in that no improvement is observed when using ranged values. The metrics for Filter 10 are comparable or better than those listed above (shown for comparison). Filter 16 shows the lowest percent CURE deviation, however, this model was limited to very low volume roads resulting in a small sample size and limited applicability. It is suspected that the low volume is a proxy for other variables as discussed earlier.

Maps showing the spatial distribution of rural, 2-lane roads by shoulder and lane widths are shown in Appendix I and Appendix K, respectively. The map of shoulder width shows a clear pattern of the topography in Kentucky. Eastern Kentucky segments have less right-of-way than the rest of Kentucky. This pattern suggests that a geographical region filter might improve the modeling process. In fact, it might be more meaningful to calibrate SPFs based on region, but this was beyond the scope of this research.

4.4.2. Additional Model Variables

This analysis compared the effect of adding attributes as variables to the model. The intention is to increase the sample of the network yet avoid omitted variable bias. Changes in an attribute can be modeled against crashes by including the attribute in the model. This is in contrast to filtering the network by that attribute. The network was filtered similar to Filter 10 (no vertical curves, no horizontal curves, no intersections, speed limit of 50 or more, no median, and no known data errors) from the previous section. The following models were compared:

- Base model (used for comparison) – 9 foot lanes and 3 foot shoulders filter
- Model 1 (shoulder and lane widths as model variables) – no additional filter
- Model 2 (shoulder width as model variable) – 9 foot lanes filter
- Model 3 (lane width as model variable) – 3 foot shoulder filter
- Model 4 (roadway width as model variable) – no additional filter
- Model 5 (shoulder and lane widths as model variables, with interaction term) – no additional filter

The metrics for each model were compared using the same convention where darker cells are less optimal. The comparison is shown in Table 11.

Table 11. SPF Metrics Compared for Various Models

	Base	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Sample	708	4829	1396	1573	4829	4829
Length	127.1	970.2	255.2	294.8	970.2	970.2
Crashes	405	5287	937	1514	5287	5287
R2	0.34	0.39	0.39	0.50	0.39	0.39
PCD	3.0	9.1	17.1	2.2	9.1	7.9
MACD	23.92	203.47	36.16	45.46	200.95	205.67
MAD	0.66	0.96	0.71	0.86	0.96	0.96
Theta	1.48	1.88	1.73	1.92	1.88	1.88
Alpha	-4.08	-4.09	-4.59	-4.09	-4.00	-4.47
Beta	0.76	0.88	0.84	0.86	0.89	0.88
*SW**		-0.08	0.01			0.01
*LW**		-0.06		-0.06		-0.02
*RW**					-0.07	
*SWxLW**						-0.01
AIC	1380.90	12503.66	2889.04	3777.85	12501.93	12504.71

*These values represent the coefficients of a specific variable

Four of the models show no improvement in any of the metrics (models 1, 2, 4, and 5). Model 3, however, shows improvements in modified R^2 and CURE Deviation Percentage yet worse MACD, MAD, and AIC. MACD and MAD measure the maximum and average deviation for the residuals. It is expected that all models would have higher deviation when compared to the base model as it is the most homogeneous network (9 foot lanes and 3 foot shoulders). This heterogeneity could lead to omitted variable bias

and outliers (as shown earlier). Furthermore, AIC penalizes models with the addition of variables to discourage overfitting. All of the models have one or more variables as compared to the base. This also explains the magnitude of AIC for models 1, 4, and 5 (each of these had more than one additional variable). CURE plots for all models are shown in Appendix L and are consistent with the PCD. As before, it is worthwhile to consider the geographical distribution of the attributes, therefore a map of roadway width (used in Model 4) is shown in Appendix M.

4.5. Conclusions and Discussion

As demonstrated, it is important to examine SPF models during their development. CURE plots can be an essential analytic tool in detecting outliers, omitted variable bias, and they indicate over what range the SPF performs well (e.g. cumulative residuals vs. AADT). These plots along with other goodness-of-fit measures can be used to improve the predictive power of the SPFs.

An advantage of the SPF automation tool developed as part of this research is that it enables nearly instant feedback when changes are made to the underlying network. This efficiency can lead to better SPFs, which will likely produce better crash predictions. Not only can SPFs be developed more quickly, which will reduce the cost, but they can be generated more easily, which can diversify the SPFs available to practitioners. This can lead to a better understanding of interaction terms, which can be difficult to identify. As the process of creating SPFs continues to improve, so too will safety professionals' ability to predict crashes and better identify more appropriate safety treatments.

Database filters were easy to apply using the automation tool. It was initially obvious that filtering by a single attribute did little to improve the models (all attributes were filtered individually but only the 4 that had a meaningful effect shown in Appendix H). Rather, filtering by a combination of attributes was required to eliminate most of the omitted variable bias. This corroborates the HSM's base condition approach. Moreover, ranges of attributes (such as lane widths from 8 to 10 feet), while increasing the sample

size, produced poorer models. This observation supports the lack of ranges for many of the base conditions in the HSM. Unlike in the previous chapter, ranges of traffic volumes were not found to improve the models on rural, 2-lane roadways. This may be caused by traffic volume serving as a proxy for some other variable not modeled for rural parkways.

The automation tool greatly aided in comparing the addition of model variables as well. Variables were added to a model from the previous section and the results were compared. The addition of variables created poorer models in four of the five models tested. Model 3 showed improvements in many of the metrics including a better CURE plot. This model filtered for 3 foot shoulders and added lane width as a variable. It is possible that some of this improvement could be attributed to 3 foot shoulders serving as a proxy for a geographic region (see Appendix J). There is an apparent clustering of 3 foot shoulders in eastern Kentucky and that coupled with the narrower lane widths (see Appendix K) in eastern Kentucky might help model a regional driver behavior or environmental effect.

Consideration should also be given to the magnitude and the range of variables used. Many of the variables estimated for shoulder and lane widths shown in Table 11 are near-zero. When coupled with a small value (such as a shoulder of 0 or 1 feet) this effectively has no impact on the model. Recall that the term is in the form: e^{b*SW} and the resultant nearly equates to one. That is, a shoulder width of 1 would be modeled to have no impact on crashes. In the other extreme, a shoulder width of 6 feet would produce a 30% reduction in crashes based on Model 1 ($e^{-0.06*6} = 0.7$). For pavement width, this impact translated to a 70% reduction for a 17 foot roadway width, using Model 4 ($e^{-0.07*17} = 0.3$). It is important to consider the length of segments by shoulder and lane width combinations that were used to create these models (recall Table 8). The small samples resulting from some of the combinations are likely contributing to the poor improvement in modeling. Finally, as noted, interaction can be difficult to anticipate. Model 5 suggests that there is little interaction observed with lane and shoulder widths as there was no model improvement, however, that is not to say

that there is no interaction among variables. The automation tool can provide an efficient way to test for interaction. The CMF Clearinghouse could also be used to help guide the selection of variables modeled as well as likely interaction terms. The magnitude of CMFs or CMFunctions can potentially help identify the most influential variables for a given facility type or crash type.

Chapter 5. Optimizing Segment Length and Roadway Attribute Specification and Aggregation

5.1. Introduction

In this chapter, both attributes and segment length were considered in SPF development. When observed on their own, length and attributes have been shown to impact the development of SPFs. It is reasonable to assume that when considered together there is likely to be some interaction.

5.2. Methodology

Observations from the previous two chapters were used to guide the evaluation of relationship between length and attributes. In this analysis, length filters were applied similarly to the way attribute filters were applied in the previous chapter. Length categories were used in conjunction with attribute filters to test the impact on SPF development. Various model forms were also tested with respect to how length is modeled and the resulting SPFs were compared. Lastly, length-based overdispersion was tested in the context in SPF development.

5.2.1. Length Filter

As discussed earlier, the database used to create segments in the previous chapter created a break whenever one of the attributes changes. As pointed out, this can create very small segments, and due to rounding, some resulting segments can be small. A length filter was applied to remove very small segments and to set a minimum length for SPF development. The following length filters were applied using the filters from Filter 10:

- No length filter (for comparison)
- Length > 0.001 miles
- Length > 0.01 miles
- Length > 0.1 miles

CURE plots and other metrics were compared for each model. CURE plots can also be used to test with which segment lengths the model performs best (similar to traffic volume). Modifications were made to SPF-R to create a CURE plot versus length (see Appendix F for details on this modification).

5.2.2. Length Categories

Analysis was performed on the rural, 2-lane dataset combining the methodologies from the two previous chapters. The rural, 2-lane dataset was re-segmented to create various length categories (fixed length) between 0.1 miles (160 m) and 1 mile (1610 m). These categories were compared to a version of the network where the length was based on changes in attributes (variable length). This is referred to as the "Raw" category as the network was segmented in its original form.

Several goodness-of-fit measures and various plots were used to compare combinations of the length categories and attribute filters. An output structure was defined to include length category, attributes (filter definition), CURE plot, scatter plot, descriptive statistics, SPF metrics, box plots, and a map. These outputs were produced for each model.

Previously, segments were discarded if they were less than the desired length category. That is, if the target length category was 0.7 miles (1.1 km), then any segment less than 0.7 was discarded (recall from 0, scenario 1). For this analysis, these segments were flagged as "remainders." The idea behind this approach was twofold. First, including these segment remainders increased the sample size due to the inclusion of previously omitted segments. Second, including remainders reduced average segment length. For example, a 1.5-mile segment in 0.2 segmentation length would result in only seven segments with the last 0.1 miles (remainder) dropped from consideration. Model comparisons were made with and without remainders in the network.

The following filters were applied to further restrict the database of rural, 2-lane roads with 9 foot lanes and 3 foot shoulders for each length category.

- Length filter
- AADT filter
- Horizontal curve filter
- Known data errors filters
- Speed limit filter
- Functional classification filter

The length filter was used to exclude "short" remainders from the analysis. The other filters were applied as before. Minor collectors are the most predominate functional classification in this dataset, however, in Kentucky, functional classification is generally not found to represent homogeneity.

5.2.3. Comparing Model Forms

Another way to improve prediction models is to alter the model's functional form. Hauer et al. (2002) implement a functional form as described in Equation 1. This is referred to as Model A. The HSM implements a similar form with a key distinction. Equation 10-6 in the HSM describes an SPF for rural, 2-lane as follows:

$$Y = AADT * L * 365 * 10^{-6} * e^{a} \textit{ (Model B)} \qquad (4)$$

It should be noted that traffic volume (AADT) is treated as an offset, similar to length in that there is no exponential term. This is referred to as Model B. Equation 1 can be rewritten similarly to Equation 4 for comparison:

$$Y = AADT^{b} * L * e^{a} \text{ (Model A)}$$

Notice that the two forms are similar with the exception of the exponential term for traffic volume (unless *b* = 1). Model B also includes a term commonly used in crash rates to normalize crashes per 100 million vehicle-miles traveled. This is term is unnecessary as the magnitude of *a* can reflect the same conversion during regression. It should also be pointed out that Equation 4 assumes a linear relationship between traffic volume and crashes – if the volume doubles, the crash prediction doubles. As pointed out in section 3.1, this is often not the case. Incidentally, the latest version of the

Interactive Highway Safety Design Model (IHSDM), which is a companion to the HSM, lacks the option to add a parameter to traffic volume, which forces it to follow Model B's form. Both functional forms were used on the same rural, 2-lane dataset with filters similar to Filter 10 and the models were compared.

Additionally, one other functional form was considered that adjusted how length is modeled. In both Models A and B, length is treated as an offset. It is likely that upcoming versions of the HSM will include a model form that treats length similarly to traffic volume in that it is not necessarily linearly related to the crash prediction. An exponential term can be added to Model A to produce a new form (Model C, Equation 5).

$$Y = AADT^{b} * L^{c} * e^{a} \text{ (Model C)} \qquad (5)$$

Similar to traffic volume in the previous models, length has a non-linear relationship with crashes in this model. It is possible that length can be a proxy for some other aspects of safety not accounted for in the model; therefore, an advantage to this model form is that the magnitude of the parameter *c* might account for the missing variables (e.g. driveway density is likely to be higher on longer segments). Examples of how to implement these model forms in SPF-R are shown in Appendix F.

5.2.4. Length-Based Overdispersion

The HSM suggests a length-based overdispersion for specific models. The distinction here is that overdispersion is estimated as a function of length. The motivation for this distinction is that overdispersion has been observed to be higher in shorter segments than in longer ones (Hauer, 2001). Cafiso et al. (2010) use Equation 6 for overdispersion:

$$k = A * L^{B} \qquad (6)$$

Where,
k=variable overdispersion
A and B are constants estimated during negative binomial regression
L=Length in miles

Notice that overdispersion is a function of length and that the sign of parameter *B* will dictate the relationship (positive or negative). Parameters *A* and *B* are estimated during the negative binomial regression. Once again, modifications were made to SPF-R to add this functionality (see Appendix F). This methodology produces a variable overdispersion that can be calculated for each segment. Recall that this methodology was unnecessary in 0 as length was constant. Recall that Filter 10 is based on the database filter from Chapter 4. This was used with the functional form from Model C to compare models with and without a variable dispersion.

5.3. Results

The following sections discuss the results from each analysis.

5.3.1. Length Filter

Three length filters were applied to a base model (Filter 10) and the models were compared. Similar to previous comparisons, goodness-of-fit measures were compared. These metrics are shown in Table 12.

Table 12. SPF Metrics for all Length Filters

	Base Model	Length > 0.001	Length > 0.01	Length > 0.1
Sample	2911	2898	2596	718
Length	225.782	225.769	224.00	129.87
Crashes	721	721	716	407
R2	0.48	0.48	0.46	0.35
PCD	6.87	6.83	6.78	2.65
MACD	37.65	37.64	36.48	22.56
MAD	0.32	0.33	0.36	0.65
Theta	1.56	1.56	1.56	1.48
Alpha	-4.81	-4.81	-4.82	-4.22
Beta	0.86	0.86	0.87	0.78

The two smaller filters had little effect on improving the model. The metrics were unchanged or worsened. The last filter (0.1 miles), however, while reducing the sample size, improved PCD and MACD. MAD did worsen but this is expected as the average deviation is likely to increase when removing smaller segments (crashes are directly proportional to length). CURE plots for the base condition are compared to the 0.1 miles length filter below.

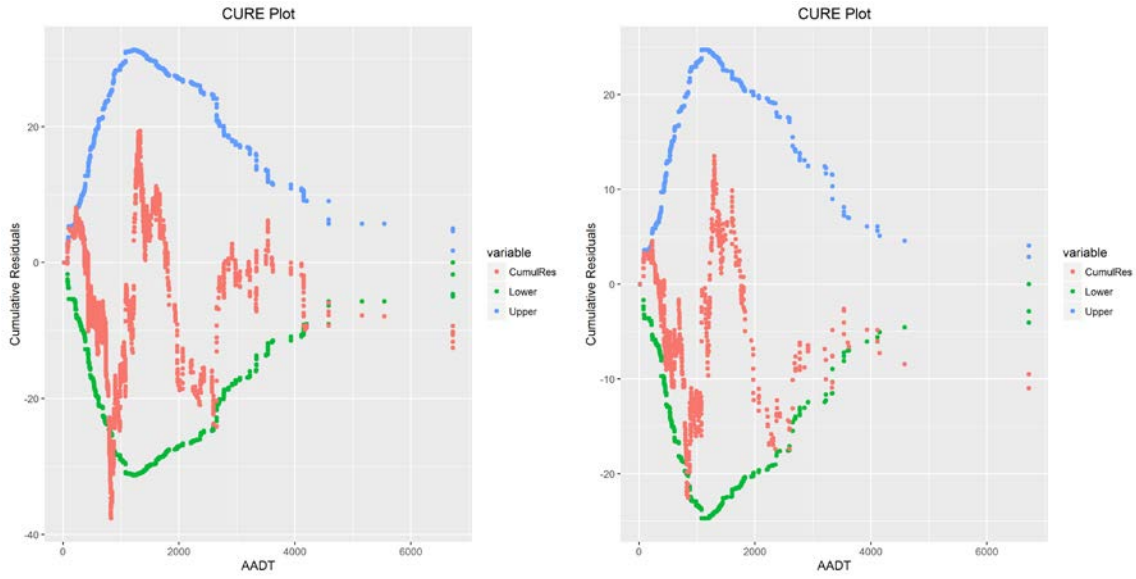


Figure 25. CURE Plots for Filter 10 (left) and with a Filter of Length > 0.1 Miles (right)

A CURE plot was created for the base condition based on length instead of traffic volume to observe if length contributed to deviation in some ranges. This plot is shown below.

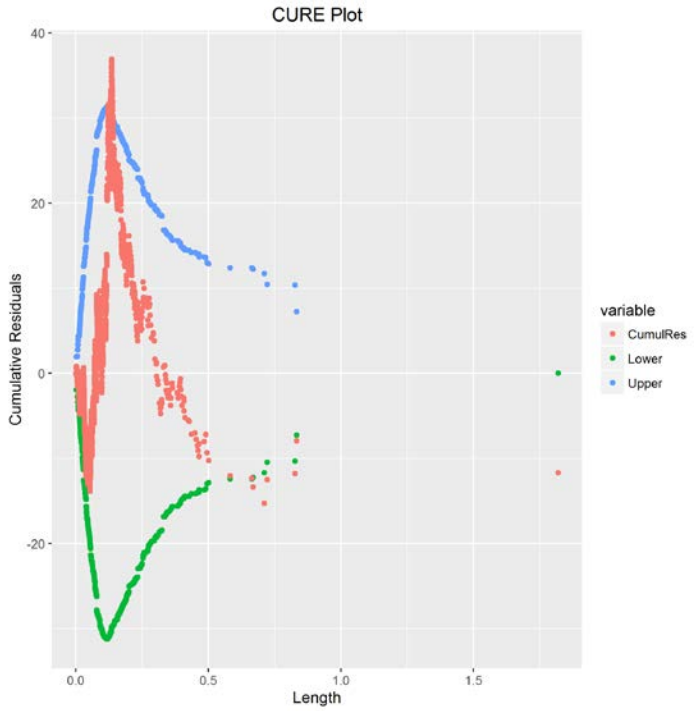


Figure 26. CURE Plot Based on Length for Filter 10

While there is a some devaiation outside of the confidance bands, there is no indication of drifting (dramatically shown in Figure 22). Instead, this devation is likely realted to the vertical jump around 0.25 miles. This CURE plot suggests that there is little bias related to segment length.

5.3.2. Length Categories

For each length, a variety of attribute filters were applied (listed in section 5.2.2). The following visualization was created for each length-attribute combination.

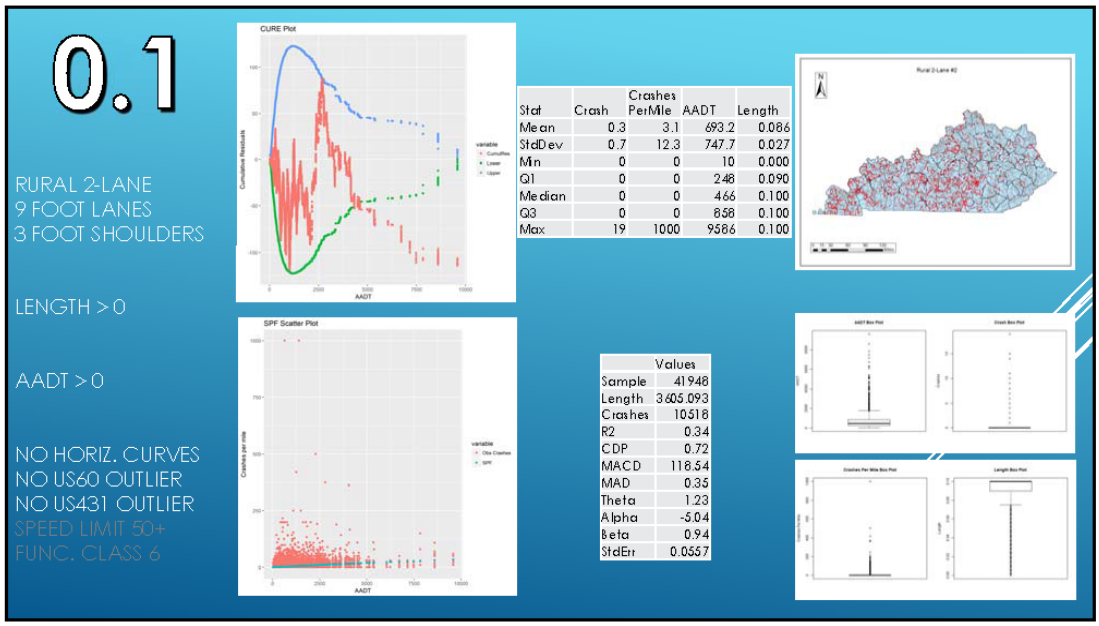


Figure 27. Example Visualization for Length-Attribute Combination

Because of the numerous combinations it was more feasible to manually evaluate small changes between combinations instead of comparing metrics in a single table. Several combinations of length and attribute filters are shown in Appendix N. Each visualization includes a list of filters, length category, CURE and scatter plots, SPF metrics, descriptive statistics, box plots, and a map.

In reviewing the visualizations, a few patterns emerge. The speed limit and functional classification (for rural, 2-lane roads) filters offer no model improvement. This is consistent with the previous finding that most rural, 2-lane segments have a speed limit of 50 mph and, therefore, the filter is unnecessary. As for functional classification, there is evidence that suggests that the existing classification does little to characterize the geometric context of the roadway (Stamatiadis et al., 2016). Therefore, it is understandable that this filter does little to improve homogeneity. Other consistent patterns are that remainders and traffic volume filters do not improve the model. Recall that traffic volume filters improved the models in 0 (scenario 1 compared to scenario 2). This implies there is less correlation between homogeneity and traffic volume for rural 2 lane roadways than for rural parkways.

Longer lengths appear to create the best models. Lengths longer than 0.7 miles have diminishing improvements. While not all length categories are shown in the appendix, there appear to be trade-offs with a length over 0.7. Table 13 compares these categories in more detail.

Table 13. SPF Metrics for Longer Length Categories

	0.3	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**1.1**
Sample	4488	2999	2140	1609	1222	979	782	646	534
Length	1345.974	1199.196	1069.78	965.28	855.298	783.097	703.762	645.982	587.351
Crashes	3214	2819	2624	2419	2148	1962	1815	1671	1551
R2	0.42	0.44	0.49	0.50	0.54	0.55	0.58	0.53	0.56
CDP	3.3	2.1	0.3	0.2	0.2	0.3	0.1	0.3	1.1
MACD	72.5	65.8	53.7	53.5	53.1	38.4	37.5	38.0	30.9
MAD	0.7	0.9	1.0	1.2	1.2	1.3	1.5	1.6	1.7
Theta	2.13	2.31	2.59	2.64	3.18	3.70	3.97	3.59	3.80
Alpha	-5.08	-5.33	-5.17	-5.24	-5.18	-5.47	-5.30	-5.28	-5.36
Beta	0.93	0.97	0.95	0.96	0.95	0.99	0.97	0.97	0.98
StdErr	0.20	0.23	0.28	0.29	0.40	0.52	0.59	0.54	0.60

Notice that for categories above 0.7 miles, some metrics improve yet there is little consistency and the effective difference is insignificant (the PCD is well below 5%, MAD indicates that the model prediction is on average between 1.3 and 1.7 from the actual crash experience). A key point in this comparison is that the regression parameters (alpha and beta) change very little. Recall that the SPF equation is:

$$y = L * e^{a} ADT^{b}$$

The difference between *a* for 0.7 miles (-5.18) and for 0.8 miles (-5.47) may seem significant, however, when evaluated using the above equation, the crash predictions are nearly identical (4.05 and 4.06, respectively) when evaluated with an AADT of 1000 and a length of 1 mile. Furthermore, the fact that beta is nearly one suggests that ADT could be modeled as an offset (discussed in the next section), which reduces the number of regression parameters.

Appendix N demonstrates that all length categories perform better than the "Raw" segmentation. Recall that "Raw" represents a variable length based on homogeneity attributes. This suggests there is an advantage to using a fixed length segmentation over variable length for this dataset.

5.3.3. Comparing Model Forms

A set of filters resulting in little omitted variable bias was used to test different functional forms (Filter 10). The metrics generated from Models A, B, and C are shown below.

Table 14. Model Form Comparison of Three Safety Performance Functions

Metric	Model A	Model B	Model C	Notes
R2	0.34	0.29	0.34	Higher values preferred
PCD	2.97%	9.60%	0.42%	Less than 5%
MACD	23.92	31.19	18.93	Lower values preferred
MAD	0.66	0.66	0.65	Lower values preferred
Theta	1.48	1.39	1.56	Higher values preferred
AIC	1380.90	1385.36	1377.16	Lower values preferred

It should be noted, in contrast to the previous finding, that parameter b (beta) is less than one. The implication here is that there is a non-linear relationship between crashes and AADT. Moreover, Model C suggests that this is also true with length. The distinction between this analysis and the previous analysis is fixed versus variable length. Based on these metrics, Model C outperforms the others in all aspects. The associated CURE plots are shown Figure 28.

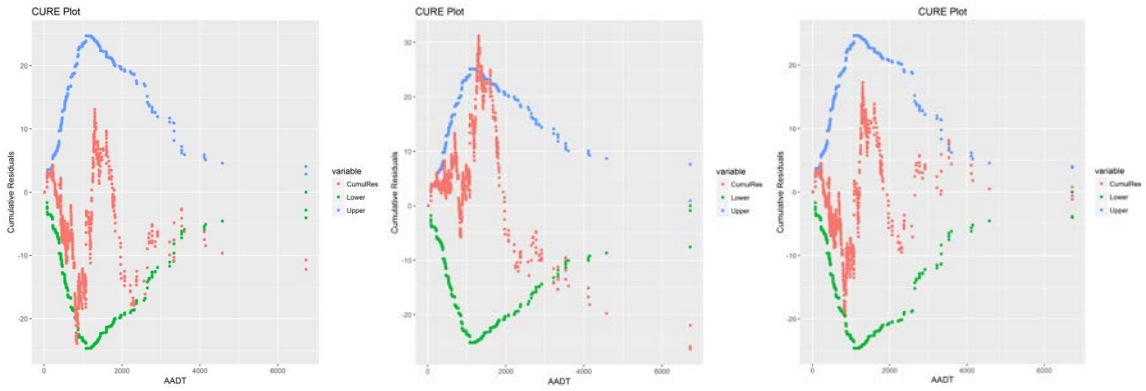


Figure 28. CURE Plots from Three Models Compared (A, B, and C, left to right)

It should be noted that Model C is just as easy to implement as the other two models. For this dataset, there is seemingly no reason not to use Model C's form. Another consideration when selecting a model form is the model prediction. Predictions using realistic values for AADT and length can be computed using the models based on the resulting regression parameters. Consider two predictions:

- Prediction 1 with an AADT of 1,000 at 0.5 miles.
- Prediction 2 with an AADT of 2,000 at 1.8 miles.

Table 15. Model Form Comparison of Prediction Results

Parameter	**Model A**	**Model B**	**Model C**
	$AADT^b * L * e^a$	$\mathrm{AADT} * \mathrm{L} * 365 * 10^{-6} * \mathrm{e}^{\mathrm{a}}$	$AADT^b * L^c * e^a$
a	-4.08	2.19	-4.46
b	0.76	n/a	0.74
c	n/a	n/a	0.68
Prediction 1	1.6	1.6	1.2
Prediction 2	9.8	11.7	4.8

Table 15 compares the resulting predictions in crashes per mile for each model. Notice for prediction 1, the model predictions are similar yet Models A and B over-predict slightly when compared to Model C. At a higher AADT and a longer length (1.8 miles is the maximum length for this dataset), the over-prediction worsens. The implication here is that while Model C might only seem marginally better than the other two models when comparing goodness-of-fit measures, however, the predictions vary wildly.

5.3.4. Length-Based Overdispersion

In contrast to the other comparisons thus far, overdispersion has a different implication on the modeling process. Two different regression model packages within R (discussed in detail in Appendix F) were used to compare variable and fixed dispersion for the same dataset. The reported regression parameters differed slightly, however, the difference was so nominal that the parameters were equal to two decimals places. As shown previously, such a small difference in regression parameter values has little

influence on the model prediction[15]. Therefore, it was determined that this difference was negligible.

The only other implication that overdispersion has in the modeling process is in the weight calculation (recall the weight component from Equation 2)[16]. When using variable dispersion, the weight equation becomes:

$$Weight_i = \frac{1}{1+A*L^B*\frac{SPF_i}{L_i}} \qquad (7)$$

When combined with Equation 3, PCR at site *i* becomes a function of overdispersion as such:

$$PCR_i = \frac{1}{1+A*L^B*\frac{SPF_i}{L_i}} * SPF_i + \left(1 - \frac{1}{1+A*L^B*\frac{SPF_i}{L_i}}\right) * OC_i \qquad (8)$$

While it was determined that length-based dispersion has no impact on the SPF development process (as described above), it could have implications on the PCR calculations. The filter process from Filter 10 and the functional form of Model C were used to compare constant and variable dispersion. Recall the functional form and dispersion formula used:

$$AADT^b * L^c * e^a$$

$$k = A * L^B$$

The overdispersion and the resulting regression parameters are shown in 0.

[15] The difference between a=-5.53 and -5.54 would equate to 0.04 crashes (compared with an AADT of 1000 and a length of 1 mile).

[16] Overdispersion also has an effect on the confidence boundaries in CURE plots but this influence was beyond the scope of this analysis.

Table 16. Constant and Variable Dispersion Parameters Compared

	Constant Dispersion	Variable Dispersion	
Theta	1.57	0.73	(average)
		0.46	(min)
		1.15	(max)
a	-4.83	-4.83	
b	0.86	0.86	
c	0.98	0.98	
A	n/a	0.50	
B	n/a	-0.12	

Two important points should be made about the dispersion parameters A and B. First, A is reported by both Stata and R as the ln(A), therefore, exponential transformation is required (i.e. $A = e^{\ln(A)}$). Furthermore, the gnlr library used in R mistakenly reports A and B as -A and -B. For this analysis; Stata was used to verify the parameters. More discussion on this issue is in Appendix F.

The above parameters were used to calculate PCR for each segment and for each model. The PCRs were ranked and Pearson's correlation and Spearman's Rho were calculated to compare the resulting lists. The values of PCRs are also compared below.

Table 17. PCRs For Constant and Variable Dispersion Compared

PCR	Constant Dispersion	Variable Dispersion
Average	0.00	-0.04
Min	-2.33	-3.12
Max	5.89	5.83

Remarkably, the PCR rankings were nearly identical with a Pearson's Correlation of 0.996 and a Spearman's Rho of 0.994. This implies that for rural, 2-lane roads there is seemingly no advantage to variable dispersion. Even the maximum PCR magnitudes were fairly comparable.

5.4. Conclusions

This chapter examined both length and attributes in the context of SPF development. The interaction between both factors uncovered aspects that can be used

to improve modeling. When considering fixed-length segments there seemed to be little need to parametrize AADT (i.e. $AADT^b$) as *b* was near one. There is a definitive advantage to reducing model complexity as the addition of variables can lead to overfitting (Srinivasan and Bauer, 2013a). In contrast, when comparing model forms using a variable length, not only did a parametrized AADT produce a better model, but length is better modeled when parametrized too. When a fixed length is used it is intuitive to expect no advantage to adding a parameter (all segments have the same length). The key point here, however, is that the functional form of an SPF may be sensitive to the segmentation of the network.

Another aspect of this analysis is the importance of checking for data errors. The advantages of an automated development process come at the price of undetected data errors. It is worth emphasizing that a variety of cross-checks can help detect such errors. Descriptive statistics and CURE plots can offer quantitative comparisons but mapping data can help cross-check geographical distributions. Furthermore, CURE plots can be helpful for detecting where modeling performs best either with respect to AADT, length, or other variables. It should be noted that it is more feasible to use a variable that has a wide range of values such as length or AADT as opposed to a narrow range like shoulder or lane widths. A CURE plot with shoulder width on the x-axis would be too coarse as to provide a meaningful assessment of model fit as shoulder width would only be comprised of about a dozen values.

Lastly, this analysis found no advantage to variable dispersion in terms of SPF development and the lists generated based on PCR. It is important to point out that this does not imply that length-based dispersion is unhelpful as others have found it to be (Hauer, 2001, Cafiso et al., 2010, Geedipally, 2008), rather, it was found unnecessary for the data analyzed.

Chapter 6. Conclusions

6.1. Summary

Experience in Kentucky and a consensus in the body of knowledge suggest that a science-based approach employing EB is more effective than traditional methods. The HSM suggests EB along with the predictive methods of SPFs for network screening. This research addressed three questions in the context network screening as a means to identify hazardous locations.

What are the statistical implications of segment length when performing safety analysis?

There are trade-offs between segment lengths when applied to near homogeneous rural parkways (multi-lane divided facilities similar to interstates) in Kentucky. Many of the goodness-of-fit metrics improve with increasing segment length, however, the applicability of the models is reduced as sample sizes become lower than HSM recommendations. Inversely, average deviation improves (lowers) with shorter segment lengths. For this application, a segment length of 2 miles seems to produce a segment length where the metrics are ideal.

This research focused on a specific crash type on a specific facility. It is recommended that the methodologies outlined in this research are applied to other facility and crash types as the results may differ. Specific crash types can be modeled to help prioritize where to apply a specific countermeasure. The optimal segment length is likely to change based on the countermeasure. Furthermore, these results suggest the need to analyze attributes with respect to roadway homogeneity.

What are the implications of roadway homogeneity on safety analysis?

Rural, 2-lane roadways were used to evaluate the role of attributes in network screening. A single attribute was understandably unable to adequately explain crash variation in the form of an SPF. Rather, combinations of attributes were needed to develop meaningful SPFs. Ranges of attributes offered no improvement to the modeling

process. In Kentucky, a set of attribute filters were identified as the ideal filter that produced the best model while maintaining practical real-world applications. Recall that too many filters limit the applicability of the SPF in that they require too many adjustments (in the form of CMFs) to apply to other segments.

Base conditions from the HSM can be used but it is more logical to determine base conditions based on the most predominate attributes for a given facility type utilizing local data. Analytical tools such as CURE plots, maps, and goodness-of-fit metrics should be utilized during this exploration in an effort to find the ideal composition of attributes. This step can also help identify data errors in either the crash or roadway databases. Interaction, while difficult to detect, can be considered during this process too. Models can be improved by including interaction terms that help explain variation caused by the combination of more than one attribute.

What are the trade-offs between homogeneity and segment length on safety analysis?

Both length and attributes can be evaluated together and, their interaction might produce different results compared to the separate analyses. The ideal model form suggested a linear relationship between traffic volume and crashes when applied to fixed length segmentation (consistent with Cafiso et al., 2013). In contrast, a non-linear relationship was found ideal when applied to variable lengths (consistent with Srinivasan et al., 2011). A key point here is that there may not be a single ideal model form. Also, shorter segments might be more sensitive to boundary effects, especially when considering the accuracy of the crash data. A non-linear length term could help account for boundary effects in short versus long segments.

The use of overdispersion as a function of length is recommended in the HSM for both rural, 2-lane and rural multilane facilities (see HSM's equations 10-7, 11-8, and 11-10). In contrast, a constant overdispersion is recommended for urban and suburban arterials. In this research, length-based overdispersion showed no effect on the SPF development or the resulting network screening. Although, other recent research (Hauer, 2001, Cafiso et al., 2010, Geedipally, 2008) does find an impact. While this step

does add some complexity, there is seemingly no downside to employing variable overdispersion. It is therefore recommended as step in the SPF development process.

6.2. Discussion

While there are a variety of segmentation techniques, this research focused on network-based approach in contrast to a crash-based approach. Crash-based segmentation, while likely to identify optimal segments for safety analysis (Lu et al., 2013; Depaire et al., 2008) is less practical than a fixed segment based on roadway data (Cafiso et al., 2013). Moreover, crash location data may not be accurate enough to warrant such segmentation techniques (Green and Agent, 2011, Ogle et al., 2011). A crash-based approach would also require the weighting of changing attributes within a segment. As shown above, safety analysis can be very sensitive to attributes, even ranges of attributes. The network-based approach is also very applicable for a network-wide countermeasure prioritization especially if the countermeasure applies to a specific geometric attribute combination (e.g. high friction surface on curves, or centerline rumble stripes on undivided roads).

The modeling process seems to be more nuanced than traditional crash analysis such as critical rate factors. When modeling, it is important to consider a variety of implications. Segmentation, model form, and attribute filters are just some of the considerations. As Hauer suggests in his title, *The Art of Regression Modeling in Road Safety*, there is "art" to the process. SPF development tools are essential to the exploratory nature of modeling. A less optimal model could be developed, for example, without an efficient way to test for interaction or omitted variable bias. Tools like The Calibrator, IHSDM, and SPF-R can offer improvements over a manual process for SPF assessment and development.

6.3. Limitations

While this research followed the recommendation to explore other fixed length sizes, future research in this area could employ other evaluation techniques used by Cafiso et al. (2013). Sensitivity, specificity, and QIC were used to compare models and

segmentation techniques. These same tools could be applied to this research and they might help to further refine the recommendations. Researchers have also used variable significance tests to determine the effect of variables while this research used CURE plots and other metrics as a proxy for significance.

This research was conducted on roadway segments particularly as length was an important factor. However, many of the tools and methods could be applied to intersections as well. Moreover, this research focused on rural roadways, yet the same principles could also be applied to urban facilities. Urban segments are typically shorter and have more complex attributes when compared to rural segments. These factors may influence the effect that segment length and length-based overdispersion may have on model development. As stated, there is likely no optimal length that applies to all facility types.

As SPF development continues to grow in the United States the demand for SPF development tools will increase as well. At the time of this writing there are many ways to develop SPFs. Excel tools have been created that use Solver to perform regression to develop SFPs. Advanced knowledge of Excel and familiarity with the SPF worksheet is helpful in developing SPFs. Workshops can also help SPF developers with the implementation of such tools. These tools offer a lot more control over SPF development, but can represent a barrier to entry for a novice at SPF development. Statisticians and programmers may be more comfortable using SAS, R, or SPSS. These solutions typically require knowledge of the software. Moreover, without advanced programming, data must be exported from a crash database, then imported into the statistics program, and finally exported into a solution such as The Calibrator to adequately evaluate SPFs. This multistep process can hinder the development process by adding complications and slowing down model development.

The model forms in this research were limited to the power function. Research has shown that other functional forms may provide a better model fit, such as the

sigmoid functional form (Kononov, 2011). Such models are not as easily implemented as they require the use of Neural Network methodologies.

The use of calibration was not employed in this research, but could easily be used to apply models to other datasets. While the site-specific SPFs are certainly ideal, many agencies lack the resources to develop SPFs. In this case calibration is very desirable alternative.

6.4. Recommendations

The research presented here was distilled down to produce a decision diagram to help with the SPF development process (Figure 29).

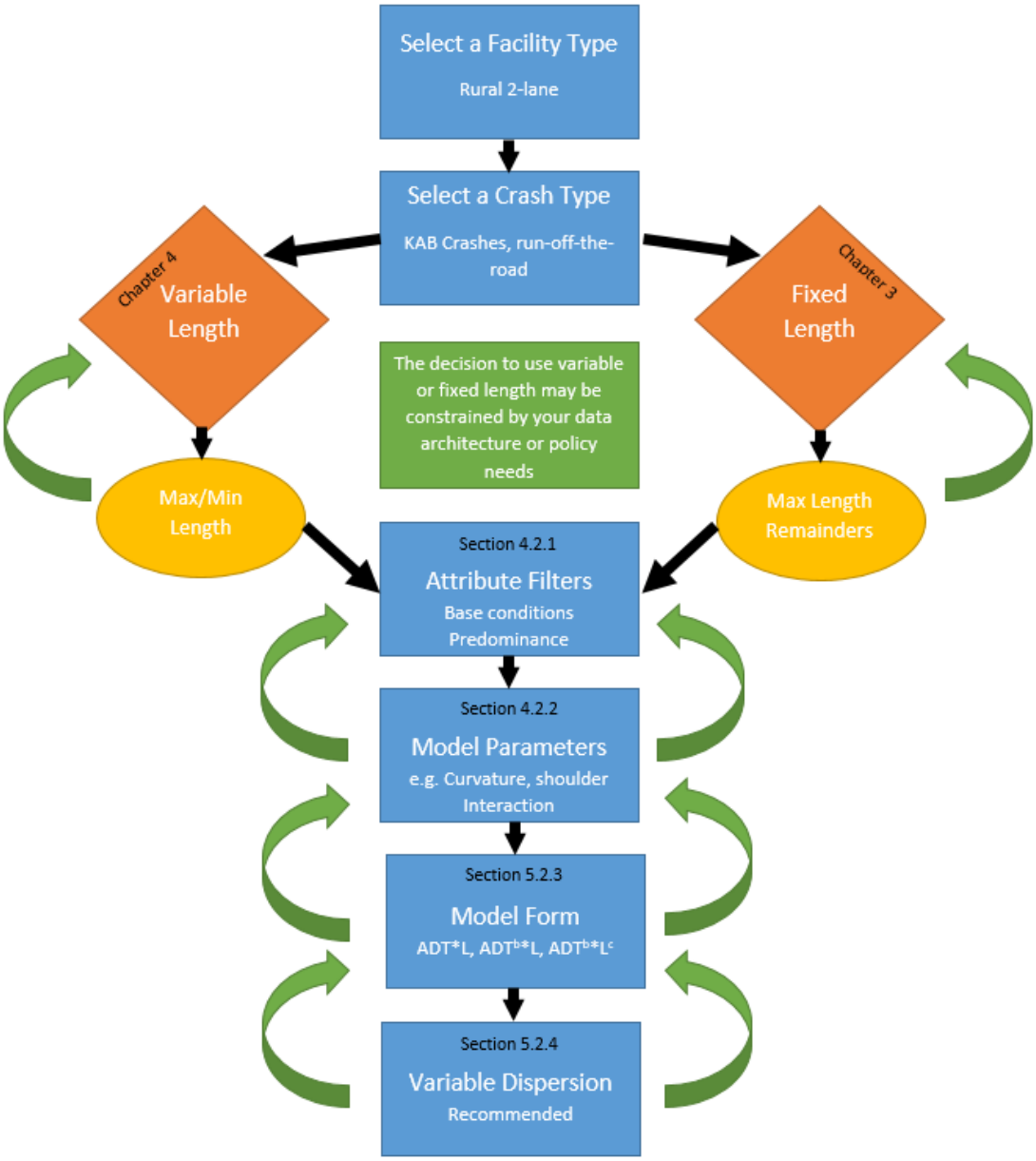


Figure 29. Framework for Analysis of Proper Segmentation for SPF Development

In each step, it is recommended to use CURE plots and goodness-of-fit metrics to evaluate the sensitivity of the decision step. For instance, the decision to select variable versus fixed length might have a strong impact on model quality. The other decisions might require more evaluation steps. Several filters can be applied and tested during the attribute filter step. As shown above in the form of green arrows, reevaluation should be considered at each step. In this context, CURE plots and other metrics should be used

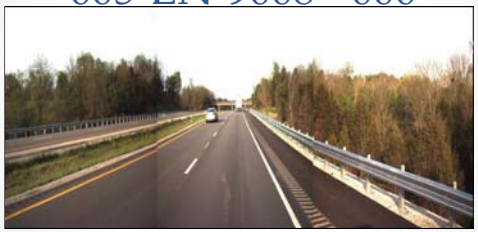
to test the effect of a decision point. For example, when considering model form, several forms should be evaluated for a given dataset and the one producing the best metrics should be selected.

The research within, while developed for Kentucky data, can be applied to other data sources. This is especially true as the framework outlined includes evaluation with each step. These evaluation steps will help identify the optimal segmentation length and attribute filters which may likely differ for other data sources.

Appendix A – Representative Images of Parkways From Kentucky's Photolog

003-BG-9002 -000

005-LN-9008 -000



024-EB-9004 -000

025-KY-9000 -000

030-AU-9005 -000

042-JC-9003 -000

043-WK-9001 -000

047-BG-9002 -000

047-WK-9001_-000

051-AU-9005_-000

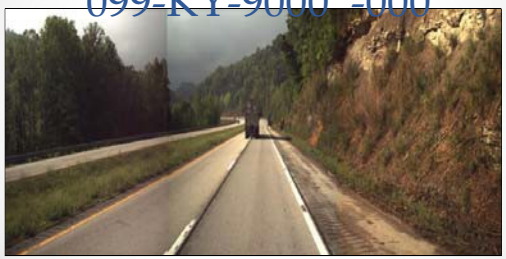
063-HR-9006_-000

079-JC-9003_-000

089-WK-9001_-000

090-BG-9002_-000

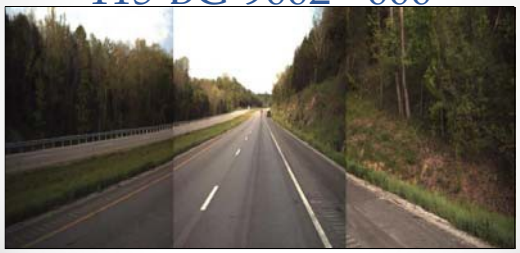
099-KY-9000_-000



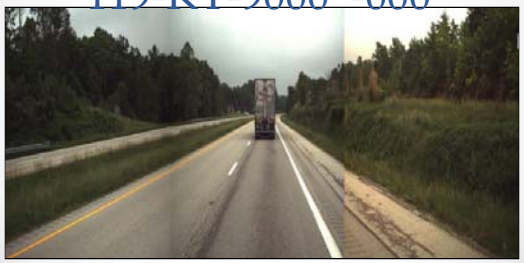
100-LN-9008_-000

114-WN-9007_-000

115-BG-9002_-000



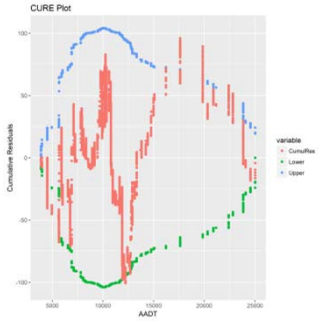
119-KY-9000_-000



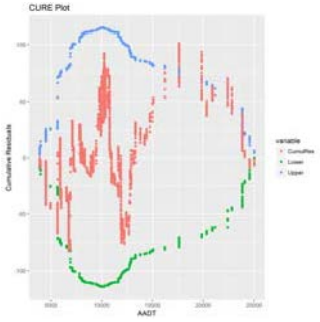
120-BG-9002_-000

Appendix B – CURE Plots for Scenario 1

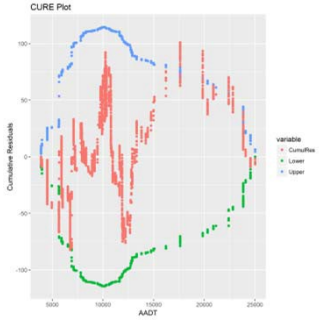
Scenario 1 0.1



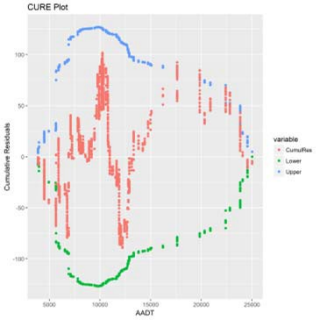
Scenario 1 0.2



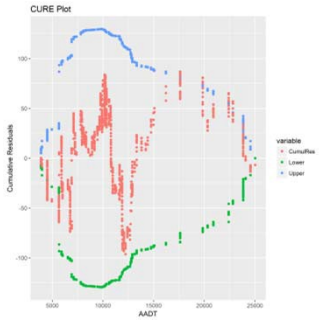
Scenario 1 0.3



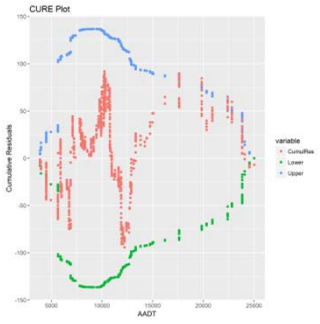
Scenario 1 0.4



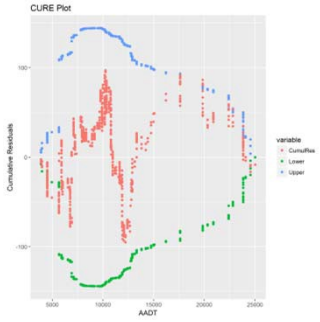
Scenario 1 0.5



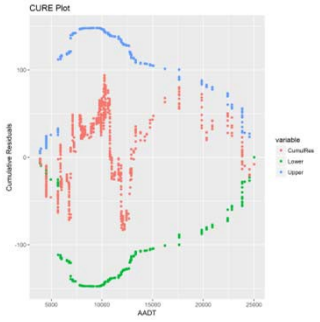
Scenario 1 0.6



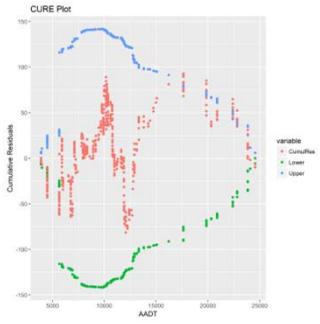
Scenario 1 0.7



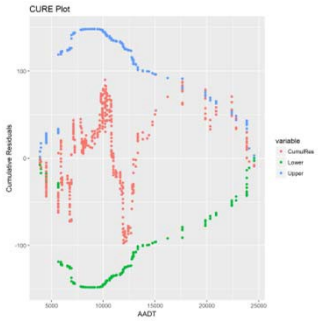
Scenario 1 0.8



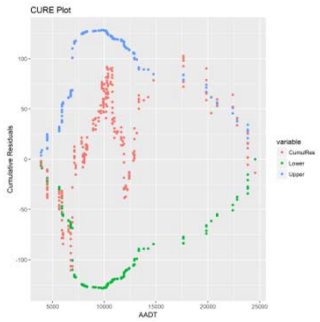
Scenario 1 0.9



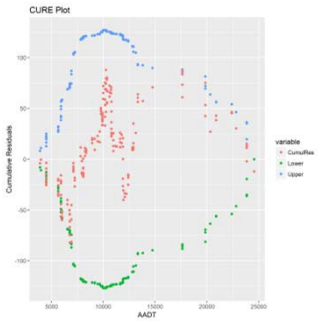
Scenario 1 1.0



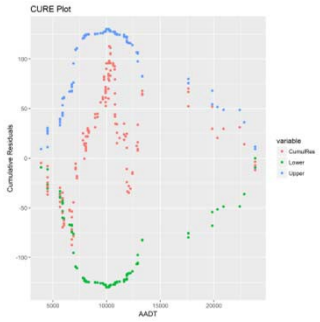
Scenario 1 1.5



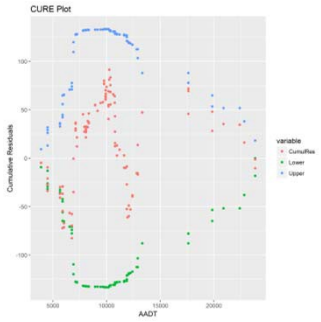
Scenario 1 2.0



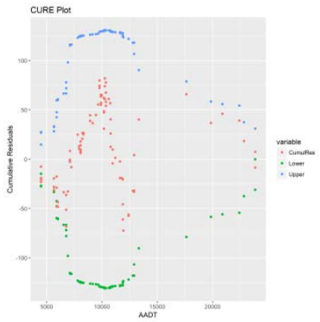
Scenario 1 2.5



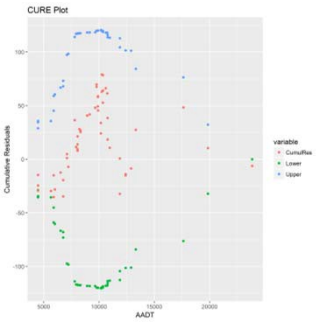
Scenario 1 3.0



Scenario 1 3.5

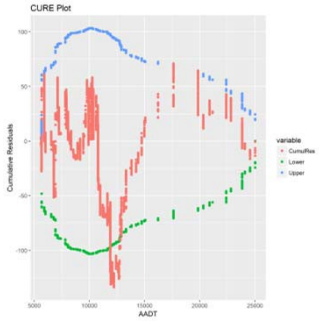


Scenario 1 5.0

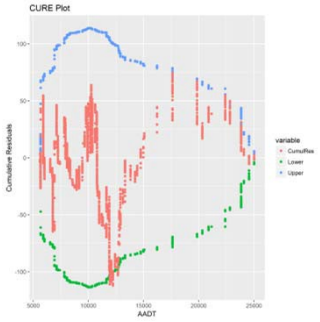


Appendix C – CURE Plots for Scenario 2

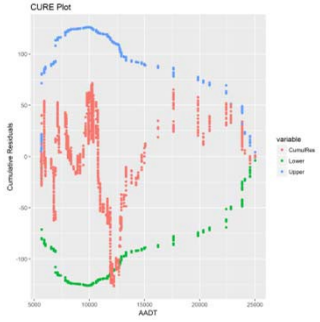
Scenario 2 0.1 Miles



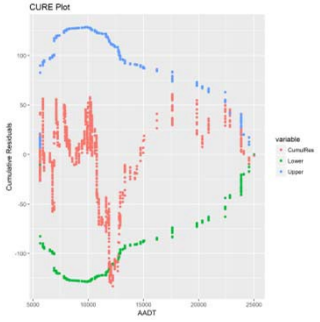
Scenario 2 0.2 Miles



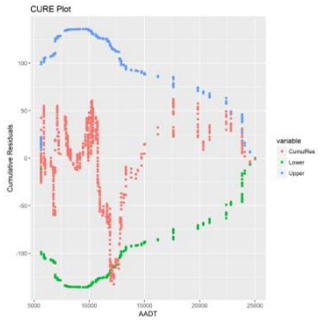
Scenario 2 0.3 Miles



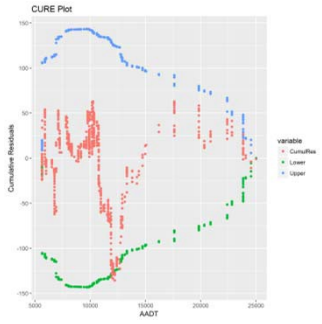
Scenario 2 0.4 Miles



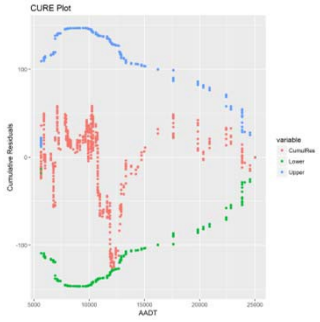
Scenario 2 0.5 Miles



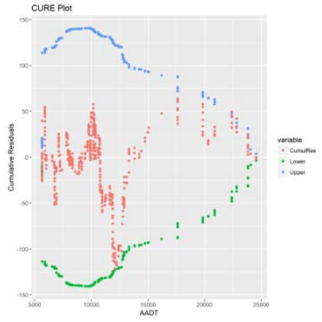
Scenario 2 0.6 Miles



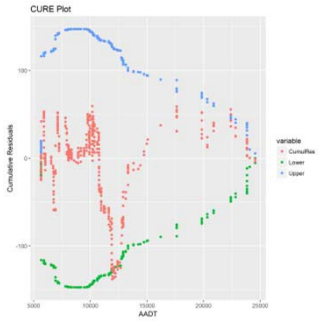
Scenario 2 0.7 Miles



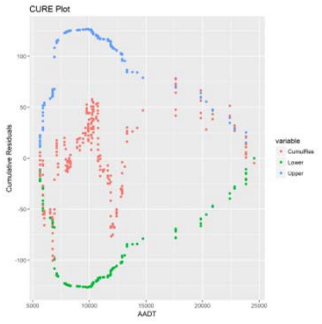
Scenario 2 0.8 Miles



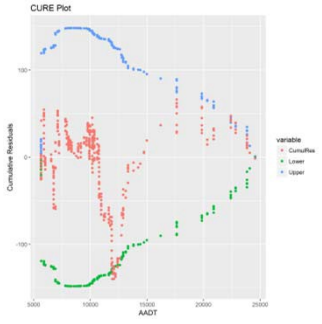
Scenario 2 0.9 Miles



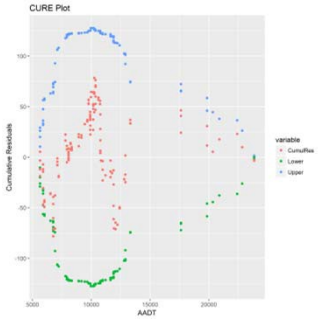
Scenario 2 1.0 Miles



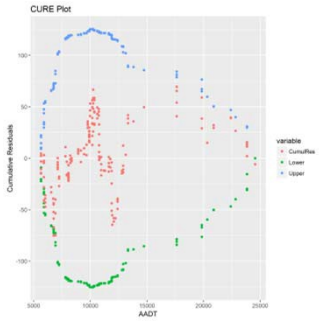
Scenario 2 1.5 Miles



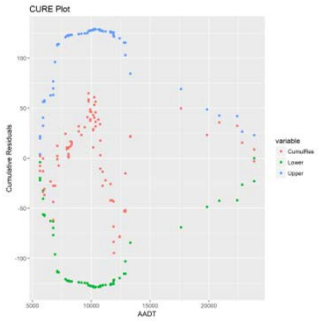
Scenario 2 2.0 Miles



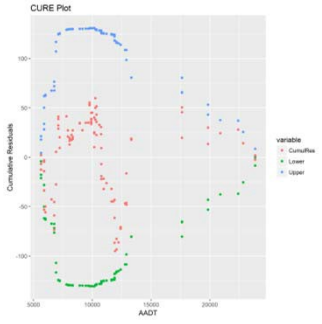
Scenario 2 2.5 Miles



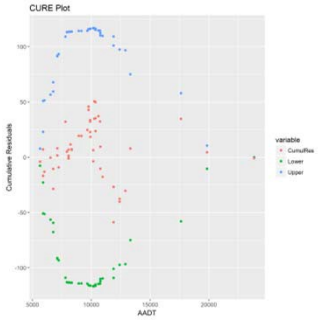
Scenario 2 3.0 Miles



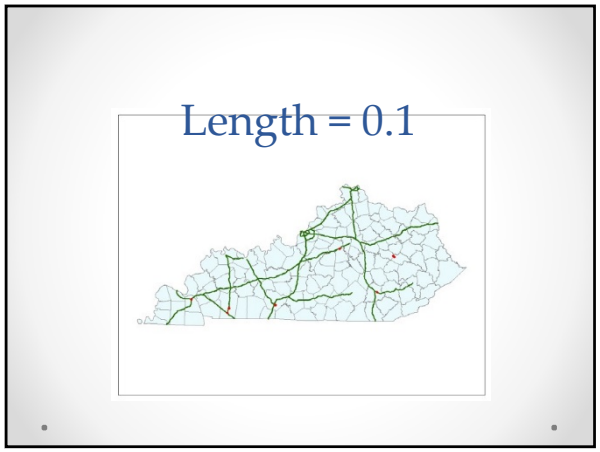
Scenario 2 3.5 Miles

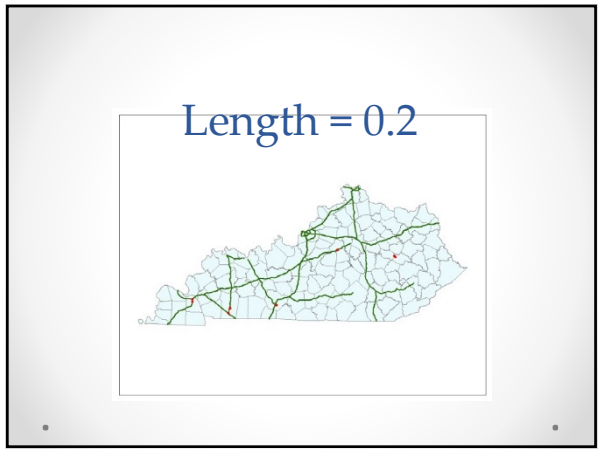


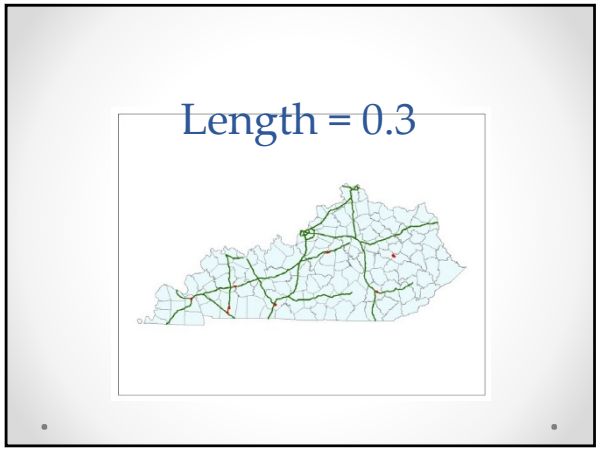
Scenario 2 5.0 Miles

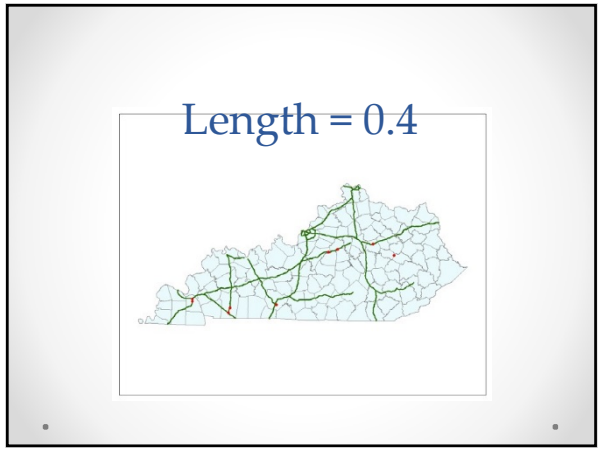


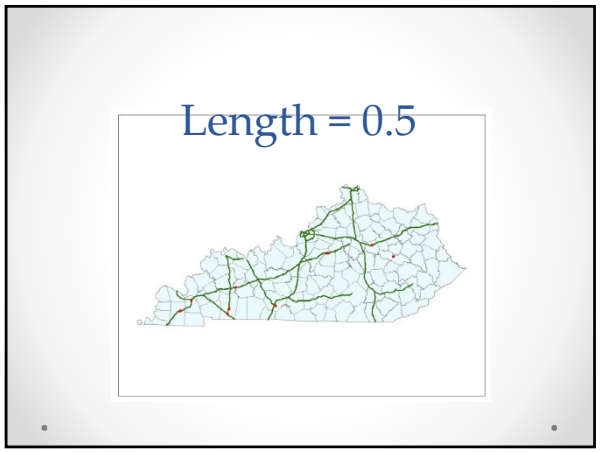
Appendix D – Top Ten PCR Segments by Segment Length

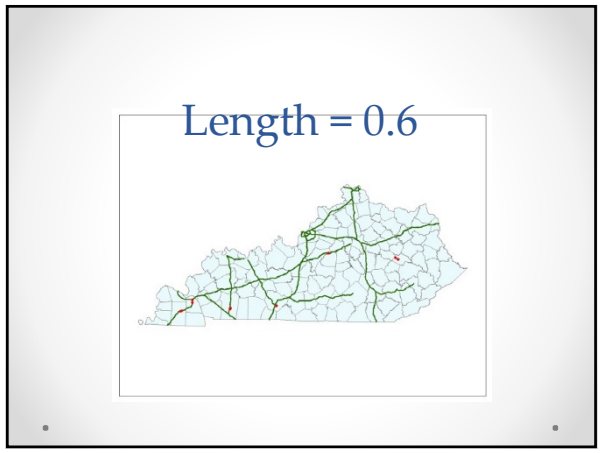


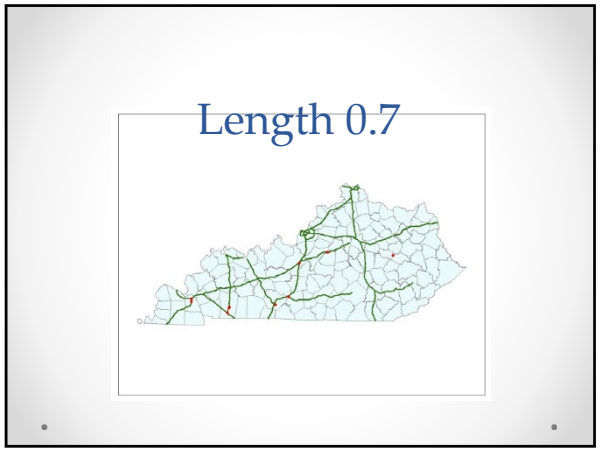


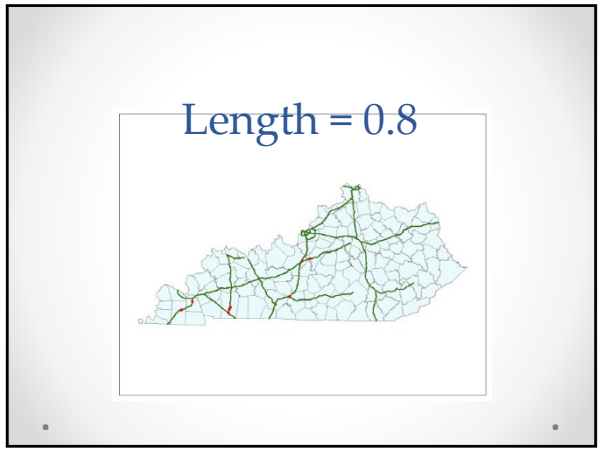





Length = 0.5


Length = 0.6


Length 0.7


Length = 0.8

Length = 0.9

Length = 1

Length = 1.5

Length = 2

Length = 2.5

Length = 3

Length = 3.5

Length = 5

Appendix E – Photolog Images of Frequently Occurring Top Ten PCRs

Warren County:
114-WN-9007-000
Mile: 0 – 1
Beginning of Route: (Approx. Mile: 0)

Right Curve: 1 Occurrences

Exit Ramp: 1 Occurrences

Christian County:
024-EB-9004-000
Mile: 0.0 – 8.5
On Ramp (Approx. Mile: 0.0)

Merge: 5 Occurrences

Bridge Entrance: 7 Occurrences

Bridge: 7 Occurrences

Overpass: 3 Occurrences

Guardrail: Right Shoulder: 4 Occurrences

Exit Ramp: 3 Occurrences

Curve/Guardrail: Median: 1 Occurrences

Powell County:
099-KY-9000-000
Mile: 32.5 – 35.8
Beginning of Route: (Approx. Mile: 32.5)

Exit Ramp: 1 Occurrences

Bridge: 2 Occurrences

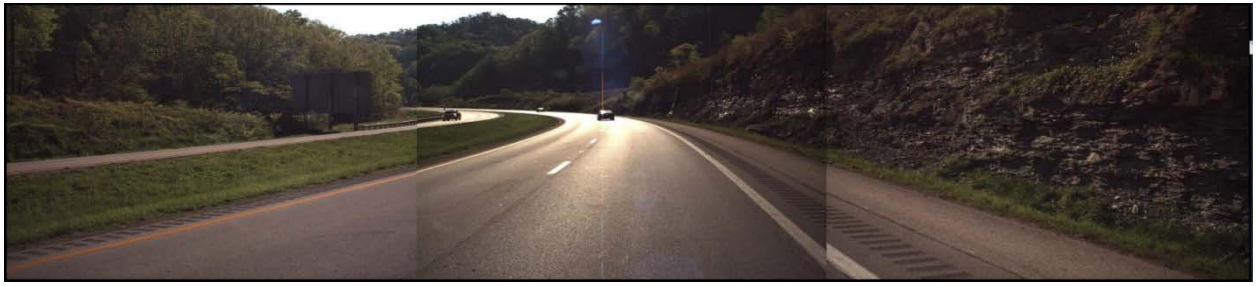
Merge: 1 Occurrences

Right Curve with Guardrail: 4 Occurrences

Overpass: 2 Occurrences

Right Guardrail: 3 Occurrences

Left Curve: 1 Occurrence



Nelson County:
090-BG-9002-000
Mile: 25.0 – 31.0
Beginning of Route: (Approx. Mile: 25.0)

Guardrail-Right Shoulder: 10 Occurrences

Right Curve: 2 Occurrences

Left Curve: 1 Occurrences

Overpass: 1 Occurrences

Graves County:
042-JC-9003 -000
Mile: 11.0 – 16.0
Beginning of Route: (Approx. Mile: 11.0)

Guardrail - Right Shoulder: 12 Occurrences

Overpass: 3 Occurrences

Bridge Entrance: 1 Occurrences

Bridge: 1 Occurrences

Overpass with On and Off Ramp: 1 Occurrences

Appendix F – SPF-R User's Guide

Introduction

The following guide describes an automation tool that helps to develop and assess Safety Performance Functions (SPFs). SPFs can be straightforward to develop. The process requires a database of roadway segments (or intersections) containing segment length, number of crashes, and traffic volumes for each site. A generalized linear model using negative binomial regression is used to create an equation that relates observed crashes to traffic volume and length (as well as other independent variables, if desired). Statistical packages such as SPSS, SAS, Stata, and R Studio perform this regression easily with built-in tools. The process can also be achieved in Microsoft Excel using solver or custom functions.

The above-mentioned tools are simple enough to generate an SPF manually but can be cumbersome when trying to improve model development, which requires several iterations while filtering the roadway dataset. Moreover, the creation of CURE Plots requires several steps and considerable amount of overhead for large database. FHWA's Calibrator tool readily generates CURE Plots but is separate from the SPF development. This separation necessitates several intermediate and repetitive steps.

The program "R Studio" can be used to simplify and streamline the SPF development and assessment process for large datasets, and code was written to automate the entire process. The following sections describe each section of the R Code – named "SPF-R." The source code is available on GitHub at: http://github.com/irkgreen/SPF-R. The code can be modified as needed and meaningful changes may be committed to the GitHub repository so that other safety professionals can benefit from the enhancements. GitHub is an online, collaborative tool that allows anyone to download the source code and contribute.

The code requires an input file in CSV-format containing roadway segments or intersections. Each record must contain, at a minimum, traffic volume (major and minor

for intersections), length (for roadway segments), and crashes. Optionally, the input file can contain data about the roadway (shoulder width, lane width, curvature, etc.) and crash counts by severity.

By default, SPF-R develops an SPF based on the input file using the model form shown in Equation 1. A CURE Plot, scatter plot, and an Excel document containing the model parameters and data are all saved to folder defined by the user. The following sections describe how to use and modify SPF-R.

SPF-R Prerequisites

The above referenced source code was intended for use with R Studio. However, it may work with other installations of R. A separate installation of Rtools as well four R Packages are required. The following list describes the required tools:

- R Studio - https://www.rstudio.com/products/rstudio/download/
- Rtools - https://cran.r-project.org/bin/windows/Rtools/ [17]
- Required packages: knitr, ggplot2, openxls, installr

An analyst may download and install both R Studio and Rtools from the links provided. To install the required packages, the user will choose run Tools>Packages from the R Studio menu and enter the comma-separated list of packages described above. R Studio provides sufficient error messaging to help with most installation errors.

SPR-R Code Description

The following describes the purpose of each section of R-code and provides advice on modification of code for other uses. Line numbers from the February 15, 2017 "commit" on GitHub will be used as references. A "commit" is an upload to the repository. It is likely that the repository will be modified after the release of this document; therefore, please refer to the SHA hash b376201f1765f3fe3b0adadbbdd794db267c2cde.

[17] When installing Rtools, make sure that the box is checked to have the installer edit your PATH.

Lines 1-17

The first few lines disable echo, clear the workspace, load libraries, and store the version number. The workspace is cleared to simplify debugging as the previous workspace memory can make it difficult to isolate errors. That said, this line can be removed if the user intends to use previously stored data (warning – clearing the workspace will delete R Studio's stored data). Edit the version number as needed; however, the other lines should stay unchanged. Editing the version is important so that results are tied to a specific version of SPF-R if changes are made.

Lines 19-27

This code is used to specify an alternate location for the Windows User's folder. For most users, the default is sufficient. However, an alternate user folder can be hardcoded using the computer's computer name as shown in lines 21 and 23. This folder is a base folder for input data as described below.

Lines 29-50

This section is used to map the data columns (from the input file – discussed below) to the variables used to develop an SPF. You must specify a data column for TotalColumn, AADTColumn, and LengthColumn. These columns represent the total crashes, traffic volume, and length, respectively, for each site. The total crashes at each site could be for all crashes or a specific crash type. TotalColumn must be used if only one specific crash severity is being analyzed (e.g. fatal only crashes). However, if SPFs are to be developed for more than one severity type then the KABCO columns can be used to simply the SPF development process. In this case, the input dataset must include a column for each severity type. For example, you can develop SPFs for five severity types by using the following mappings:

- TotalColumn = "Total" #The title of the column containing All Crashes (KABCO)
- KABCColumn = "KABC" #The title of the column containing KABC Crashes
- KABColumn = "KAB" #The title of the column containing KAB Crashes
- KAColumn = "KA" #The title of the column containing KA Crashes
- KColumn = "Fatal" #The title of the column containing K Only Crashes

Spaces should be avoided in all column names, however, you can replace spaces with a period: "Total.Crashes"

Classes can be used if your dataset contains more than one group of roadway segments or intersection types. For example, the dataset may contain several districts across a state. SPF-R can be used to build a separate SPF for each district. The mapped ClassColumn must contain a positive integer (e.g. district number). The lowest and highest integers must be defined with ClassStart and ClassEnd. Gaps in the range should be avoided. For instance, a dataset might include data for two highway types: rural, 2-lane roads and urban 4-lane divided roads. In this dataset, all of the rural, 2-lane roads could be coded as HighwayType = 1 and the others as HighwayType =2. ClassColumn would be set to “HighwayType” with ClassStart = 1 and ClassEnd = 2.

The CSVPath variable is used to set the location of the input CSV file. This file must contain all of the fields mapped above. The CSV must have a title row. The location is relative to the folder set in line 26. Notice that R uses forward slashes (“/”) for file paths.

The OutputProject_Base is used to define the name of the output folder. The myFilter_Base is used to apply a global filter to the data. Generally, it is good practice to specify that traffic volume and length are both greater than zero to avoid errors in the regression. You can reference a field in two ways:

- Directly – data$FieldName where FieldName is the name of the field in the input CSV
- Using pre-defined variables – data[[VariableName]] where VariableName is TotalColumn or another previously defined field (ideal for dynamic assignment of a variable throughout the code)

It is important to change the OutputProject_Base anytime the myFilter_Base is changed. This will ensure that the modified SPF is saved to another folder instead of overwriting the previous analysis. There is no warning about overwriting folders or files.

The InputData_Base is used to uniquely identify the analysis type. It is recommended that the crash time period and crash type are described in this text

string. This description will be included in the output file. Lastly, initTheta is used to specify a starting point for the overdispersion parameter. This can be adjusted if the regression model is not able to converge. R Code uses *Theta* as opposed to *k* for the overdispersion parameter. *Theta* is the reciprocal of *k*.

Lines 52-55

These comments simply show examples of advanced filters using AND (&) and OR (|) operators. Notice that the presence of parenthesizes is important in developing filters. Text string filters require the use of a single quote (apostrophe). R uses a single equal sign (=) to set a variable, but double equal signs (==) to set a filter to an exact match (as opposed to an inequality such as greater than).

Lines 57-92

These lines simply check for the input dataset and attempt to bind the data. A flag is set to TRUE, if successful.

Lines 94-193

This section represents the main function to develop the model – RunSPF. These statements are not actually executed until called upon later in the code. This may seem a bit counterintuitive, but these lines will be explained in a later section.

Line 196

This line merely checks that the input dataset (CSV) was bound successfully. The following lines will not execute if unsuccessful.

Lines 198-213

This section checks if the user has defined a column of classes. If a class column is set, then the remaining code will loop through each class. In each loop, the base filter is limited to class *i* where *i* is the current class. If no class is defined, then no filter is applied and the loop is only executed once.

Lines 215-222

This section represents the primary SPF initialization. Three variables are temporarily assigned to identify the crash column, the input dataset description, the output folder. The RunSPF function is executed using the temporally assigned variables. Lastly, a message is printed indicating that this code has completed.

Lines 227-272

This section executes the same code as in the previous section however the variables are changed to reference the predefined severity columns, if enabled. The same three variables are used but this time the crash columns are assigned accordingly. Similarly, the severity type is indicated in the description variables.

Lines 94-193 (revisited)

This section develops the SPF and creates the output files. It should be more intuitive now that the other sections have been explained. This function uses temporary variables such that it can be called several times throughout the code. Care has been taken to make all of the inputs and outputs generic. Line numbers are indicated where appropriate below.

A filter is applied using data from the base filter (line 43) and using a defined class (line 208), if applicable (line 97). This new data table is then sorted by the traffic volume column (line 100). The crash column is set to a variable to be used negative binomial model development (line 103). A generalized linear model is used to compute the regression parameters. The natural log is used to generalize the functional form of the SPF so that the parameters are coefficients instead of exponents. As such, the natural log of traffic volume and length are computed (lines 104-105). Optionally, length can be calculated directly from beginning and ending points; however, segments with a length of zero will cause an error in the SPF function. It is therefore recommended that length is included in the input file so that a simple filter can be applied. Theta is initialized on line 45. An effort was made to group all user-defined settings into a few sections of the code.

Line 112 executes regression based on the SPF model form. This code can be altered to support other model forms. A few notes about the syntax:

- The variable to the left of the tilde (~) is the dependent variable – crashes.
- The plus sign is used to separate the independent variables. These are variables that are affected by regression parameter as an exponent (e.g. $AADT^b$ or e^{SW*b}).
- Any additional independent variable need to be added to lines 104-105 so that the column titles are mapped to variables to be used in the glm.nb function.
- A natural log transformation must be computed for any variables lacking the exponent (Euler's number, *e*). Traffic volume (AADT) typically requires this transformation as shown in Equation 1. Variable names that have been transformed should start with "ln" to indicate the transformation.
- Advanced users can modify the code to include interaction terms
- Offset() is used to isolate variables that are not affected by a regression parameter (e.g. Length). These variables should also be transformed using the natural log. Although the current edition of the HSM (AAHSTO, 2010) treats length this way, there is some recent evidence that Length should be modeled similar to AADT. In this case offset() can simply be removed from the R code.

The following table lists three common SPF models and their R Code syntax.

Table F-1. Various SPF Forms and the Corresponding R Code Syntax

Description	Functional Form**	R Code
Typical	$L * e^{a} AADT^{b}$	SPF=glm.nb(crash~lnADT+offset(lnL))
Alternate	$L^{c} * e^{a} AADT^{b}$	SPF=glm.nb(crash~lnADT+lnL)
HSM	$AADT * L * 365 * 10^{-6} * e^{a}$	SPF=glm.nb(crash~offset(HSM*))
Intersection	$L * e^{a} AADT\_Maj^{b1} AADT\_Min^{b2}$	SPF=glm.nb(crash~lnADT1+lnADT2)
Shoulder	$L * e^{a} AADT^{b1} e^{SW*b2}$	SPF=glm.nb(crash~lnADT+SW+offset(lnL))
Interaction	$L * e^{a} AADT^{b1} e^{SW*b2+LW*b3+LW*SW*b}$	SPF=glm.nb(crash~lnADT+SW+LW+SW*LW+offset(lnL))

*HSM = log(data2[[AADTColumn]]*data2[[LengthColumn]]*365*10^-6)

**LW = lane width, SW = shoulder width

Terms that are in exponential functional form (such as e^{b} and e^{SW*b2}) do not require a transformation; however, length, power functions (such as $AADT^{a}$), and any

other terms require a natural log transformation. Transformation is required so that the exponents (a, b, b2) can be treated as coefficients and computed using linear regression. Consider the following transformation:

$Y = L * e^{a} AADT^{b}$
$\ln(Y) = \ln(L * e^{a} * AADT^{b})$
$\ln(Y) = \ln(L) + \ln(e^{a}) + \ln(AADT^{b})$*
$\ln(Y) = \ln(L) + a + b * \ln(AADT)$
where,
$\ln(e^{a}) = a * \ln(e^{1})$
$\ln(e^{1}) = 1$
*natural log identity

Notice that *a* and *b* can now be computed using linear regression with ln(L) as an offset. In this model form *a* is the intercept and *b* is the regression coefficient for AADT. The same transformation can be applied to other model forms using the same natural log identities. All natural log transformations must be computed in the section of code starting at line 104. Moreover, additional parameters (such as b_1 and b_2) must be referenced in the output section near line 167 as discussed later.

More complicated model forms can also be used. In this case, it is advisable to check the R-code syntax using Excel. This is easily accomplished by calculating the prediction using the intended model form from within Excel. From here, the independent variables and model parameters can be referenced directly. The resulting prediction can be compared to the fitted result provided by R – conveniently stored in Excel as well. A perfect match (to several decimals) confirms that the model form was properly converted. For example, consider the fatal and injury SPF for two-lane rural road by Bauer and Harwood as described in the *SPF Development Guide*:

$$N_{FI} = e^{b_0 + b_1 * \ln(AADT) + b_2 * G + b_3 * \ln\left(2 * \frac{5730}{R}\right) * I_{HC} + b_4 \left(\frac{1}{R}\right)\left(\frac{1}{L_c}\right) * I_{HC}}$$

The equivalent R syntax for this model is:

```
 #Point to variables
crash=data2[[CrashColumn]]
lnADT=log(data2[[AADTColumn]])
IHC=data2$IHC
ln2CD=ifelse(data2$CURVEDEG == 0 ,0,log(2*data2$CURVEDEG)*data2$IHC) # omit if DegreeOfCurve is zero**
G=data2$G
CD_L=data2$CURVEDEG*data2$IHC/(5730*data2[[LengthColumn]])

init.theta = initTheta

#####################################################################
SPF=glm.nb(crash~lnADT+G+ln2CD+CD_L)
#####################################################################
```

(Recall that CurveDegree=5730/R)

A variable dispersion can also be used but it requires an additional library. This library will require significant modifications to the remainder of the code, however. The creation of CURE plots, scatter plots, and SPFs metrics are all based on the glm output format. While some of the code might work, much of it will require adjustments. As an alternative, these lines can be commented out and a manually summary can be used to view the model results. The following code shows the essential lines required to employ a variable dispersion.

```
library(gnlm)

#Point to variables
crash=data2[[CrashColumn]]
lnADT=log(data2[[AADTColumn]])
lnL=log(data2[[LengthColumn]])

SPF = gnlr(crash, dist="negative binomial", mu=~exp(a+b*lnADT+c*lnL), shape=~(const+b1*lnL), pmu=list(a=0,b=0,c=0),
pshape=c(0,0))
```

It should be noted that the results of this methodology have been compared to another statistical package (Stata) and there are some discrepancies. The resulting parameters differ slightly (likely variations in the way they are estimated) but not enough to change the predictions. More importantly, the sign of the parameters are opposite. This may imply there is a bug in R's gnlm library (the results from Stata are more intuitive and are likely correct). Validation should be used with other statistical packages before employing this feature. This was observed when both reported parameters were found to be negative in Stata. While this was consistent, it was not exhaustively tested and may not apply in all cases.

Line 116 adds the SPF predictions, residuals, and cumulative residuals to the recently sorted table. The SPF prediction is simply the predicted crashes using the fitted SPF for each record in the dataset. The residuals are the difference between the actual crash experience and the prediction.

The next section (lines 118-146) calculates the information needed to create the CURE Plot. The CURE Plot is a scatter plot of the cumulative residuals versus a sorted variable (typically traffic volume). A standard deviation computation is used to create upper and lower bounds for residuals exceeding 95% confidence boundaries. This section also flags road segments that are outside of the bounds so that the Percent CURE Deviation (PCD) can be computed. The ggplot2 library is used to generate the CURE plot and add labels. The resulting graph is saved as a PNG file to the output folder.

CURE plots can also be generated for other variables. To accomplish this, the data must be sorted by the variable of choice. It is common for length to be used in CURE plots as well as traffic volume. The following code shows how to implement this change (underlined statements can be changed to reference a variable other than length).

```
 #sort by Length
 data3 <- dataout[ order(dataout[[LengthColumn]]),]
 #add new cumul
 dataout2 <- cbind(data3,CumulRes2=cumsum(data3$Residuals))

 #calculate data for CURE plot
 datalimits2 <- data.frame(dataout2$Residuals)
 datalimits2["Length"] <- NA
 datalimits2$Length <- dataout2[[LengthColumn]]
 datalimits2["CumulRes"] <- NA
 datalimits2$CumulRes <- dataout2$CumulRes2
 datalimits2["Squared_Res"] <- NA
 datalimits2$Squared_Res <- datalimits2$dataout2.Residuals^2
 datalimits2["CumulSqRes"] <- NA
 datalimits2$CumulSqRes <- cumsum(datalimits2$Squared_Res)
 datalimits2["SigmaSum"] <- NA
 datalimits2$SigmaSum <- sqrt(datalimits2$CumulSqRes)
 datalimits2["StdDev"] <- NA
 datalimits2$StdDev <- datalimits2$SigmaSum*sqrt(1-datalimits2$CumulSqRes/sum(datalimits2$Squared_Res))
 datalimits2["UpperLimit"] <- NA
 datalimits2$UpperLimit <- datalimits2$StdDev * 1.96
 datalimits2["LowerLimit"] <- NA
 datalimits2$LowerLimit <- datalimits2$StdDev * (-1.96)
 datalimits2["Per_CURE"] <- NA
 datalimits2$Per_CURE <-
ifelse(datalimits2$CumulRes<=datalimits2$UpperLimit,ifelse(datalimits2$CumulRes>=datalimits2$LowerLimit,1,0),0)

 #create CURE plot
 CUREPlot2 <- ggplot(datalimits2, aes(datalimits2$Length, y = value, color = variable)) +
   geom_point(aes(y = UpperLimit, col = "Upper")) +
   geom_point(aes(y = LowerLimit, col = "Lower")) +
   geom_point(aes(y = CumulRes, col = "CumulRes")) +
   ggtitle("CURE Plot") +
   labs(x="Length",y="Cumulative Residuals")
 ggsave(file=paste0(OutPath,OutputProject,"_CURE_L.png"))
```

The same library is used to plot traffic volume versus crashes (actual) per mile (lines 148-154). The SPF predictions are also divided by segment length and plotted to visualize the SPF model. This plot indicates the relative amount of dispersion in the data and is saved to the output folder as a PNG. The scatter plot will include a curve represented by points that describes the shape of the SPF normalized by length. When additional variables are added to the SPF, this curve is obfuscated as each point is affected by more than just AADT (such as lane or shoulder width). In this case it would be more appropriate to plot the SPF at various combinations of the additional variables (e.g. SPFs for lane width of 9 feet, 10 feet, and 11 feet); each with a slightly different shape. This can be added to the output but was beyond the scope of this guide.

The next section (lines 156-170) calculates basic descriptive statistics about the data such as total crashes, mileage, and number of records. Goodness-of-fit measures are also calculated so that similar models can be compared and improved:

- An equivalent analog to R-squared does not exist for negative binomial regression; however, a pseudo-R-squared can be computed.
- PCD is calculated by computing the percentage of segments that are outside of the upper and lower confidence bands from the CURE Plot.
- The Maximum Absolute CURE Deviation is simply the largest (positive or negative) cumulative residual. As described earlier, this can be useful in outlier and data error detection.
- Lastly, the Mean Absolute Deviation (MAD) is computed as the average of the absolute values of the residuals.

These metrics are stored into three arrays including the metric name, the value, and a description. The descriptions, in many cases, include helpful comments such as if higher or lower values are preferred or if there are recommended limits. For instance, the HSM has recommendations for the number of crashes per year and miles in a network for SPF development. It is important to note that these arrays must be altered if there are any changes to the SPF functional form (as described in Table F-1). That is, if a minor AADT is added to the SPF then the corresponding regression coefficient must also be added to the three arrays. The coefficient is referenced using the following code:

```
coef(summary(SPF))["VariableName","Estimate"]
```

The term "VariableName" must be replaced with the variable used in line 112 that corresponds to the coefficient. For instance, the following three lines of code would be used to report the five regression coefficients described in Equation 2 (the altered and added code is underlined).

```
datametrics <- data.frame(Values = c(Sample,Mileage,Crashes,RSquared,PCD,MACD,MAD,SPF$theta
,coef(summary(SPF))["(Intercept)","Estimate"],coef(summary(SPF))["lnADT","Estimate"],coef(summary(SPF))["G","Estimate"],
coef(summary(SPF))["ln2CD","Estimate"],coef(summary(SPF))["CD_L","Estimate"], SPF$SE.theta, SPF$aic, "", "", ""))

 datametrics$Notes <- c("100-200 intersections*","100-200 miles*","300 crashes per year*","Higher values preferred","Less than
5%","Smaller values preferred","Smaller values preferred","Higher values preferred","(b0)","(b1)","(b2)","(b3)","(b4)", "", "",
myFilter, InputData,"*As recommended by FHWA-SA-14-004")

 attr(datametrics, "row.names") <-
c("Sample","Length","Crashes","R2","PCD","MACD","MAD","Theta","Intercept","lnADT","G","ln2CD","CD_L","StdErr","AIC",
"Filter","Input Data","")
```

Care must be taken to ensure that each line is altered similarly such that each array reports the data in the same order.

The next section (lines 172-180) calculates the Potential for Crash Reduction (PCR) using the Empirical Bayes (EB) method as outlined in the HSM. The equation for the Empirical Bayes estimate is:

$$EB[N] = w * E[N] + (1 - w) N$$

where:

EB[N] = EB estimate for site N
E[N]= predicted number of crashes for site N based on SPF
N = number of observed crashes at site N
w = weight equation defined as: $1 / [1 + (E[N]/\theta)]$
θ = over-dispersion parameter (reciprocal of *k*)

It should be noted that R terminology and the above methodology differs slightly from the HSM. R reports the over-dispersion parameter as *theta* which is the reciprocal of *k* as designated by the HSM and most other statistical packages (SPSS, SAS, etc.) Also, the input files used for SPF development are typically created for a five-year period. That is, there is one record per segment with a single traffic volume and an aggregated total of crashes for the entire period. As such, there is no need to total the predicted number of crashes as shown in the HSM in equation 3-10.

The EB estimate is a critical step in the network screening process as it addresses regression-to-the-mean bias. An analyst may be tempted to compare the observed crashes (N) to the prediction from the SPF (E[N]); however, this can potentially be misleading if the observed crashes are uncharacteristically high or low. The EB estimate estimates the magnitude of expected crashes by using the above weight equation.

PCR is then calculated by the following equation:

$$PCR = EB[N] - E[N]$$

This number represents the potential benefit that can be expected if the target crash type is addressed such that the segment of roadway (or intersection) is to become more like the average segment in the road type. That is, if an SPF was developed for

lane departure crashes and a PCR at a site was calculated to be 20.6 crashes, then installing rumble stripes could be expected to eliminate nearly 21 crashes over 5-year period. A Crash Modification Factor (CMF) could be used to quantify this reduction in crashes based on a specific countermeasure.

The final section (lines 182-192) creates an Excel file with the metrics and goodness-of-fit information. Original input data along with all site-specific data (e.g. PCR, weight, SPF prediction, etc.) are also written out to the same Excel document in a separate sheet.

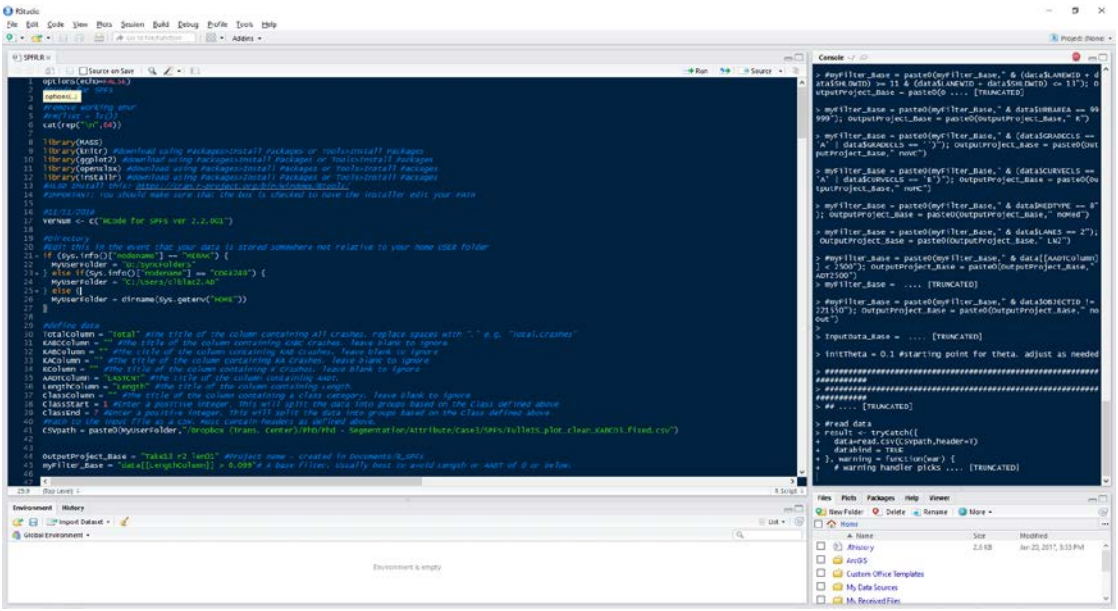
Configuring and Running SPF-R

The SPF development tool can easily be configured to work for a variety of SPF models. Filters can be applied to develop SPFs for specific crash types or to change the roadway geometry. In addition, classes can be used to develop SPFs for several subsets of data. The following is a summary of the lines that are typically changed:

- Line 17 – Version number – It is good practice update this number to indicate significant changes to the code base (please consider sharing any advancements on GitHub as well).
- Line 26 – User folder – This variable is based on the current Windows User's folder. This is helpful as this path is different for every user.
- Lines 30-45 – Main Settings – As discussed earlier, these settings specify column names, classes, severity outputs, main filter, and the input path (line 41). The input path can be hard coded and will ignore the User Folder if convenient (e.g. CSVpath = "C/Temp/Input.csv").
- Line 112 – SPF Model Form – This line allows the user to specify a different model form. Be sure to add statements under line 102 if any additional variables are added to the model. For instance, a variable for the natural log of traffic volume on the minor approach would need to be added if you were developing an intersection SPF.

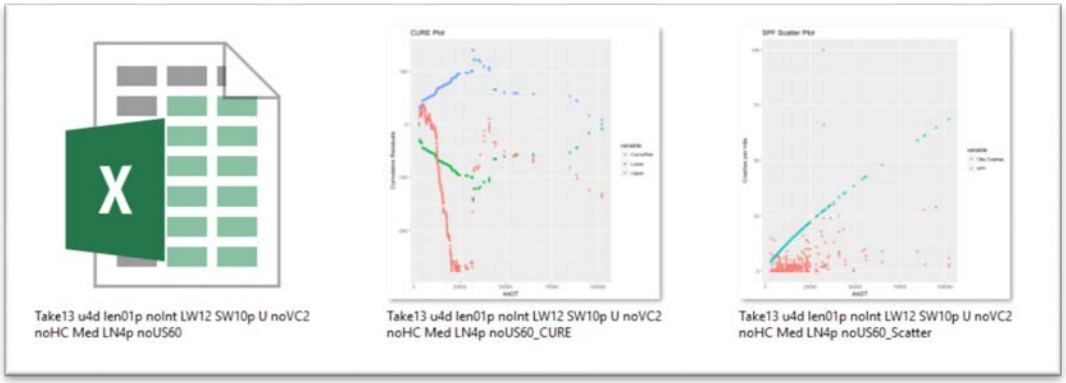
Generally, all other sections of the code should remain unchanged.

Once configured, a user simply executes the script using Code>Run Region>Run All (or using the hotkey Ctrl+Alt+R). The code includes several printed statements that will appear in the Console that can help with debugging. The following figure shows a typical R Studio layout.



SPF-R Output

After a successful execution, a folder called “R_SPFs” will be created in the designated output folder (a warning that this folder already exists will appear after the successive executions). In this folder, a project folder will be created containing three files: and Excel workbook with two worksheets, an image of a crash scatter plot, and an image of the CURE Plot. Windows Explorer provides an easy way to view the output quickly if the thumbnails are enlarged as shown below.



R Studio is able to process a large database with several classes (recall that classes are groups of roadway segments or intersections) resulting in several SPFs in just

a few minutes (on a modern computer at the time of writing this dissertation). In fact, typical SPF development takes only a few seconds.

Conclusions

This SPF development tool presented above is useful when trying to improve SPF development. The effect that the roadway network's heterogeneity has on SPF development can be quickly explored by simply adjusting the output folder (line 42) and the base filter (line 43). Consider the following example:

- Base condition #1
    - OutputProject_Base = "BC1-SW_2_LW_9"
    - myFilter = "data$SHLDWID == 2 & data$LANEWID == 9
- Base condition #2
    - OutputProject_Base = "BC1-SW_3_LW_10"
    - myFilter = "data$SHLDWID == 3 & data$LANEWID == 10

In the above example, two SPFs can quickly be developed for the same roadway network but for different specifications for shoulder and lane widths. Each SPF will be saved to separate folders, named accordingly. The CURE Plots can be compared and further assessment can be performed by opening the respective Excel files. Sample sizes and goodness-of-fit measures can be compared as well to decide which SPF is more appropriate for the dataset. The CURE plots provide a quick and visual screening process while other goodness-of-fit measures allow the user to objectively compare SPFs.

Resources

The following resources offer information on SPF development and calibration.

- The Highway Safety Manual, First Edition
- NCHRP Project 20-7 (Task 332): User's Guide to Develop Highway Safety Manual Safety Performance Function (SPF) Calibration Factors.
- SPF Decision Guide: SPF Calibration vs. SPF Development.
    - https://safety.fhwa.dot.gov/rsdp/downloads/spf_decision_guide_final.pdf
- SPF Development Guide: Developing Jurisdiction-Specific SPFs.
    - https://safety.fhwa.dot.gov/rsdp/downloads/spf_development_guide_final.pdf
- The Art of Regression Modeling in Road Safety by Ezra Hauer
    - http://www.springer.com/us/book/9783319125282

Appendix G – SPF-R RStudio Code

```
options(echo=FALSE)
#RCode for SPFs

#remove working envr
#rm(list = ls())
cat(rep("\n",64))

library(MASS)
library(knitr) #download using Packages>Install Packages or Tools>Install Packages
library(ggplot2) #download using Packages>Install Packages or Tools>Install Packages
library(openxlsx) #download using Packages>Install Packages or Tools>Install Packages
library(installr) #download using Packages>Install Packages or Tools>Install Packages
#ALSO Install this: https://cran.r-project.org/bin/windows/Rtools/
#IMPORTANT: You should make sure that the box is checked to have the installer edit
your PATH

#4/21/2017
VerNum <- c("RCode for SPFs ver 2.2.003")

#Directory
#Edit this in the event that your data is stored somewhere not relative to your home
USER folder
if (Sys.info()["nodename"] == "MERAK") {
  MyUserFolder = "D:/SyncFolders"
} else if(Sys.info()["nodename"] == "COE4249") {
  MyUserFolder = "C:/Users/clblac2.AD"
} else {
  MyUserFolder = dirname(Sys.getenv("HOME"))
}

#define data
TotalColumn = "Total" #The title of the column containing All Crashes. replace spaces
with "." e.g. "Total.Crashes"
KABCColumn = "" #The title of the column containing KABC Crashes. leave blank to ignore
KABColumn = "KAB" #The title of the column containing KAB Crashes. leave blank to ignore
KAColumn = "" #The title of the column containing KA Crashes. leave blank to ignore
KColumn = "" #The title of the column containing K Crashes. leave blank to ignore
AADTColumn = "LASTCNT" #The title of the column containing AADT.
LengthColumn = "Length" #The title of the column containing Length.
ClassColumn = "Class" #The title of the column containing a class category. leave blank
to ignore
ClassStart = 1 #Enter a positive integer. This will split the data into groups based on
the Class defined above
ClassEnd = 7 #Enter a positive integer. This will split the data into groups based on
the Class defined above
#Path to the input file as a CSV. Must contain headers as defined above.
CSVpath = paste0(MyUserFolder,"/Dropbox (Trans. Center)/~Annual
Projects/HSIP/2016/Cable/Crash Data/OverlayResultsWithClassesClean.csv")
OutputProject_Base = "Cable class Test" #Project name - created in Documents/R_SPFs
myFilter_Base = "data[[AADTColumn]] > 0  & data[[LengthColumn]] > 0" # A base filter.
Usually best to avoid Length or AADT of 0 or below.
InputData Base = "Median crossover 2011-2015" # be sure to uniquely describe the data
so it can be traced back to the source
initTheta = 1.0 #starting point for theta. adjust as needed
#########################################################################
#########################################################################
##### Be sure to edit the model form in the glm.nb  if different #####
#########################################################################
#########################################################################

#Examples
#CSVPath = "D:/SyncFolders/Dropbox (Trans. Center)/PhD/Phd -
Segmentation/Attribute/Case 1/SPFs/ParkwayTest/S4_1.csv"
#myFilter = "data[[AADTColumn]] > 0 & data[[LengthColumn]]"
#myFilter = "data[[AADTColumn]] < 500 & data$SHLDWID == 2 & data$LANEWID == 9 &
(data$CURVECLS == 'A' | data$CURVECLS == 'B') & data$MEDTYPE == 8 & (data$GRADECLS ==
```

```
    'A' | data$GRADECLS == '')"
56
57  #flag to test if data is bound
58  databind = FALSE
59
60  #read data
61  result <- tryCatch({
62    data=read.csv(CSVpath,header=T)
63    databind = TRUE
64  }, warning = function(war) {
65    # warning handler picks up where error was generated
66    print(paste("MY_WARNING:  ",war))
67    databind = FALSE
68  }, error = function(err) {
69    # error handler picks up where error was generated
70    print(paste("MY_ERROR:  ",err))
71    databind = FALSE
72  }, finally = {
73    # cleanup
74  })
75
76  #bind data
77  if (databind) {
78  result <- tryCatch({
79    exists("data")
80    databind = TRUE
81  }, warning = function(war) {
82    # warning handler picks up where error was generated
83    print(paste("MY_WARNING:  ",war))
84    databind = FALSE
85  }, error = function(err) {
86    # error handler picks up where error was generated
87    print(paste("MY_ERROR:  ",err))
88    databind = FALSE
89  }, finally = {
90    # cleanup
91  })
92  }
93
94  #Main SPF function
95  RunSPF <- function() {
96    #filter based on users' base filter
97    data_temp <- data[ which(eval(parse(text = myFilter))),]
98
99    #sort by AADT
100   data2 <- data_temp[ order(data_temp[[AADTColumn]]),]
101
102   #Point to variables
103   crash=data2[[CrashColumn]]
104   lnADT=log(data2[[AADTColumn]])
105   lnL=log(data2[[LengthColumn]])
106   #Calculate length if it doesn't exists - this will make zero length filter difficult
107   #lnL=log(EMP-BMP)
108
109   init.theta = initTheta
110
111 ###################################################################
112   SPF=glm.nb(crash~lnADT+offset(lnL))
113 ###################################################################
114
115   #add results from GLM
116   dataout <-
      cbind(data2,Predicted=SPF$fitted.values,Residuals=resid(SPF,type="resp"),CumulRes=cumsu
      m(resid(SPF,type="resp")))
117
118   #calculate data for CURE plot
```

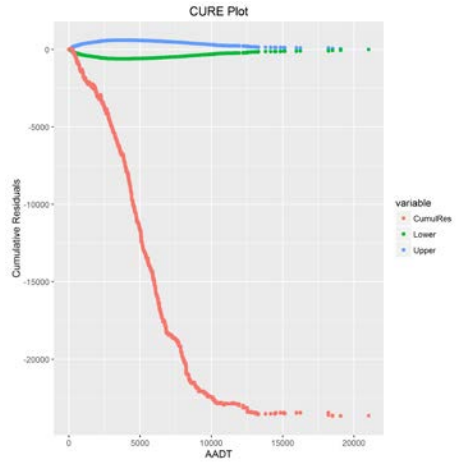
```
    datalimits <- data.frame(dataout$Residuals)
    datalimits["AADT"] <- NA
    datalimits$AADT <- data2[[AADTColumn]]
    datalimits["CumulRes"] <- NA
    datalimits$CumulRes <- dataout$CumulRes
    datalimits["Squared_Res"] <- NA
    datalimits$Squared_Res <- datalimits$dataout.Residuals^2
    datalimits["CumulSqRes"] <- NA
    datalimits$CumulSqRes <- cumsum(datalimits$Squared_Res)
    datalimits["SigmaSum"] <- NA
    datalimits$SigmaSum <- sqrt(datalimits$CumulSqRes)
    datalimits["StdDev"] <- NA
    datalimits$StdDev <-
    datalimits$SigmaSum*sqrt(1-datalimits$CumulSqRes/sum(datalimits$Squared_Res))
    datalimits["UpperLimit"] <- NA
    datalimits$UpperLimit <- datalimits$StdDev * 1.96
    datalimits["LowerLimit"] <- NA
    datalimits$LowerLimit <- datalimits$StdDev * (-1.96)
    datalimits["Per_CURE"] <- NA
    datalimits$Per_CURE <-
    ifelse(datalimits$CumulRes>datalimits$UpperLimit,1,ifelse(datalimits$CumulRes<datalimit
    s$LowerLimit,1,0))

    #create CURE plot
    CUREPlot <- ggplot(datalimits, aes(datalimits$AADT, y = value, color = variable)) +
      geom_point(aes(y = UpperLimit, col = "Upper")) +
      geom_point(aes(y = LowerLimit, col = "Lower")) +
      geom_point(aes(y = CumulRes, col = "CumulRes")) +
      ggtitle("CURE Plot") +
      labs(x="AADT",y="Cumulative Residuals")
    ggsave(file=paste0(OutPath,OutputProject,"_CURE.png"))

    #Scatter Plot with SPF
    ScatterPlot <- ggplot(dataout, aes(dataout[[AADTColumn]], y = value, color =
    variable)) +
      geom_point(aes(y = dataout[[CrashColumn]] / dataout[[LengthColumn]], col = "Obs
      Crashes")) +
      geom_point(aes(y = dataout$Predicted / dataout[[LengthColumn]], col = "SPF")) +
      ggtitle("SPF Scatter Plot") +
      labs(x="AADT",y="Crashes per mile")
    ggsave(file=paste0(OutPath,OutputProject,"_Scatter.png"))

    #Metrics/Stats
    Sample = nrow(dataout)
    Mileage = sum(dataout[[LengthColumn]])
    Crashes = sum(dataout[[CrashColumn]])
    ObsAvg = mean(dataout[[CrashColumn]])
    tmpTerm = sum((dataout[[CrashColumn]]-ObsAvg)^2)
    tmpTerm2 = sum((dataout[[CrashColumn]]-dataout$Predicted)^2)
    RSquared = (tmpTerm-tmpTerm2)/(tmpTerm-sum(dataout$Predicted))
    CDP = sum(datalimits$Per_CURE)/length(datalimits$Per_CURE)*100
    MACD = max(abs(datalimits$CumulRes))
    MAD = mean(abs(dataout$Residuals))
    datametrics <- data.frame(Values =
    c(Sample,Mileage,Crashes,RSquared,CDP,MACD,MAD,SPF$theta,coef(summary(SPF))["(Intercept
    )","Estimate"],coef(summary(SPF))["lnADT","Estimate"], SPF$SE.theta, SPF$aic, "", "",
    ""))
    datametrics$Notes <- c("100-200 intersections*","100-200 miles*","300 crashes per
    year*","Higher values preferred","Less than 5%","Smaller values preferred","Smaller
    values preferred","Higher values preferred","(Intercept)","(lnAADT)", "", "",
    myFilter, InputData,"*As recommended by FHWA-SA-14-004")
    attr(datametrics, "row.names") <-
    c("Sample","Length","Crashes","R2","CDP","MACD","MAD","Theta","Alpha","Beta","StdErr","
    AIC", "Filter","Input Data","")
    datametrics$Values = as.numeric(as.character(datametrics$Values))

```

```
  #PCR (potential for crash reduction)
  # NOTE: the weight equation is based on a 5-year period. That is, the number of crashes in the input file is for
  #a 5-year period therefore year is not in the equation!
  dataout["Weight"] <- NA
  dataout$Weight <- 1/(1+dataout$Predicted/dataout[[LengthColumn]]/SPF$theta)
  dataout["EB_Estimate"] <- NA
  dataout$EB_Estimate <- dataout[[CrashColumn]]*(1-dataout$Weight) + dataout$Predicted*(dataout$Weight)
  dataout["PCR"] <- NA
  dataout$PCR <- dataout$EB_Estimate - dataout$Predicted

  #save results to Excel
  wb <- createWorkbook()
  options("openxlsx.borderStyle" = "thin")
  options("openxlsx.borderColour" = "#4F81BD")
  addWorksheet(wb, "Metrics")
  addWorksheet(wb, "Data")
  writeData(wb, "Metrics", datametrics, startCol = 2, startRow = 3, rowNames = TRUE)
  writeData(wb, "Metrics", VerNum, startCol = 1, startRow = 1)
  writeData(wb, "Metrics", CSVpath, startCol = 1, startRow = 2)
  writeData(wb, "Data", dataout)
  saveWorkbook(wb, paste0(OutPath,OutputProject,".xlsx"), overwrite = TRUE)
}

#Check if input data is valid
if (databind) {

  if (ClassColumn == "") {
    # this will disable the loop for classes
    ClassStart=0
    ClassEnd=0
    }

  for(i in ClassStart:ClassEnd) {

      # add a filter and change output path for classes if needed
      if (ClassColumn == "") {
        myFilter = myFilter_Base
        ClassOut = ""
      } else {
        myFilter = paste0(myFilter_Base," & data[[ClassColumn]] == ",i)
        ClassOut = paste0(" - Class ",i)
      }

      # All crashes
      CrashColumn = TotalColumn
      InputData = paste0(ClassOut,InputData_Base)
      OutputProject = paste0(OutputProject_Base,ClassOut)
      #create folders
      dir.create(file.path(Sys.getenv("HOME"), "R_SPFs"))
      dir.create(file.path(paste0(Sys.getenv("HOME"),"/R_SPFs"),OutputProject))
      OutPath = paste0(Sys.getenv("HOME"),"/R_SPFs/",OutputProject,"/")

      RunSPF()
      print(paste0("All crashes finished",ClassOut))

      if (KABCColumn != "") {
        #KABC
        CrashColumn = KABCColumn
        InputData = paste0(InputData_Base," - KABC",ClassOut)
        OutputProject = paste0(OutputProject_Base," - KABC",ClassOut)
        #create folders
        dir.create(file.path(paste0(Sys.getenv("HOME"),"/R_SPFs"), OutputProject))
        OutPath = paste0(Sys.getenv("HOME"),"/R_SPFs/",OutputProject,"/")
        RunSPF()
```

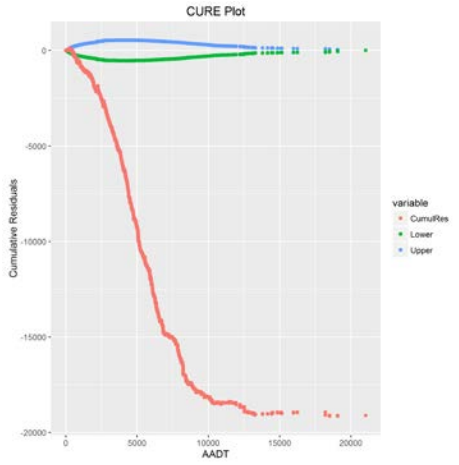
```
        print(paste0("KABC crashes finished",ClassOut))
      }

      if (KABColumn != "") {
        #KAB
        CrashColumn = KABColumn
        InputData = paste0(InputData_Base," - KAB",ClassOut)
        OutputProject = paste0(OutputProject_Base," - KAB",ClassOut)
        #create folders
        dir.create(file.path(paste0(Sys.getenv("HOME"),"/R SPFs"), OutputProject))
        OutPath = paste0(Sys.getenv("HOME"),"/R_SPFs/",OutputProject,"/")
        RunSPF()
        print(paste0("KAB crashes finished",ClassOut))
      }

      if (KAColumn != "") {
        #KA
        CrashColumn = KAColumn
        InputData = paste0(InputData_Base," - KA",ClassOut)
        OutputProject = paste0(OutputProject_Base," - KA",ClassOut)
        #create folders
        dir.create(file.path(paste0(Sys.getenv("HOME"),"/R SPFs"), OutputProject))
        OutPath = paste0(Sys.getenv("HOME"),"/R_SPFs/",OutputProject,"/")
        RunSPF()
        print(paste0("KA crashes finished",ClassOut))
      }
      if (KColumn != "") {
        #K
        CrashColumn = KColumn
        InputData = paste0(InputData_Base," - K",ClassOut)
        OutputProject = paste0(OutputProject_Base," - K",ClassOut)
        #create folders
        dir.create(file.path(paste0(Sys.getenv("HOME"),"/R_SPFs"), OutputProject))
        OutPath = paste0(Sys.getenv("HOME"),"/R SPFs/",OutputProject,"/")
        RunSPF()
        print(paste0("K crashes finished",ClassOut))
      }

    }
    print("finished")

  } else {

    print("Check for error.")

  }
```

Appendix H – CURE Plots with Increasing Homogeneity

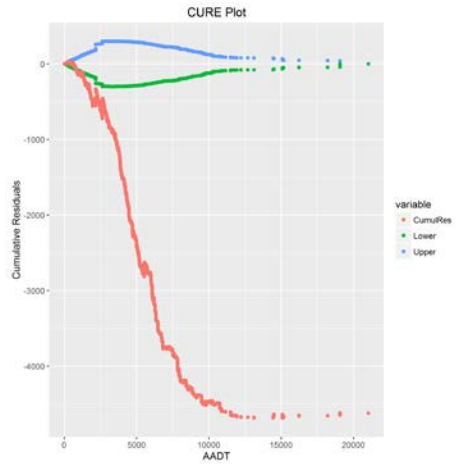
1



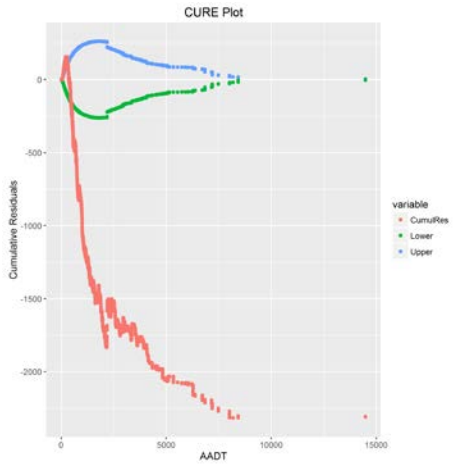
2



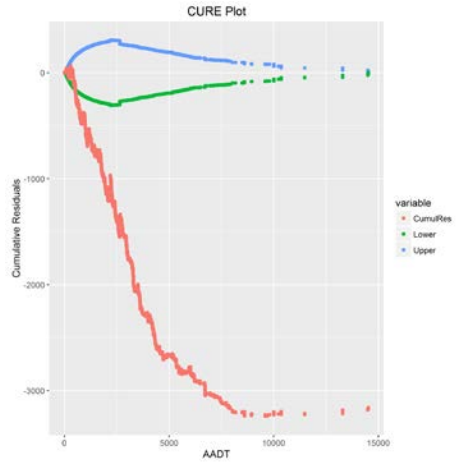
3



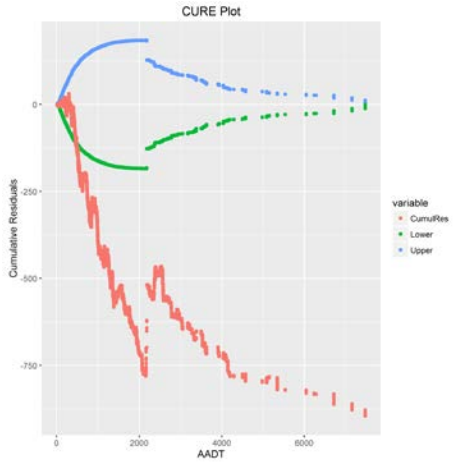
4



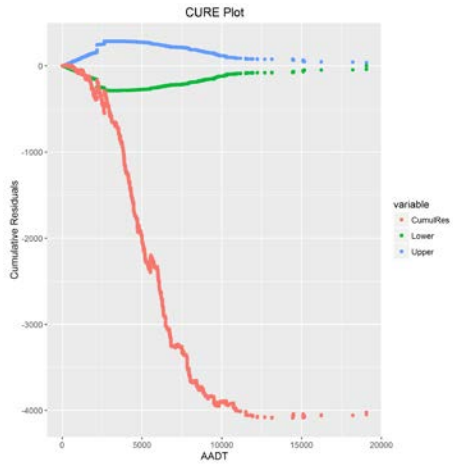
5



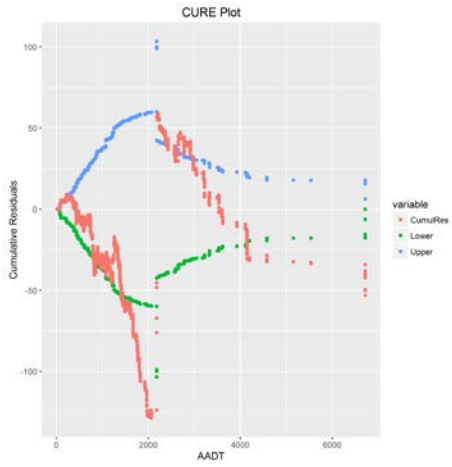
6



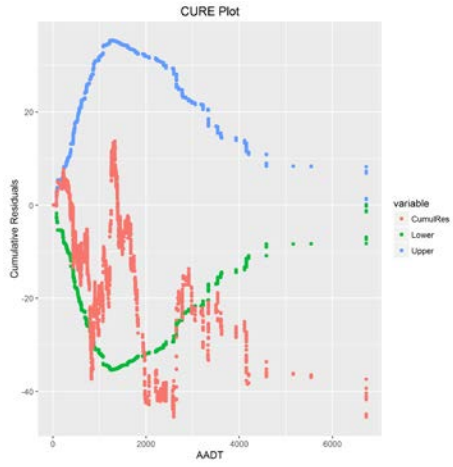
7


CURE Plot
Cumulative Residuals
AADT
variable
CumulRes
Lower
Upper

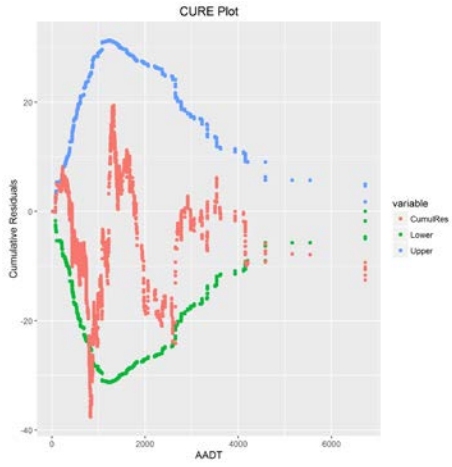
8


CURE Plot
Cumulative Residuals
AADT
variable
CumulRes
Lower
Upper

9

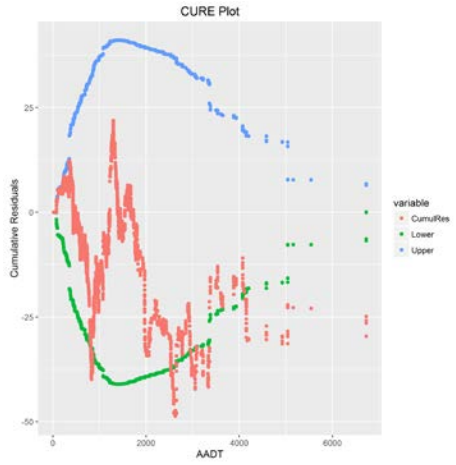

CURE Plot
Cumulative Residuals
AADT
variable
CumulRes
Lower
Upper

10

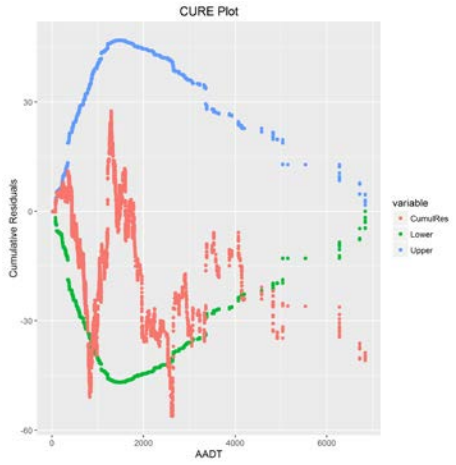

CURE Plot
Cumulative Residuals
AADT
variable
CumulRes
Lower
Upper

Appendix I – CURE Plots with Increasing Homogeneity with Ranges

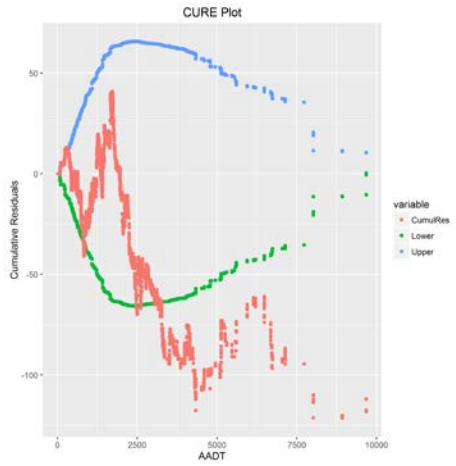
11



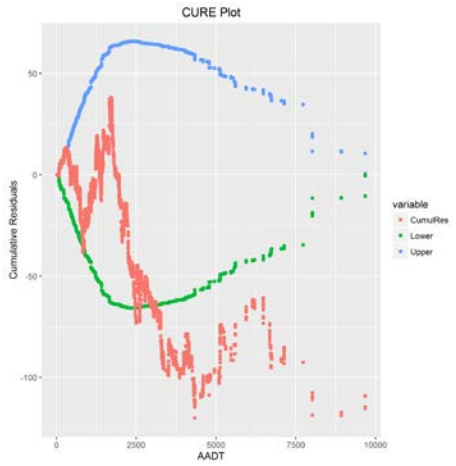
12



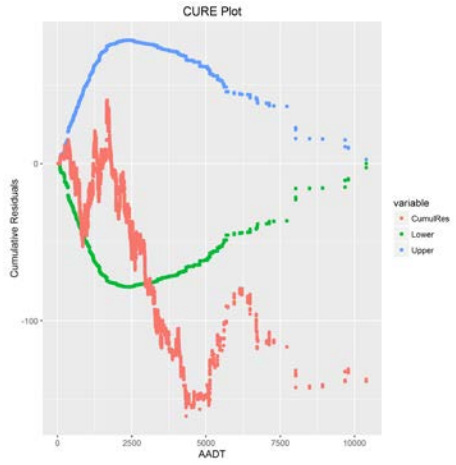
13



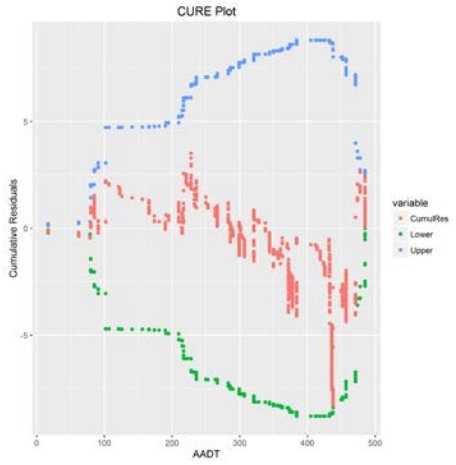
14



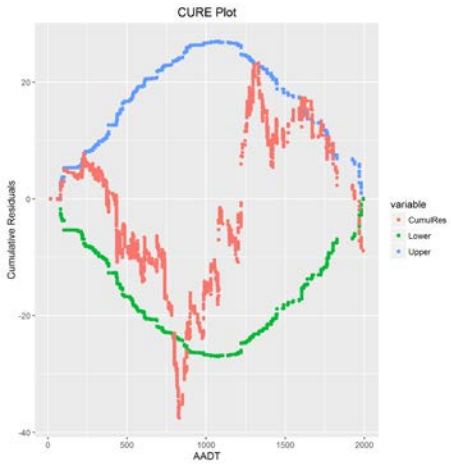
15



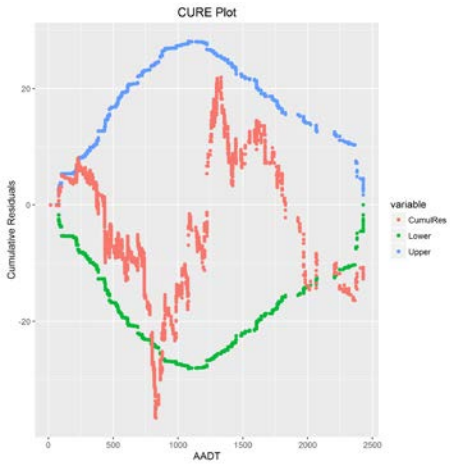
16



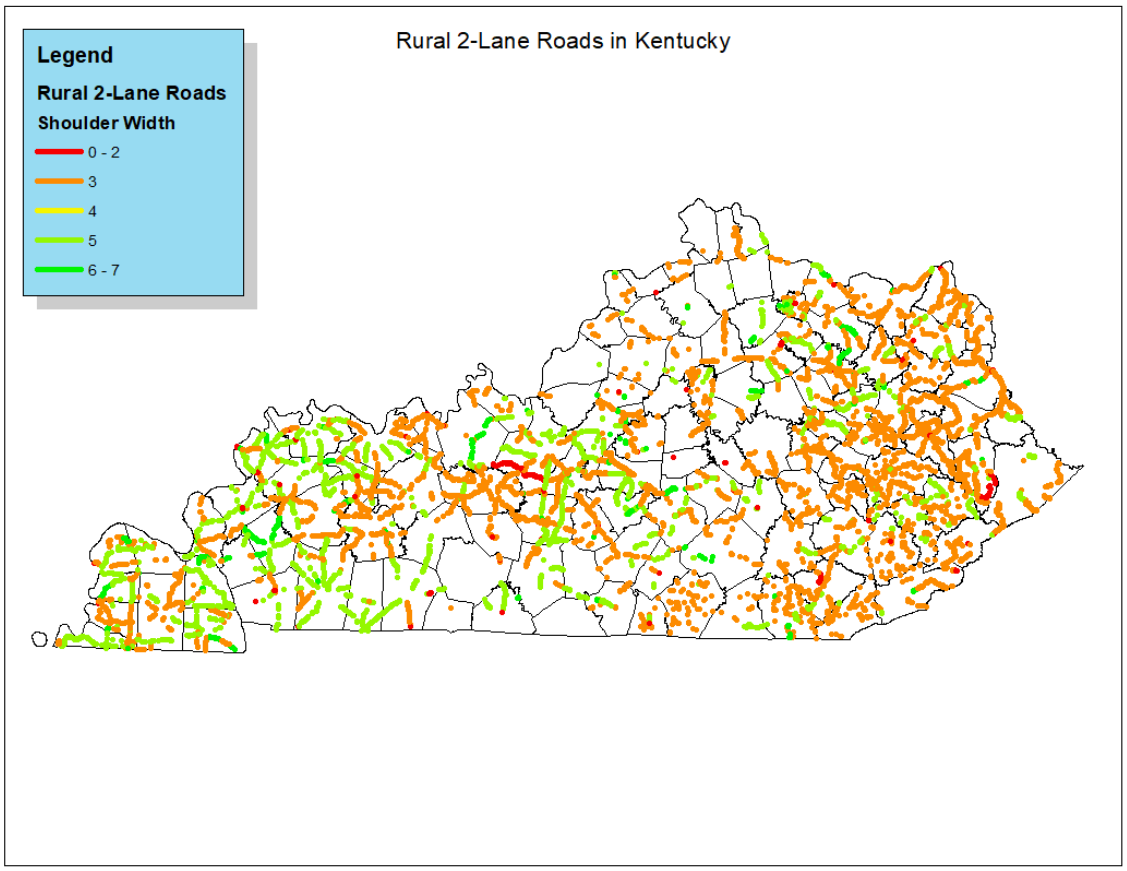
17


CURE Plot
Cumulative Residuals
AADT
variable
CumulRes
Lower
Upper

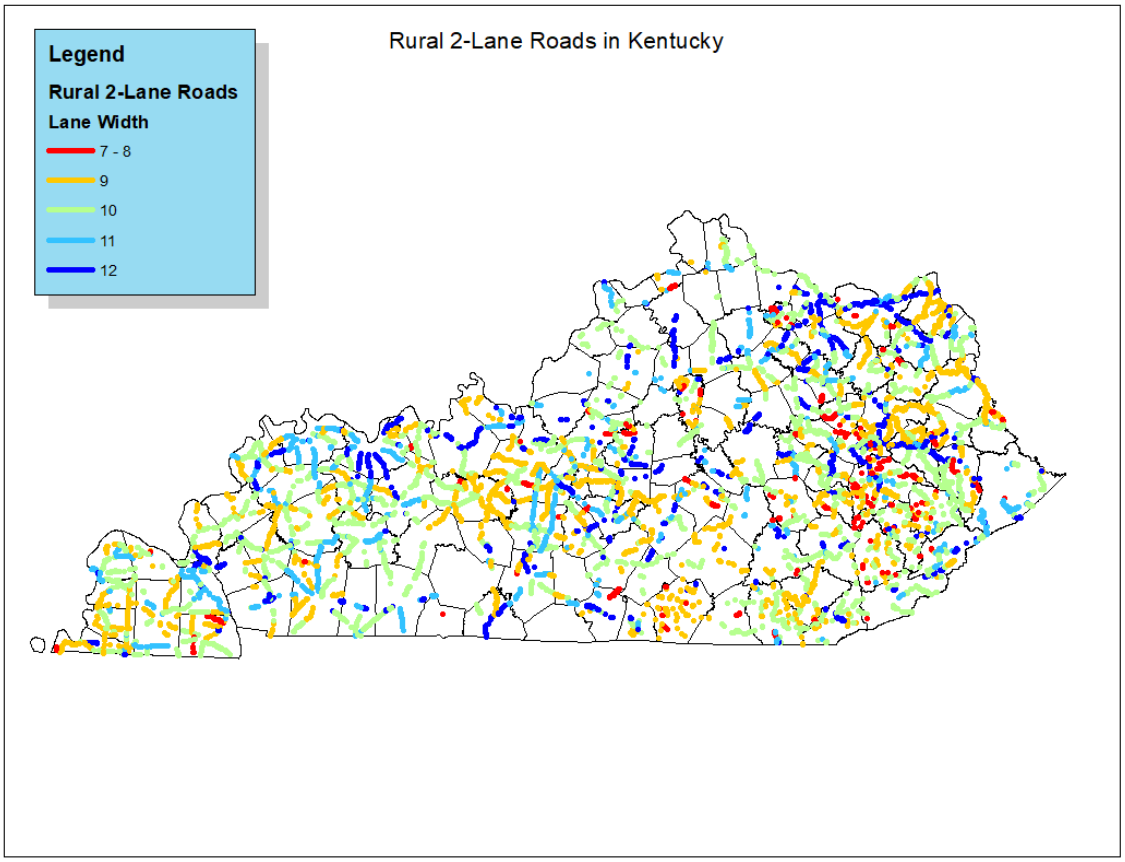
18


CURE Plot
Cumulative Residuals
AADT
variable
CumulRes
Lower
Upper

Appendix J – Map of Rural 2-Lane by Shoulder Widths

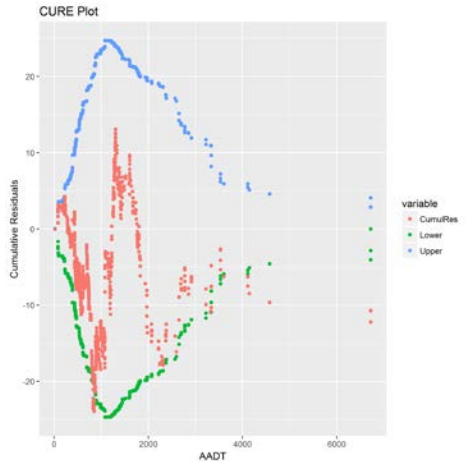


Appendix K – Map of Rural 2-Lane by Lane Widths

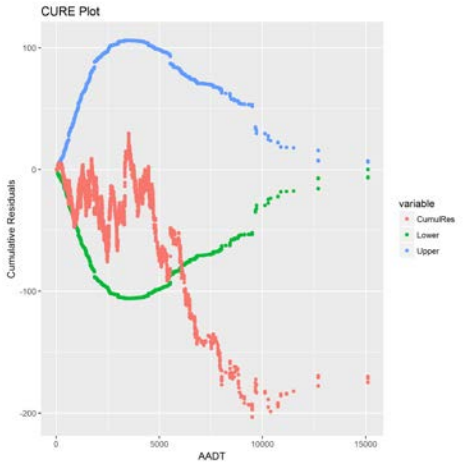


Appendix L – CURE Plots For Comparing Models

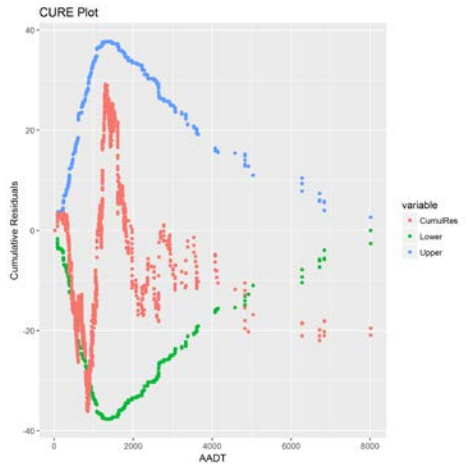
Base



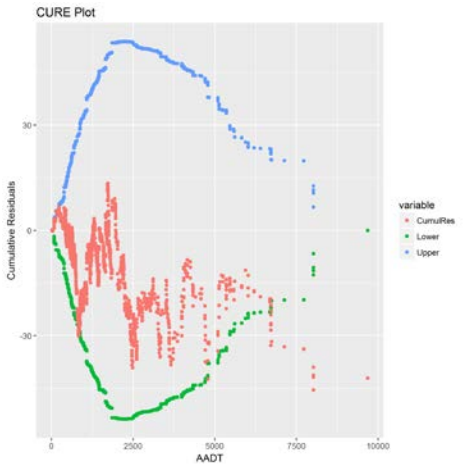
Model 1



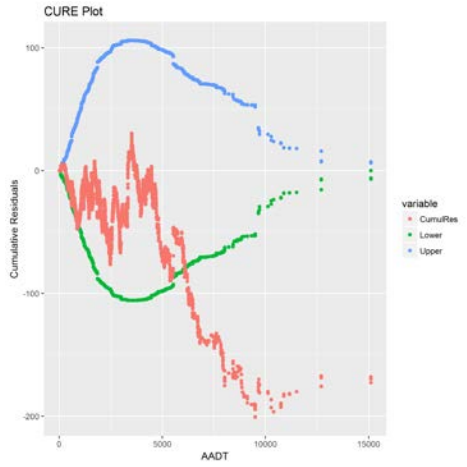
Model 2



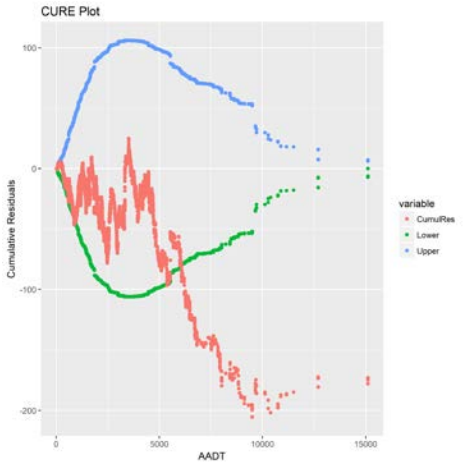
Model 3



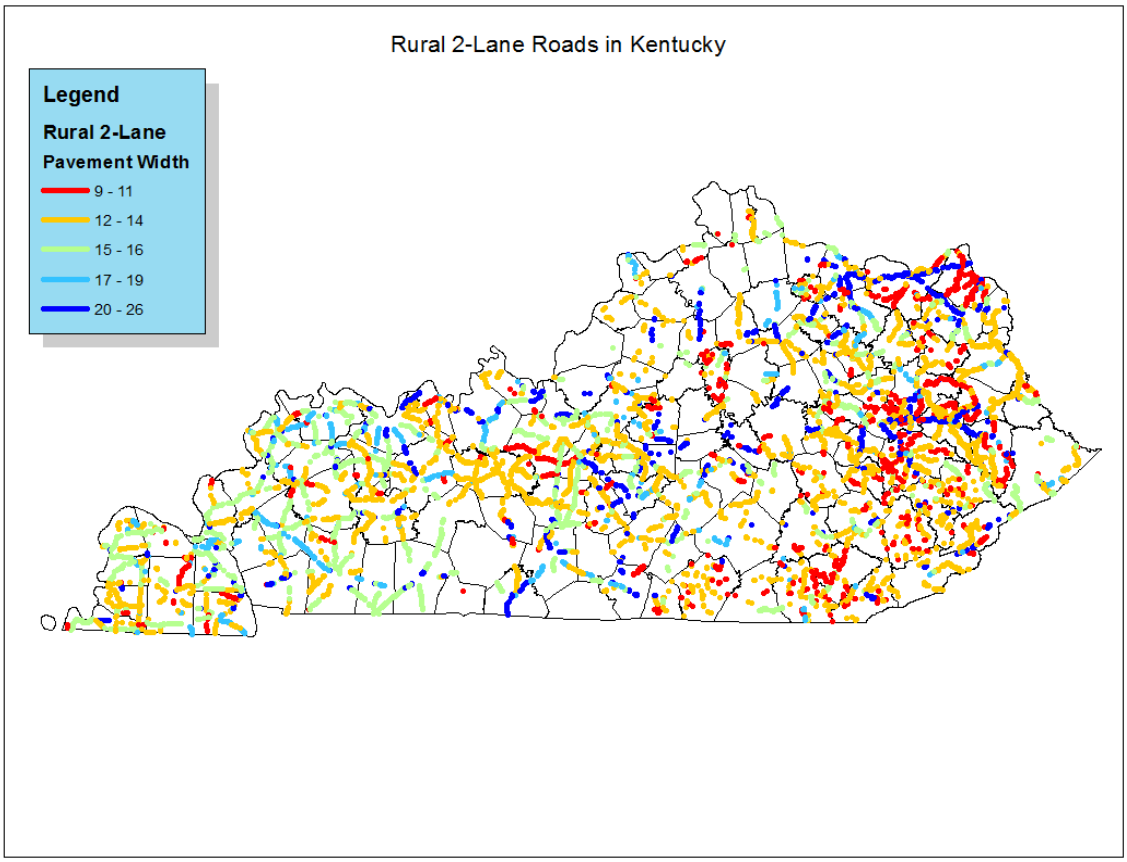
Model 4



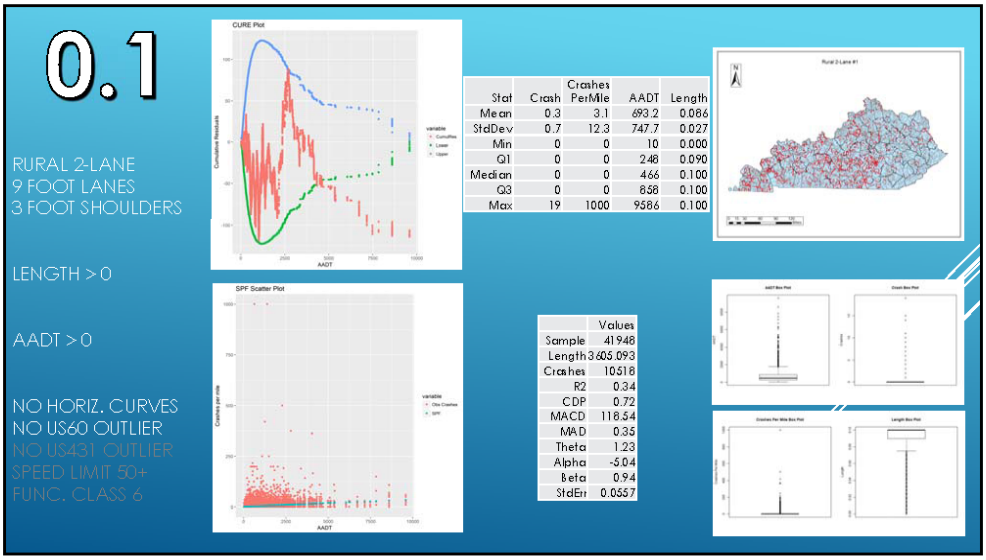
Model 5

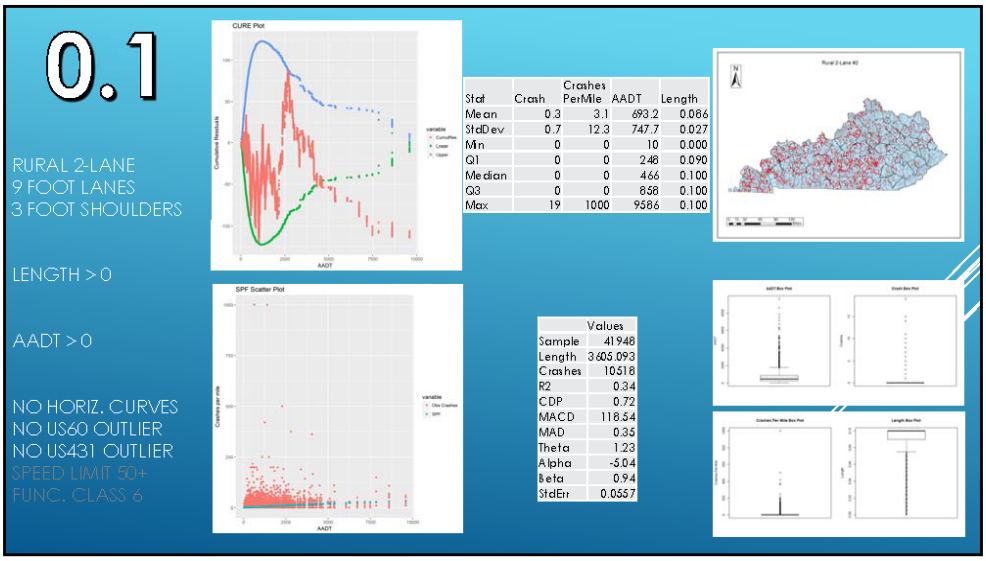


Appendix M– Map of Rural 2-Lane by Roadway Widths



Appendix N – Visualization Comparing Changes in Length and Attributes





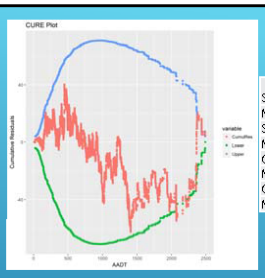
0.1

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

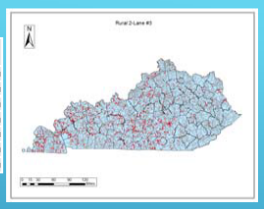
LENGTH >= 0.1

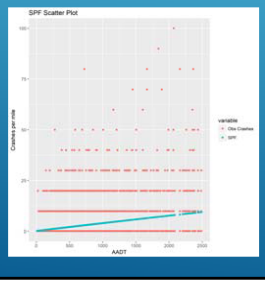
AADT < 2500

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

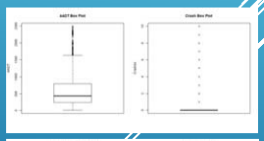


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.2	2.4	593.5	0.100
StdDev	0.6	6.0	494.9	0.000
Min	0	0	10	0.100
Q1	0	0	240	0.100
Median	0	0	435	0.100
Q3	0	0	796	0.100
Max	10	100	2489	0.100





	Values
Sample	16388
Length	1638.8
Crashes	3898
R2	0.32
CDP	0.93
MACD	62.37
MAD	0.34
Theta	1.37
Alpha	-5.27
Beta	0.96
StdErr	0.1121



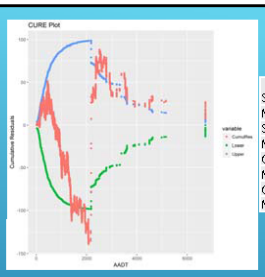
0.1

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

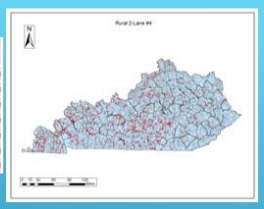
LENGTH = 0.1

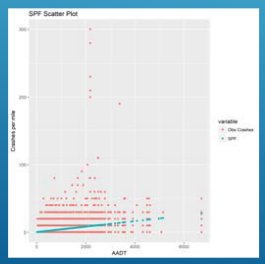
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

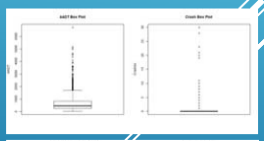


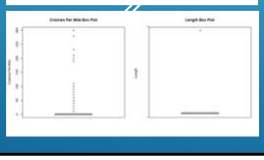
Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.3	2.7	665.6	0.100
StdDev	0.8	8.2	664.6	0.000
Min	0	0	10	0.100
Q1	0	0	244	0.100
Median	0	0	448	0.100
Q3	0	0	835	0.100
Max	30	300	6720	0.100





	Values
Sample	16856
Length	1685.6
Crashes	4614
R2	0.18
CDP	7.27
MACD	138.98
MAD	0.37
Theta	1.08
Alpha	-5.48
Beta	1.00
StdErr	0.0639





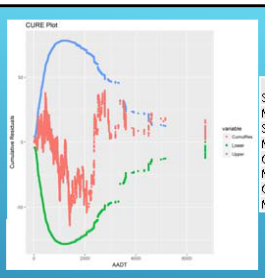
0.1

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

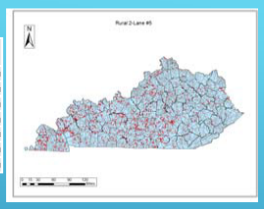
LENGTH = 0.1

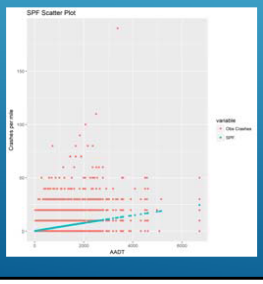
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

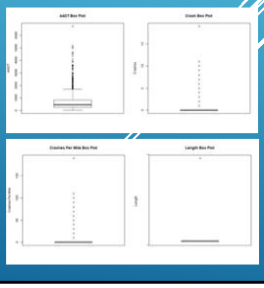


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.3	2.6	664.8	0.100
StdDev	0.7	6.6	663.9	0.000
Min	0	0	10	0.100
Q1	0	0	243.75	0.100
Median	0	0	448	0.100
Q3	0	0	835	0.100
Max	19	190	6720	0.100





	Values
Sample	16848
Length	1684.8
Crashes	4448
R2	0.35
CDP	0.43
MACD	63.87
MAD	0.36
Theta	1.44
Alpha	-5.24
Beta	0.96
StdErr	0.1085



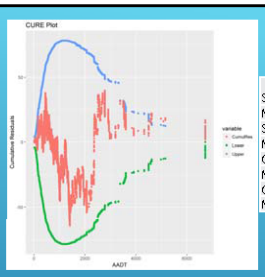
0.1

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

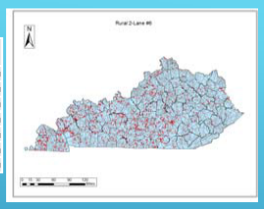
LENGTH = 0.1

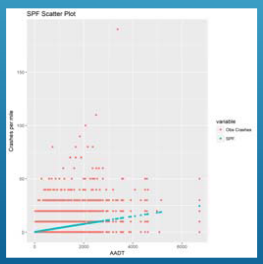
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

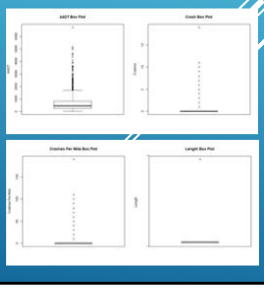


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.3	2.6	664.8	0.100
StdDev	0.7	6.6	663.9	0.000
Min	0	0	10	0.100
Q1	0	0	243.75	0.100
Median	0	0	448	0.100
Q3	0	0	835	0.100
Max	19	190	6720	0.100





	Values
Sample	16848
Length	1684.8
Crashes	4448
R2	0.35
CDP	0.43
MACD	63.87
MAD	0.36
Theta	1.44
Alpha	-5.24
Beta	0.96
StdErr	0.1085



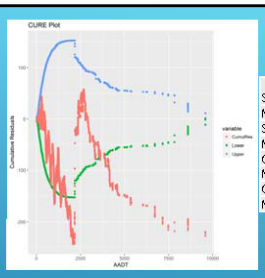
0.2

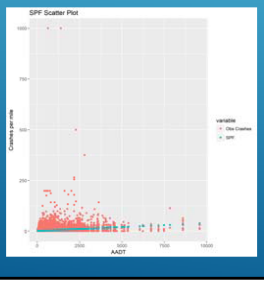
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH > 0

AADT > 0

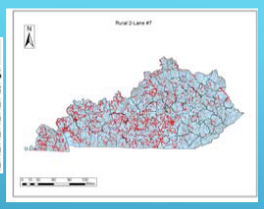
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

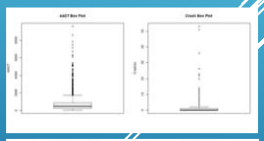


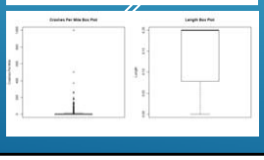


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.4	3.3	679.5	0.146
StdDev	1.1	13.5	738.5	0.068
Min	0	0	10	0.000
Q1	0	0	238	0.079
Median	0	0	450	0.200
Q3	1	5	838	0.200
Max	53	1000	9586	0.200

	Values
Sample	24693
Length	3605.893
Crashes	10684
R2	0.29
CDP	6.86
MACD	243.34
MAD	0.49
Theta	1.42
Alpha	-5.18
Beta	0.96
StdErr	0.0599







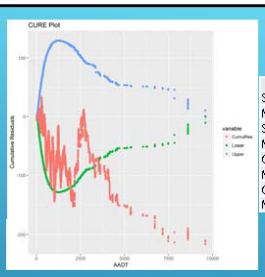
0.2

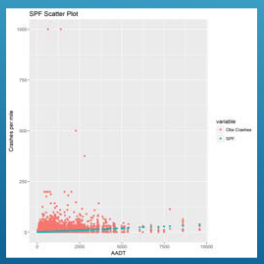
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH > 0

AADT > 0

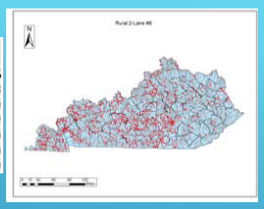
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

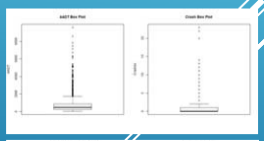


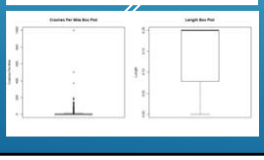


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.4	3.2	679.3	0.146
StdDev	1.0	13.3	738.3	0.068
Min	0	0	10	0.000
Q1	0	0	238	0.079
Median	0	0	450	0.200
Q3	1	5	838	0.200
Max	23	1000	9586	0.200

	Values
Sample	24689
Length	3605.093
Crashes	10518
R2	0.44
CDP	3.90
MACD	216.46
MAD	0.49
Theta	1.59
Alpha	-5.08
Beta	0.94
StdErr	0.0734







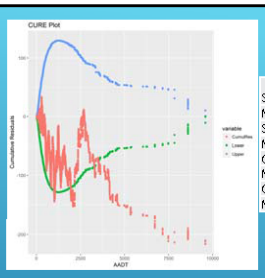
0.2

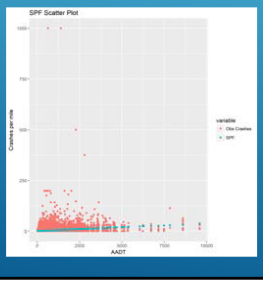
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH > 0

AADT > 0

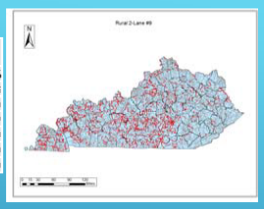
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

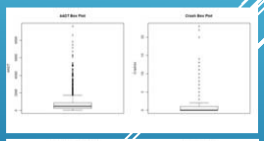


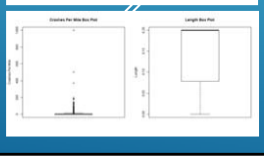


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.4	3.2	679.3	0.146
StdDev	1.0	13.3	738.3	0.068
Min	0	0	10	0.000
Q1	0	0	238	0.079
Median	0	0	450	0.200
Q3	1	5	838	0.200
Max	23	1000	9586	0.200

	Values
Sample	24689
Length	3605.093
Crashes	10518
R2	0.44
CDP	3.90
MACD	216.46
MAD	0.49
Theta	1.59
Alpha	-5.08
Beta	0.94
StdErr	0.0734







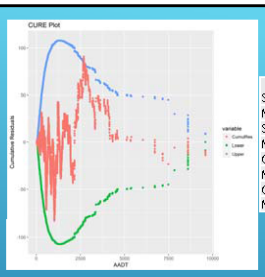
0.2

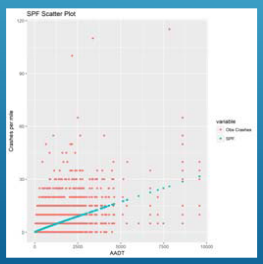
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH = 0.2

AADT > 0

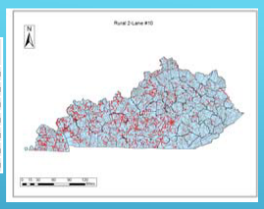
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

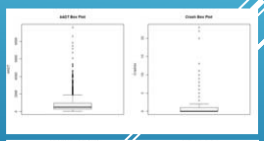


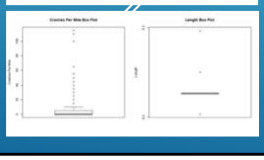


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.6	2.8	739.8	0.200
StdDev	1.1	5.4	783.2	0.000
Min	0	0	13	0.200
Q1	0	0	278	0.200
Median	0	0	506	0.200
Q3	1	5	930	0.200
Max	23	115	9586	0.200

	Values
Sample	13274
Length	2654.8
Crashes	7488
R2	0.45
CDP	1.16
MACD	90.72
MAD	0.60
Theta	2.06
Alpha	-5.10
Beta	0.93
StdErr	0.1251







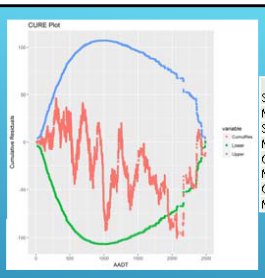
0.2

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

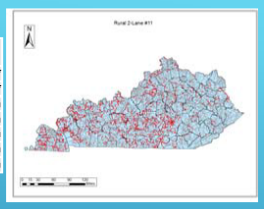
LENGTH >= 0.1

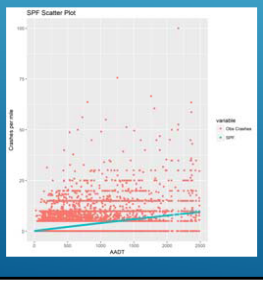
AADT < 2500

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

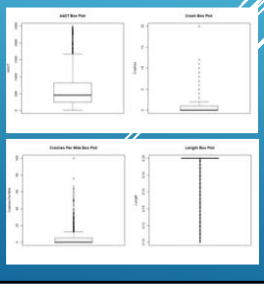


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.5	2.5	610.0	0.187
StdDev	0.9	5.0	488.1	0.027
Min	0	0	10	0.100
Q1	0	0	254	0.200
Median	0	0	459	0.200
Q3	1	5	820	0.200
Max	20	100	2489	0.200





	Values
Sample	16650
Length	3118.783
Crashes	7768
R2	0.35
CDP	1.58
MACD	100.28
MAD	0.55
Theta	1.67
Alpha	-5.19
Beta	0.95
StdErr	0.0919



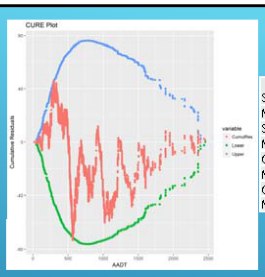
0.2

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

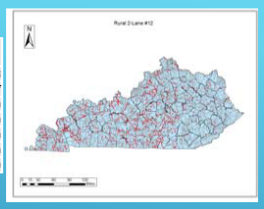
LENGTH <= 0.1

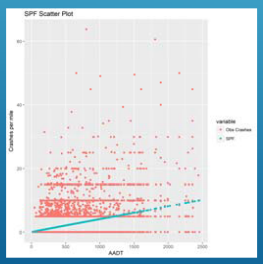
AADT < 2500

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

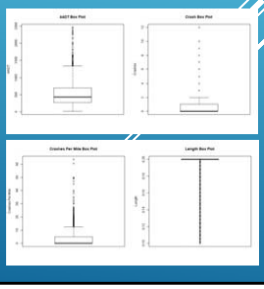


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.4	2.3	547.9	0.188
StdDev	0.9	4.6	396.4	0.027
Min	0	0	17	0.100
Q1	0	0	275	0.200
Median	0	0	437	0.200
Q3	1	5	697	0.200
Max	12	63.69427	2456	0.200





	Values
Sample	9628
Length	1805.69
Crashes	4153
R2	0.34
CDP	0.44
MACD	73.21
MAD	0.53
Theta	1.62
Alpha	-5.34
Beta	0.98
StdErr	0.1242



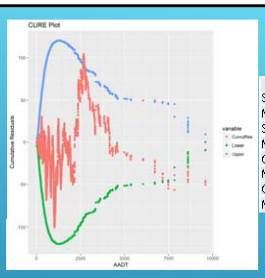
0.2

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

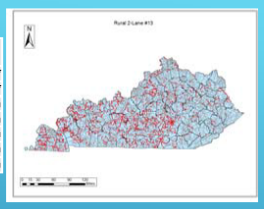
LENGTH >= 0.1

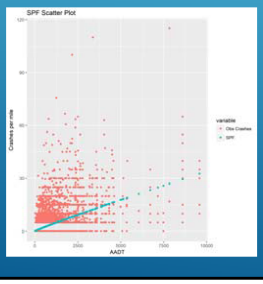
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

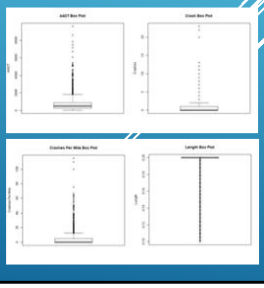


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.5	2.8	713.1	0.187
StdDev	1.1	5.7	760.5	0.027
Min	0	0	10	0.100
Q1	0	0	262	0.200
Median	0	0	484	0.200
Q3	1	5	892	0.200
Max	23	115	9586	0.200





	Values
Sample	17261
Length	3235.697
Crashes	9162
R2	0.43
CDP	0.74
MACD	104.33
MAD	0.58
Theta	1.81
Alpha	-5.06
Beta	0.93
StdErr	0.0932



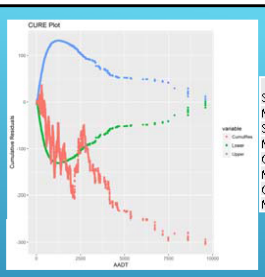
0.3

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

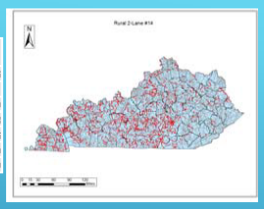
LENGTH > 0

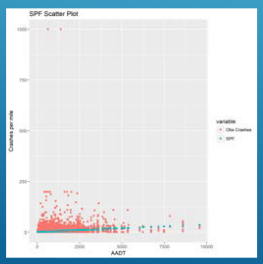
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

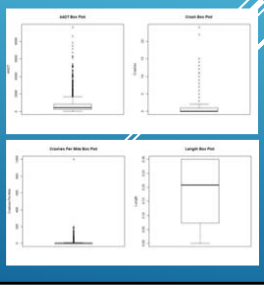


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.5	3.3	666.0	0.188
StdDev	1.2	13.9	727.5	0.111
Min	0	0	10	0.000
Q1	0	0	232	0.073
Median	0	0	441	0.208
Q3	1	3.333333	816	0.300
Max	24	1000	9586	0.300





	Values
Sample	19187
Length	3605.093
Crashes	10518
R2	0.52
CDP	9.59
MACD	304.35
MAD	0.56
Theta	1.85
Alpha	-5.11
Beta	0.95
StdErr	0.0895



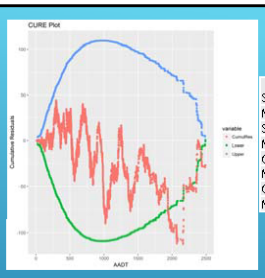
0.3

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

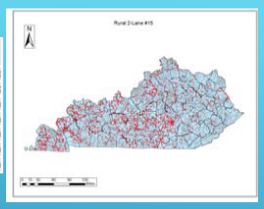
LENGTH >= 0.1

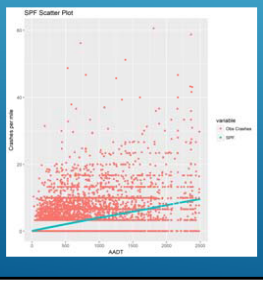
AADT < 2500

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

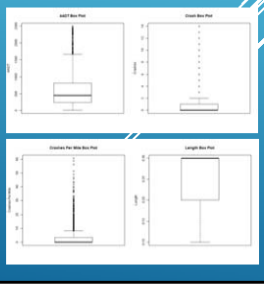


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.6	2.5	602.8	0.253
StdDev	1.1	4.5	485.1	0.068
Min	0	0	10	0.100
Q1	0	0	245	0.200
Median	0	0	454	0.300
Q3	1	3.333333	811	0.300
Max	14	60.60606	2489	0.300





	Values
Sample	12546
Length	3171.051
Crashes	7895
R2	0.43
CDP	2.12
MACD	112.46
MAD	0.65
Theta	2.00
Alpha	-5.17
Beta	0.95
StdErr	0.1142



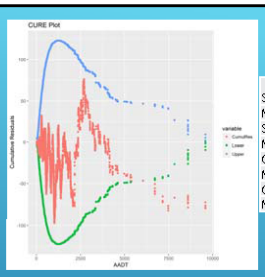
0.3

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

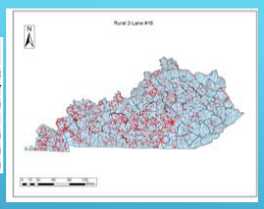
LENGTH >= 0.1

AADT > 0

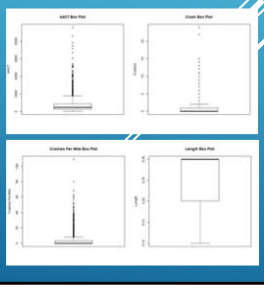
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6



Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.7	2.8	703.1	0.253
StdDev	1.3	5.3	753.0	0.067
Min	0	0	10	0.100
Q1	0	0	256	0.201
Median	0	0	474	0.300
Q3	1	3.333333	865	0.300
Max	24	108.9109	9586	0.300



	Values
Sample	12992
Length	3289.874
Crashes	9334
R2	0.52
CDP	0.61
MACD	97.12
MAD	0.70
Theta	2.17
Alpha	-5.08
Beta	0.94
StdErr	0.1176



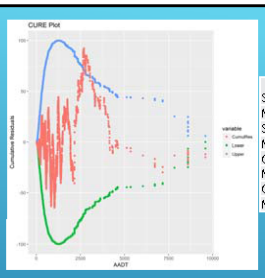
0.3

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

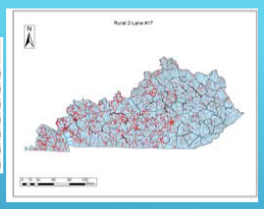
LENGTH = 0.3

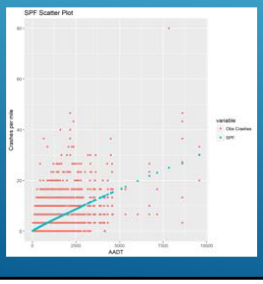
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

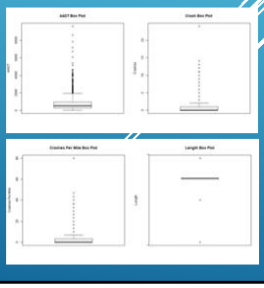


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.8	2.8	750.1	0.300
StdDev	1.4	4.6	790.3	0.000
Min	0	0	13	0.300
Q1	0	0	282	0.300
Median	0	0	518	0.300
Q3	1	3.333333	953	0.300
Max	24	80	9586	0.300





	Values
Sample	7762
Length	2328.6
Crashes	6553
R2	0.55
CDP	2.60
MACD	92.56
MAD	0.77
Theta	2.64
Alpha	-5.02
Beta	0.92
StdErr	0.1872



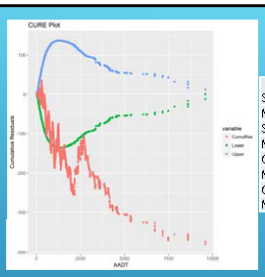
0.4

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

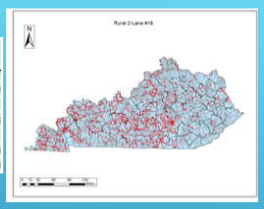
LENGTH > 0

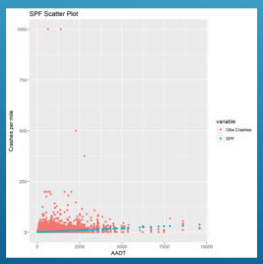
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

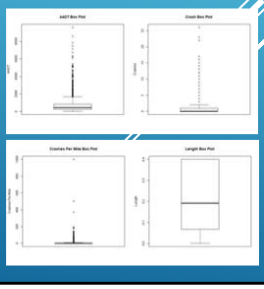


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.6	3.4	656.5	0.217
StdDev	1.3	15.1	717.8	0.150
Min	0	0	10	0.001
Q1	0	0	227	0.068
Median	0	0	436	0.191
Q3	1	2.5	809	0.400
Max	26	1000	9586	0.400





	Values
Sample	16597
Length	3605.093
Crashes	10518
R2	0.56
CDP	17.49
MACD	379.43
MAD	0.61
Theta	1.93
Alpha	-5.13
Beta	0.95
StdErr	0.0941



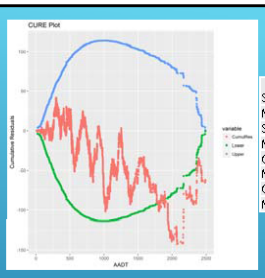
0.4

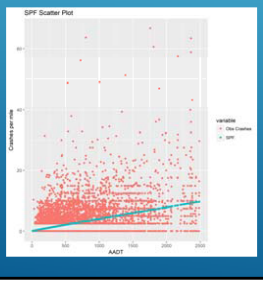
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

AADT < 2500

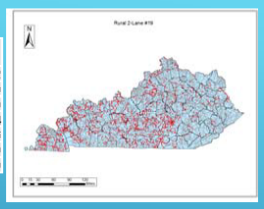
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

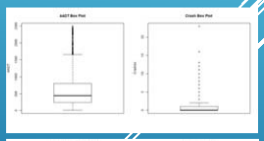


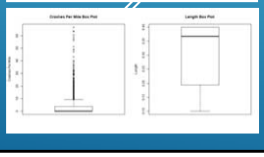


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.8	2.5	597.2	0.302
StdDev	1.3	4.6	483.8	0.110
Min	0	0	10	0.100
Q1	0	0	243	0.194
Median	0	0	446	0.368
Q3	1	3.745318	807	0.400
Max	23	66.66667	2489	0.400

	Values
Sample	10573
Length	3195.557
Crashes	7992
R2	0.47
CDP	3.55
MACD	142.48
MAD	0.72
Theta	2.08
Alpha	-5.20
Beta	0.96
StdErr	0.1178







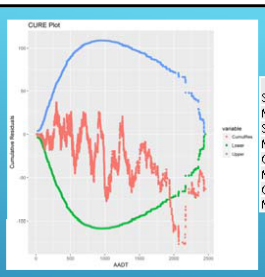
0.4

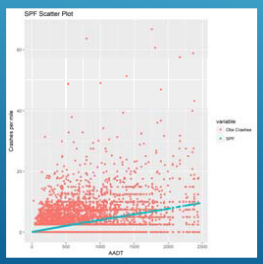
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

AADT < 2500

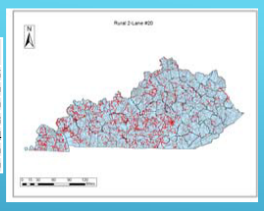
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

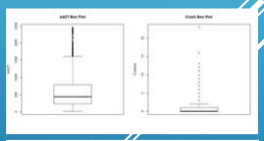


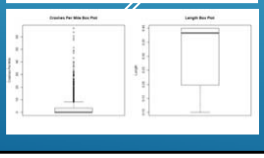


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.7	2.4	575.6	0.305
StdDev	1.3	4.4	465.2	0.110
Min	0	0	10	0.100
Q1	0	0	234	0.198
Median	0	0	434	0.384
Q3	1	3.297616	785	0.400
Max	23	66.66667	2456	0.400

	Values
Sample	9998
Length	3051.957
Crashes	7323
R2	0.46
CDP	2.33
MACD	126.62
MAD	0.71
Theta	2.05
Alpha	-5.19
Beta	0.95
StdErr	0.1209







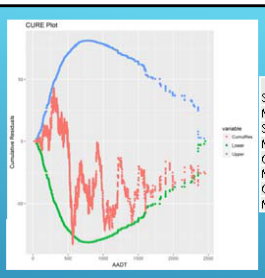
0.4

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

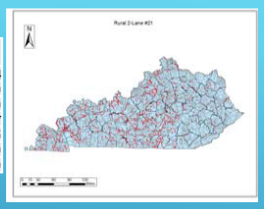
LENGTH >= 0.1

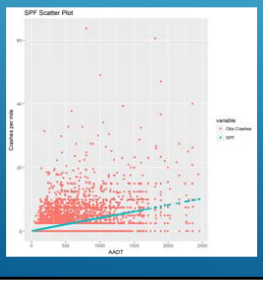
AADT < 2500

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

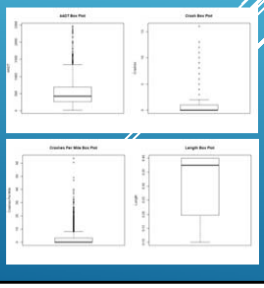


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.7	2.3	540.0	0.304
StdDev	1.2	4.2	393.6	0.110
Min	0	0	17	0.100
Q1	0	0	265	0.197
Median	0	0	429	0.375
Q3	1	3.236246	695	0.400
Max	16	63.69427	2456	0.400





	Values
Sample	6081
Length	1849.668
Crashes	4274
R2	0.45
CDP	1.36
MACD	82.36
MAD	0.70
Theta	1.99
Alpha	-5.32
Beta	0.98
StdErr	0.1542



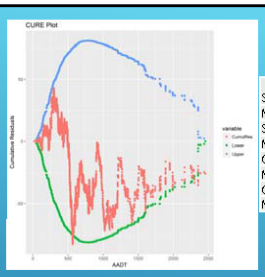
0.4

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

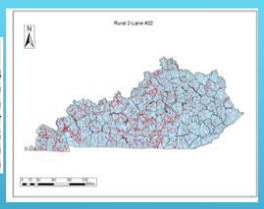
LENGTH >= 0.1

AADT < 2500

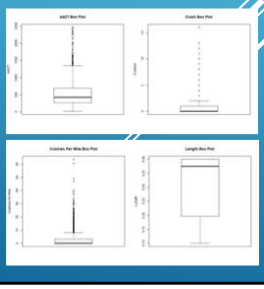
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6



Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.7	2.3	540.0	0.304
StdDev	1.2	4.2	393.6	0.110
Min	0	0	17	0.100
Q1	0	0	265	0.197
Median	0	0	429	0.375
Q3	1	3.236246	695	0.400
Max	16	63.69427	2456	0.400



	Values
Sample	6081
Length	1849.668
Crashes	4274
R2	0.45
CDP	1.36
MACD	82.36
MAD	0.70
Theta	1.99
Alpha	-5.32
Beta	0.98
StdErr	0.1542



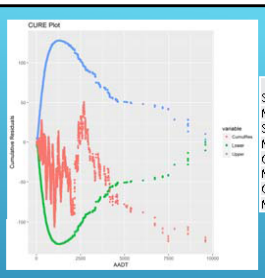
0.4

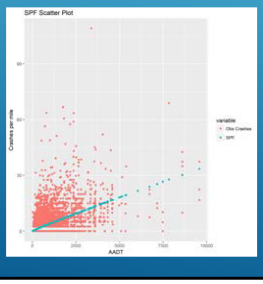
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

AADT > 0

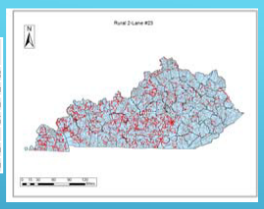
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

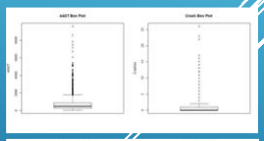


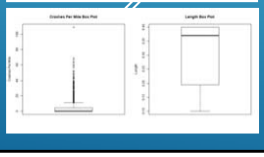


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.9	2.8	694.3	0.303
StdDev	1.5	5.3	744.1	0.110
Min	0	0	10	0.100
Q1	0	0	250	0.195
Median	0	0	470	0.371
Q3	1	4.524887	858	0.400
Max	26	108.9109	9586	0.400

	Values
Sample	10937
Length	3314.874
Crashes	9430
R2	0.55
CDP	0.96
MACD	124.51
MAD	0.78
Theta	2.20
Alpha	-5.08
Beta	0.94
StdErr	0.1176







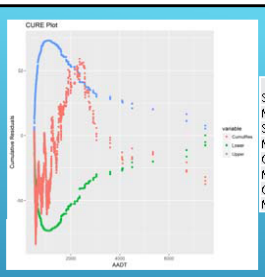
0.4

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

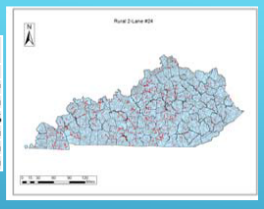
AADT > 500

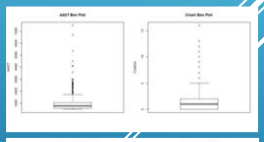
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

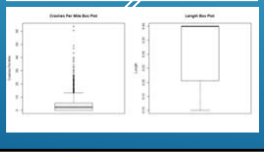


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.2	3.9	951.7	0.311
StdDev	1.6	5.5	607.7	0.109
Min	0	0	504	0.100
Q1	0	0	604.5	0.206
Median	1	2.5	784	0.400
Q3	2	5.361969	1070	0.400
Max	16	63.69427	7481	0.400

	Values
Sample	2667
Length	828.512
Crashes	3162
R2	0.33
CDP	27.63
MACD	83.43
MAD	1.00
Theta	2.31
Alpha	-4.82
Beta	0.90
StdErr	0.2053







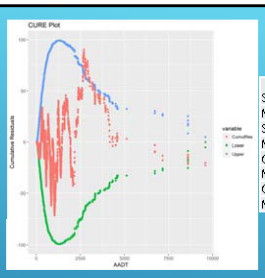
0.4

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

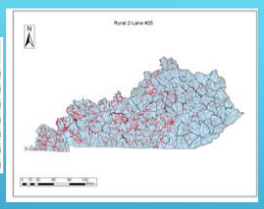
LENGTH = 0.4

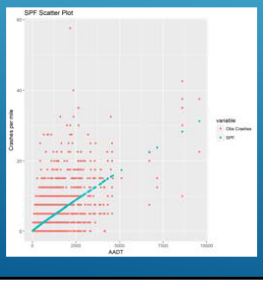
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

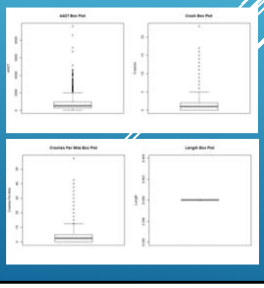


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.1	2.9	762.3	0.400
StdDev	1.7	4.4	793.6	0.000
Min	0	0	13	0.400
Q1	0	0	286	0.400
Median	1	2.5	531	0.400
Q3	2	5	984.5	0.400
Max	23	57.5	9586	0.400





	Values
Sample	5171
Length	2068.4
Crashes	5907
R2	0.56
CDP	2.86
MACD	90.77
MAD	0.93
Theta	2.75
Alpha	-5.12
Beta	0.93
StdErr	0.2009



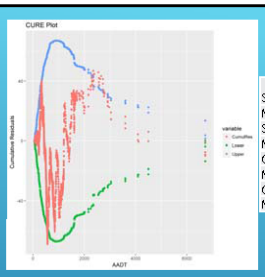
0.4

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

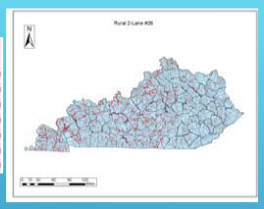
LENGTH = 0.4

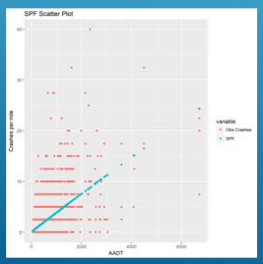
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

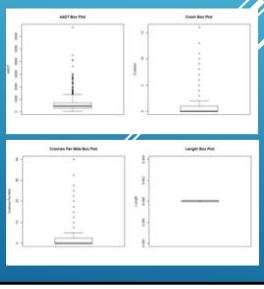


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.9	2.4	606.4	0.400
StdDev	1.5	3.6	531.2	0.000
Min	0	0	46	0.400
Q1	0	0	290	0.400
Median	0	0	470	0.400
Q3	1	2.5	740	0.400
Max	16	40	6720	0.400





	Values
Sample	2931
Length	1172.4
Crashes	2769
R2	0.44
CDP	2.97
MACD	68.71
MAD	0.86
Theta	2.31
Alpha	-5.31
Beta	0.96
StdErr	0.2300



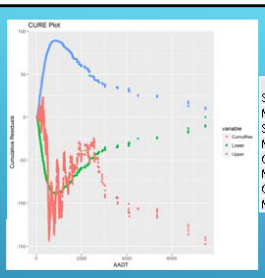
0.5

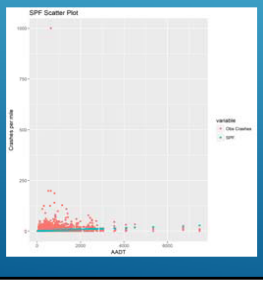
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH > 0

AADT > 0

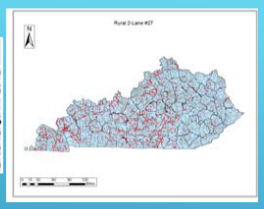
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

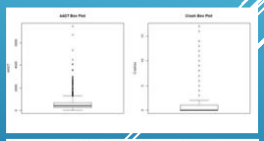


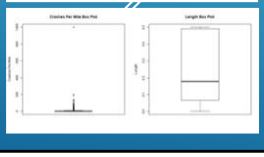


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.6	2.9	546.1	0.239
StdDev	1.2	13.6	513.6	0.185
Min	0	0	17	0.001
Q1	0	0	243	0.066
Median	0	0	412	0.179
Q3	1	2.136752	665	0.492
Max	17	1000	7481	0.500

	Values
Sample	8514
Length	2036.725
Crashes	4977
R2	0.55
CDP	25.59
MACD	147.41
MAD	0.58
Theta	1.97
Alpha	-5.00
Beta	0.93
StdErr	0.1446







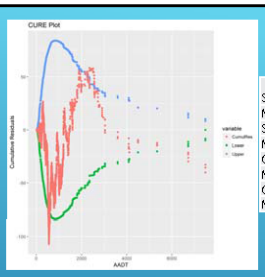
0.5

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

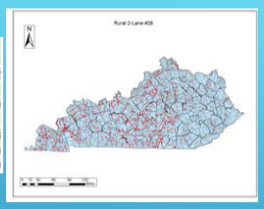
AADT > 0

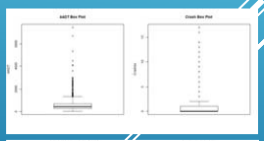
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6



Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.8	2.4	570.4	0.342
StdDev	1.4	4.2	530.0	0.151
Min	0	0	17	0.100
Q1	0	0	265	0.191
Median	0	0	433	0.358
Q3	1	4	697	0.500
Max	17	60.60606	7481	0.500

	Values
Sample	5498
Length	1880.189
Crashes	4459
R2	0.52
CDP	6.49
MACD	107.12
MAD	0.76
Theta	2.23
Alpha	-4.98
Beta	0.92
StdErr	0.1781





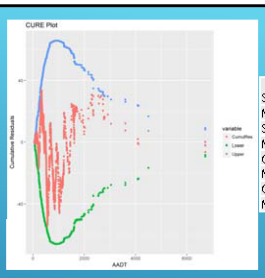
0.5

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

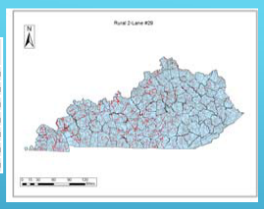
LENGTH = 0.5

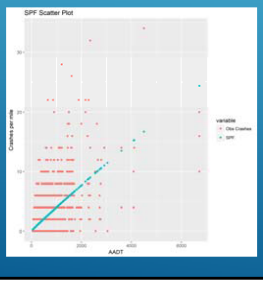
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

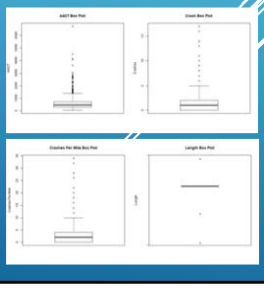


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.2	2.5	609.8	0.500
StdDev	1.7	3.5	529.5	0.000
Min	0	0	46	0.500
Q1	0	0	295	0.500
Median	1	2	471	0.500
Q3	2	4	743	0.500
Max	17	34	6720	0.500





	Values
Sample	2099
Length	1049.5
Crashes	2583
R2	0.48
CDP	0.29
MACD	53.53
MAD	1.02
Theta	2.57
Alpha	-5.17
Beta	0.95
StdErr	0.2750



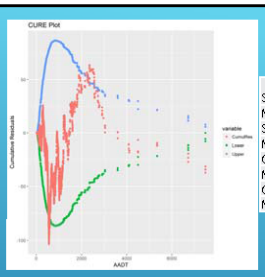
0.6

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

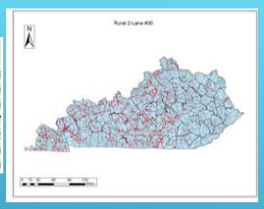
LENGTH >= 0.1

AADT > 0

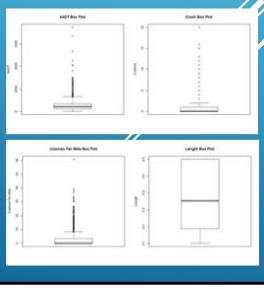
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6



Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.9	2.4	567.8	0.373
StdDev	1.5	4.1	526.9	0.190
Min	0	0	17	0.100
Q1	0	0	264	0.187
Median	0	0	430	0.352
Q3	1	3.333333	697	0.600
Max	20	60.60606	7481	0.600



	Values
Sample	5051
Length	1883.237
Crashes	4479
R2	0.54
CDP	5.44
MACD	103.50
MAD	0.80
Theta	2.27
Alpha	-4.98
Beta	0.92
StdErr	0.1793



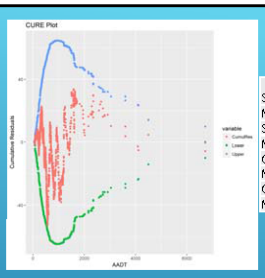
0.6

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

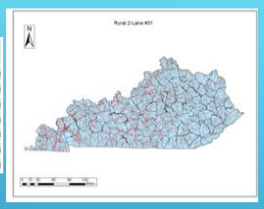
LENGTH = 0.6

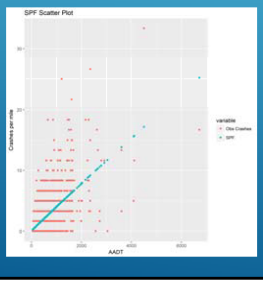
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

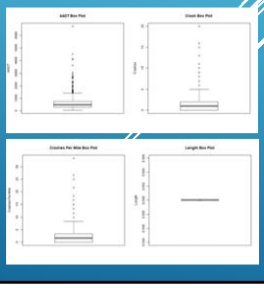


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.5	2.5	614.0	0.600
StdDev	2.0	3.3	526.3	0.000
Min	0	0	46	0.600
Q1	0	0	299	0.600
Median	1	1.666667	476	0.600
Q3	2	3.333333	750	0.600
Max	20	33.33333	6720	0.600





	Values
Sample	1585
Length	951
Crashes	2360
R2	0.51
CDP	0.25
MACD	52.28
MAD	1.15
Theta	2.68
Alpha	-5.27
Beta	0.96
StdErr	0.3001



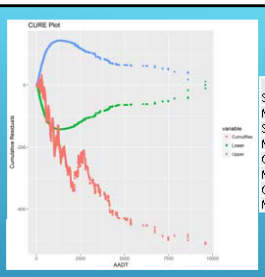
0.7

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

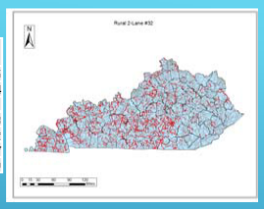
LENGTH > 0

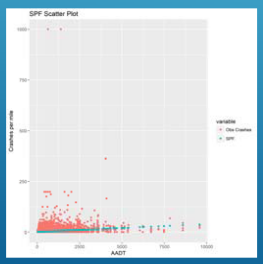
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

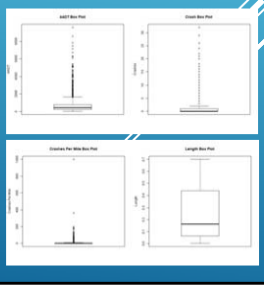


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.8	3.4	636.4	0.265
StdDev	1.7	15.8	701.0	0.244
Min	0	0	10	0.001
Q1	0	0	217	0.063
Median	0	0	424	0.162
Q3	1	28.57143	789	0.437
Max	32	1000	9586	0.700





	Values
Sample	13598
Length	3605.093
Crashes	10518
R2	0.64
CDP	24.89
MACD	511.63
MAD	0.67
Theta	2.20
Alpha	-5.16
Beta	0.96
StdErr	0.1141



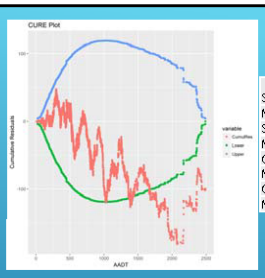
0.7

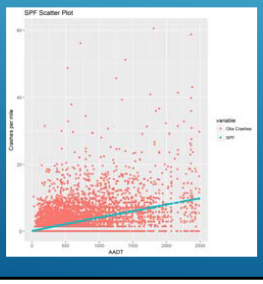
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

AADT <= 2500

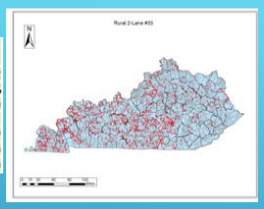
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

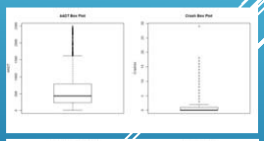


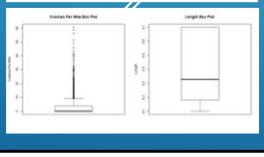


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.0	2.5	582.8	0.392
StdDev	1.7	4.4	477.3	0.226
Min	0	0	10	0.100
Q1	0	0	234	0.181
Median	0	0	436	0.329
Q3	1	3.717472	788	0.700
Max	29	60.60606	2489	0.700

	Values
Sample	8227
Length	3221.69
Crashes	8057
R2	0.57
CDP	5.28
MACD	180.65
MAD	0.83
Theta	2.43
Alpha	-5.22
Beta	0.96
StdErr	0.1454







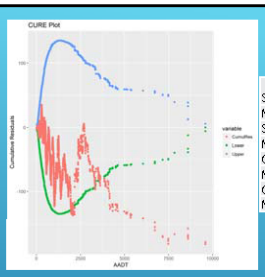
0.7

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

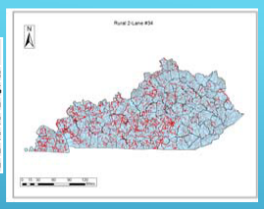
AADT > 0

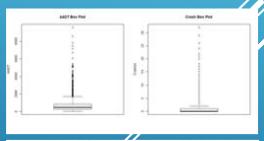
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

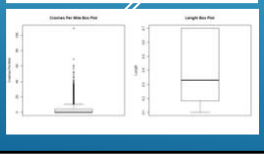


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.1	2.8	676.7	0.393
StdDev	2.0	5.2	736.5	0.226
Min	0	0	10	0.100
Q1	0	0	240	0.182
Median	0	0	450	0.332
Q3	1	4.285714	829	0.700
Max	32	108.9109	9586	0.700

	Values
Sample	8496
Length	3342.308
Crashes	9528
R2	0.63
CDP	1.72
MACD	180.89
MAD	0.90
Theta	2.55
Alpha	-5.13
Beta	0.94
StdErr	0.1449







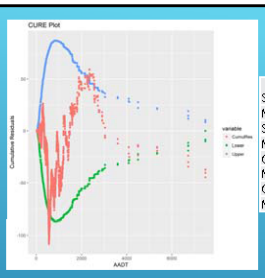
0.7

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

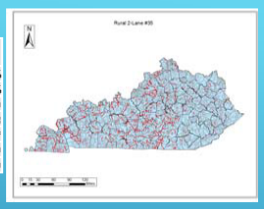
LENGTH >= 0.1

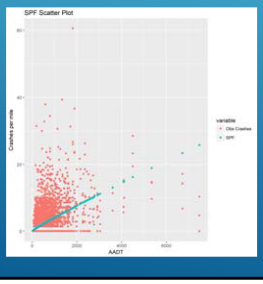
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

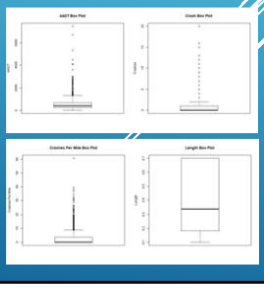


Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.9	2.4	566.4	0.396
StdDev	1.6	4.1	529.3	0.226
Min	0	0	17	0.100
Q1	0	0	262	0.183
Median	0	0	427	0.339
Q3	1	3.533569	696	0.700
Max	20	60.60606	7481	0.700





	Values
Sample	4765
Length	1887.823
Crashes	4495
R2	0.58
CDP	5.83
MACD	108.47
MAD	0.82
Theta	2.44
Alpha	-4.98
Beta	0.92
StdErr	0.1999



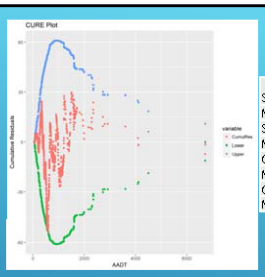
0.7

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

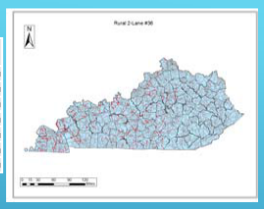
LENGTH = 0.7

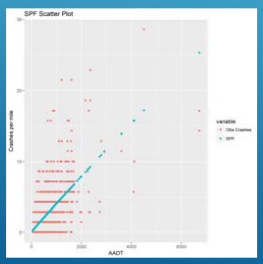
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

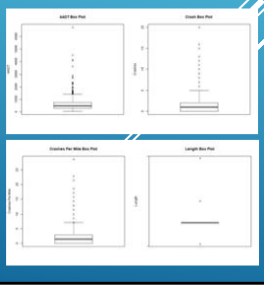


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.7	2.5	614.3	0.700
StdDev	2.2	3.2	536.0	0.000
Min	0	0	46	0.700
Q1	0	0	298	0.700
Median	1	1.428571	478	0.700
Q3	2	2.857143	750	0.700
Max	20	28.57143	6720	0.700





	Values
Sample	1205
Length	843.5
Crashes	2108
R2	0.54
CDP	0.25
MACD	53.94
MAD	1.24
Theta	3.22
Alpha	-5.25
Beta	0.96
StdErr	0.4127



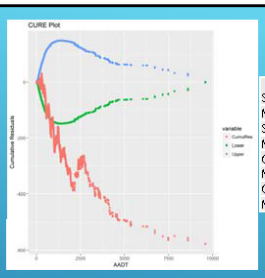
1.0

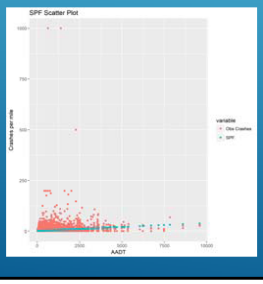
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH > 0

AADT > 0

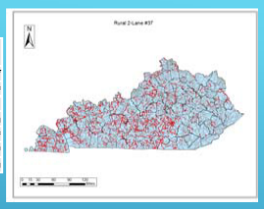
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

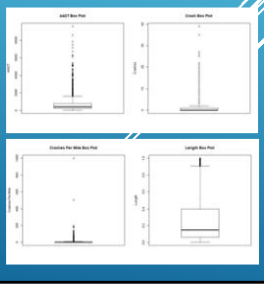




Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.8	3.5	627.1	0.287
StdDev	2.0	16.6	692.9	0.309
Min	0	0	10	0.001
Q1	0	0	215	0.060
Median	0	0	416	0.149
Q3	1	2.868413	776.75	0.399
Max	39	1000	9586	1.000

	Values
Sample	12582
Length	3605.093
Crashes	10518
R2	0.68
CDP	27.76
MACD	577.87
MAD	0.69
Theta	2.31
Alpha	-5.17
Beta	0.96
StdErr	0.1247





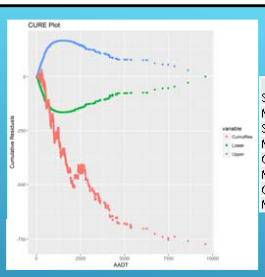
Raw

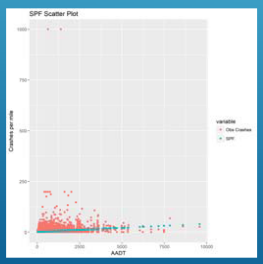
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH > 0

AADT > 0

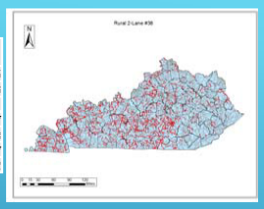
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

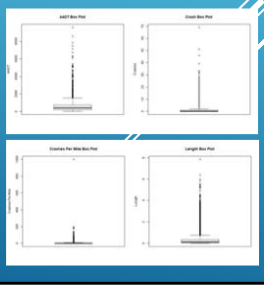




Stat	Crash	Crashes PerMile	AADT	Length
Mean	0.9	3.4	608.8	0.315
StdDev	2.7	16.6	675.3	0.518
Min	0	0	10	0.001
Q1	0	0	205	0.057
Median	0	0	405	0.133
Q3	1	2.442301	742	0.337
Max	69	1000	9586	7.832

	Values
Sample	11435
Length	3605.093
Crashes	10518
R2	0.75
CDP	30.47
MACD	772.88
MAD	0.71
Theta	2.35
Alpha	-5.22
Beta	0.97
StdErr	0.1383





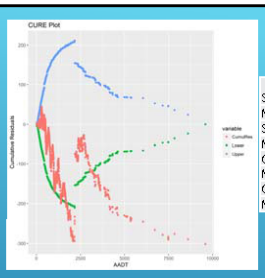
Raw

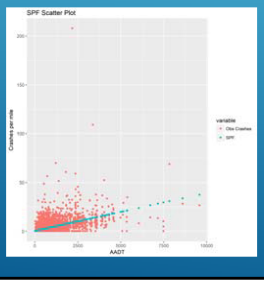
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

AADT > 0

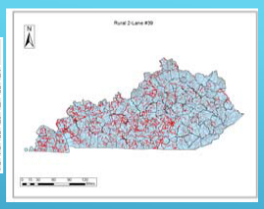
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

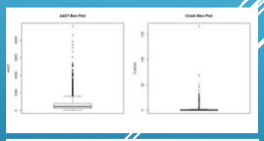


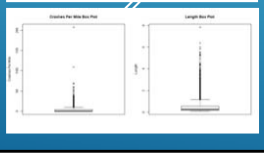


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.5	2.8	644.9	0.505
StdDev	3.9	5.8	709.7	0.613
Min	0	0	10	0.100
Q1	0	0	226	0.168
Median	0	0	429	0.283
Q3	1	3.896104	789	0.572
Max	166	207.5	9586	7.832

	Values
Sample	6649
Length	3359.041
Crashes	9764
R2	0.56
CDP	4.53
MACD	301.65
MAD	1.03
Theta	2.47
Alpha	-5.27
Beta	0.97
StdErr	0.1428







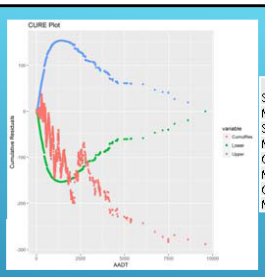
Raw

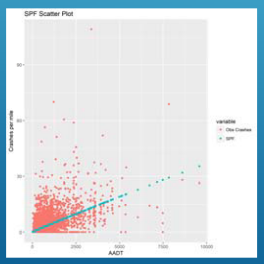
RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

LENGTH >= 0.1

AADT > 0

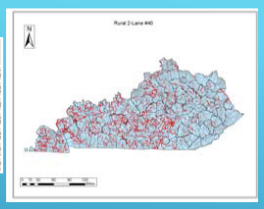
NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

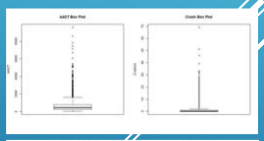


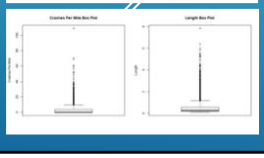


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.4	2.7	644.7	0.505
StdDev	3.4	5.3	709.5	0.613
Min	0	0	10	0.100
Q1	0	0	226	0.168
Median	0	0	429	0.283
Q3	1	3.894839	789	0.572
Max	69	108.9109	9586	7.832

	Values
Sample	6648
Length	3358.241
Crashes	9598
R2	0.77
CDP	3.44
MACD	288.02
MAD	1.00
Theta	2.82
Alpha	-5.16
Beta	0.95
StdErr	0.1811







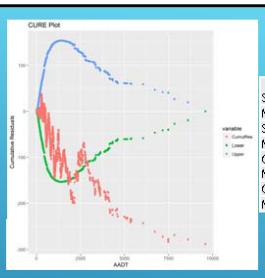
Raw

RURAL 2-LANE
9 FOOT LANES
3 FOOT SHOULDERS

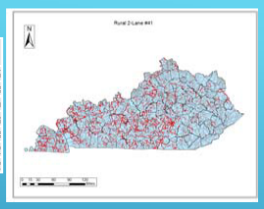
LENGTH >= 0.1

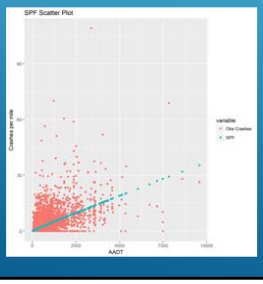
AADT > 0

NO HORIZ. CURVES
NO US60 OUTLIER
NO US431 OUTLIER
SPEED LIMIT 50+
FUNC. CLASS 6

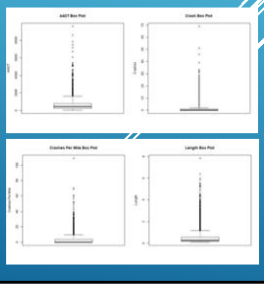


Stat	Crash	Crashes PerMile	AADT	Length
Mean	1.4	2.7	644.7	0.505
StdDev	3.4	5.3	709.5	0.613
Min	0	0	10	0.100
Q1	0	0	226	0.168
Median	0	0	429	0.283
Q3	1	3.694839	789	0.572
Max	69	108.9109	9586	7.832





	Values
Sample	6648
Length	3358.241
Crashes	9598
R2	0.77
CDP	3.44
MACD	288.02
MAD	1.00
Theta	2.82
Alpha	-5.18
Beta	0.95
StdErr	0.1811



References

AASHTO. Highway Safety Manual. First Edition. American Association of State Highway and Transportation Officials. Washington, D.C., 2010. Print.

Borsos, Attila, John N. Ivan, and Gyula Orosz. "Development of safety performance functions for two-lane rural first-class main roads in Hungary." *Transport Research Arena (TRA) 5th Conference: Transport Solutions from Research to Deployment*, 2014.

Cafiso, Salvatore, Alessandro Di Graziano, Giacomo Di Silvestro, and Grazia La Cava. "Safety Performance Indicators for Local Rural Roads: Comprehensive Procedure from Low-Cost Data Survey to Accident Prediction Model." *Transportation Research Board 87th Annual Meeting*, No. 08-2542, 2008.

Cafiso, Salvatore, Giacomo Di Silvestro, Bhagwant Persaud and Morjina Begum. "Revisiting Variability of Dispersion Parameter of Safety Performance for Two-Lane Rural Roads." *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2148, no. 05, 2010, pp. 38-46.

Cafiso, Salvatore, Carmelo D'Agostino, and Bhagwant Persaud. "Investigating the influence of segmentation in estimating safety performance functions for roadway sections." *Transportation Research Board 92nd Annual Meeting,* 2013.

Cook, Dan, Reginald Souleyrette and Justin Jackson. "Effect of Road Segmentation on Highway Safety Analysis." *Transportation Research Board 90th Annual Meeting*, no. 11-1995, 2011.

Depaire, Benoit, Geert Wets and Koen Vanhoof. "Traffic Accident Segmentation by Means of Latent Class Clustering." *Accident Analysis and Prevention*, vol. 40, no. 4, 2008, pp. 1257-1266.

Elvik, Rune. "Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads." *Transportation Research Record: Journal of the Transportation Research Board*, No. 2083, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 72–75.

ESRI. "GIS Dictionary: *MAUP" Esri Support GIS Dictionary*, support.esri.com/en/other-resources/gis-dictionary/term/maup. Accessed 28 Oct. 2017.

FHWA. "Assessment of Geographic Information Systems' (GIS) Needs and Obstacles in Traffic Safety." *U.S. Department of Transportation: Federal Highway Administration*. 2013. Print.

FHWA. “How to Develop and Use SPFs.” *Crash Modification Factors Clearinghouse*, http://www.cmfclearinghouse.org/resources_spf.cfm, 2016.

Geedipally, Srinivas R., Dominique Lord, and Byung-Jung Park. “Analyzing Different Parametrizations of the Varying Dispersion Parameter as a Function of Segment Length.” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2103, no. 13, 2008.

Geyer, Judy, Elena Lankina, Ching Yao Chan, David Ragland, Trinh Pham and Ashkan Sharafsale. “Methods for identifying high collision concentration locations for potential safety improvements.” *California PATH Program, Institute of Transportation Studies*, University of California at Berkeley, 2008.

Green, Eric R., and Kenneth R Agent. “Evaluation of the Locations of Kentucky's Traffic Crash Data.” *Proc. of 3rd International Conference on Road Safety and Simulation,* Indianapolis, IN. 2011

Green, Eric R., Kenneth R. Agent, Jerry G. Pigman and Michael A. Fields. “Analysis of Traffic Crash Data in Kentucky (2010 - 2014).” *Kentucky Transportation Center Research Report,* 2015.

Green, Eric R., C. Blackden, M.A. Fields. "Spatial Database For Intersections", 95th Annual TRB, Washington, DC; Transportation Research Board: Washington, DC, 2016

Green, Eric, Nikiforos Stamatiadis, and Reginald Souleyrette. "Segment Length and Highway Safety Analysis: Does It Matter?," 96th Annual TRB, Washington, DC; Transportation Research Board: Washington, DC, 2017

Harwood, Douglas W., Karin M. Bauer, David K. Gilmore, Reginald Souleyrette and Zachary N. Hans. “Validation of U.S. Road Assessment Program Star Ratting Protocol.” 2015

Hauer, Ezra and Joseph Bamfo. “Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables.” *University of Toronto*, Proceedings ICTCT (International Cooperation on Theories and Concepts in Traffic Safety), Lund, Sweden, 1997.

Hauer, Ezra. “Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation.” *Accident Analysis & Prevention*, vol. 33, no. 6, Nov. 2001, pp. 799-808.

Hauer, Ezra, Douglas Harwood, Forrest Council and Michael Griffith. “Estimating Safety by the Empirical Bayes Method: A Tutorial” *Transportation Research Record*, vol. 1784, 2002.

Hauer, Ezra. “Statistical Road Safety Modeling.” *Transportation Research Record* vol. 1897, 2004, pp. 81-87.

Hauer, Ezra. *The Art of Regression Modeling in Road Safety*. Springer International Publishing Switzerland, 2015.

Kononov, Jake, Craig Lyon, and Bryan K. Allery. "Relation of Flow, Speed, and Density of Urban Freeways to Functional Form of a Safety Performance Function" *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2236, 2011, pp. 11-19.

Koorey, Glen. “Road data aggregation and sectioning considerations for crash analysis.” *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2103 no. 1, 2009, pp. 61-68.

Kwon, Oh Hoon, Min Ju Park, Hwasoo Yeo and Koohong Chung. "Evaluating the performance of network screening methods for detecting high collision concentration locations on highways." *Accident Analysis & Prevention*, vol. 51, 2013, pp. 141-149.

Lee, Jaeyoung, Abdel-Aty, Mohamed and Jiang, Ximiao. “Development of zone system for macro-level traffic safety analysis.” *Journal of Transport Geology*, vol 38, 2014, pp. 13-21.

Lu, Jinyan, Albert Gan, Kirolos Haleem and Wanyang Wu. “Clustering-based roadway segment division for the identification of high-crash locations.” *Journal of Transportation Safety & Security,* vol. 5, no. 3, 2013, pp. 224-239.

Lyon, Craig, Bhagwant Persaud and Frank Gross. “The Calibrator – Calibrate, Critique, CURE: An SPF Calibration Tool User Guide.” *U.S. Department of Transportation: FHWA Roadway Safety Data Program*, 2016.

“Quick Facts 2016.” *U.S. Department of Transportation: National Highway Traffic Safety Administration*, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812451, accessed 2017 October 28th.

Ogle, Jennifer H., Priyanka Alluri and Wayne Sarasua. “MMUCC and MiRE: The role of segmentation in safety analysis.” Transportation Research Board Annual Meeting, Washington D.C., 2011.

Openshaw, S., “Ecological Fallacies and the Analysis of Areal Census Data”, *Environment and Planning A*, vol. 16, no. 1, 1984, pp. 17-31.

Persaud, Bhagwant. “The Problem of Regression-to-the-Mean in Estimating the Effectiveness of Road Safety Countermeasures.” *Insurance Institute for Highway Safety*, March 1984, http://www.iihs.org/frontend/iihs/documents/masterfiledocs.ashx?id=729

Persaud, Bhagwant and Craig Lyon. “Bayes Before-after safety studies: Lessons learned from two decades of experience and future directions.” *Accident Analysis & Prevention*, vol. 39, 2007, pp. 546-555.

Qin, Xiao and Adam Wellner. “Segment length impact on highway safety screening analysis.” *Transportation Research Record: Journal of the Transportation Research Board.* 2011, *No. 12-0644*. 2012.

Souleyrette, Reginald R., Robert P Haas., and Thomas H Maze. “Validation and implication of segmentation on Empirical Bayes for highway safety studies.” *University of Kentucky: From the Selected Works of Reginald R. Souleyrette*, 2007, pp. 85-94.

Srinivasan, Bhagwant, Craig Lyon, Bhagwant Persaud, Carol Martell and Jongdae Baek. “Methods for Identifying High Collision Concentration Locations (HCCL) for Potential Safety Improvements: Phase II, Evaluation of Alternative Methods for Identifying HCCL.” *California Department of Transportation,* 2011.

Srinivasan, Raghavan and Karin Bauer. “Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs.” *U.S. Department of Transportation: Federal Highway Administration*, 2013a.

Srinivasan, Raghavan, Daniel Carter and Karin Bauer. “Safety Performance Function Decision Guide: SPF Calibration vs SPF Development.” *U.S. Department of Transportation: Federal Highway Administration*, 2013b.

Stamatiadis, Nikiforos, Adam Kirk, Jeff Jasper, and Samantha Wright. “Development of a Context Sensitive Multimodal Functional Classification System.” *Transportation Research Record: Journal of the Transportation Research Board* 2017 2638:, 18-25

Tegge, Robert A., Jang-Hyeon Jo and Yanfeng Ouyang. “Development And Application Of Safety Performance Functions For Illinois.” *Illinois Center for Transportation*, 2010 .

usRAP. Guide to Producing usRAP Star Ratings and Safer Roads Investment Plans. AAA Foundation for Traffic Safety. Not For Release. November 2012, p 8.

Valent, Francesca, Flavio Schiava, Cecilia Savonitto, Tolinda Gallo, Silvio Brusaferro and Fabio Barbone. "Risk factors for fatal road traffic accidents in Udine, Italy." *Accident Analysis & Prevention,* vol. 34, 2002, pp. 71-84.

Wu, Kun-Feng, Scott C. Himes and Martin T. Pietrucha. "Evaluation of Effectiveness of the Federal Highway Safety Improvement Program." *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2318 no. 1, 2012, pp. 23-34

Zhang, Yunlong, Zhirui Ye and Dominique Lord. " Estimating Dispersion Parameter of Negative Binomial Distribution for Analysis of Crash Data: Bootstrapped Maximum Likelihood Method." *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2019, no. 03, 2007, pp. 15-21.

Zhang, Yunlong, Yuanchang Xie and Linhua Li. "Crash Frequency Analysis of Different Types of Urban Roadway Segments Using Generalized Addictive Model." *Transportation Research Board 88th Annual Meeting,* 2009.

Vita

Eric Green

Education

Master of Science, Civil Engineering, University of Kentucky, 2002
Bachelor of Science, Civil Engineering, University of Kentucky, 2000

Professional Experience

2017-Present Associate Program Manager, Traffic Operations and Safety
Kentucky Transportation Center, University of Kentucky
2002-2017 Research Engineer, Traffic Operations and Safety
Kentucky Transportation Center, University of Kentucky
1998-2002 Research Assistant, Traffic Operations and Safety
Kentucky Transportation Center, University of Kentucky

Honors and Awards

Professional Engineer (KY, No. 23983)
Certified Geographic Information Systems (GIS) Professional (GISP) (No. 22980)
Kentucky Transportation Cabinet for Excellence and Dedication to Highway Safety
Award, Lifesavers Conference 2006, April 2006
Kentucky Section of the Institute of Transportation Engineers Individual Activity Award,
KYSITE Annual Banquet, November 2008
4 E's Of Highway Safety Award – Engineering
Lifesavers Conference 2016, March 2016

Publications

- E.R. Green, K.R. Agent. "Evaluation of the Locations of Kentucky's Traffic Crash Data", Roadway Safety and Simulation Conference, Indianapolis, IN; Transportation Research Board: Washington, DC, 2011
- E.R. Green, J.G. Pigman, J.R. Walton, S. McCormack. "Identification of Secondary Crashes and Recommended Countermeasures", 91st Annual TRB, Washington, DC; Transportation Research Board: Washington, DC, 2012
- E.R. Green, J.R. Ripy, X. Xu, M. Chen. "Conflation Methodologies to Incorporate Consumer Travel Data into State HPMS Data Sets", 92nd Annual TRB, Washington, DC; Transportation Research Board: Washington, DC, 2013
- D.W. Harwood, R.R. Souleyrette, E.R. Green, M.A. Fields. "Comparison of Countermeasure Selection Methods for Use in Road Safety Management", 93rd Annual TRB, Washington, DC; Transportation Research Board: Washington, DC, 2014
- E. C. Davis, E.R. Green, N. Stamatiadis, G. A. Winchester, R. R. Souleyrette, J.G. Pigman. "Highway Safety Manual Methodologies and Benefit-Cost Analysis In Program-Level Segment Selection And Prioritization", Roadway Safety and Simulation Conference, Orlando, FL; University of Central Florida & the University of Tennessee, 2015
- E.R. Green, M.A. Fields. "A Methodology to Prioritize the Locations of Cable Barrier

Installations in Kentucky", Roadway Safety and Simulation Conference, Orlando, FL; University of Central Florida & the University of Tennessee, 2015

- E.R. Green, C. Blackden, M.A. Fields. "Spatial Database For Intersections", 95th Annual TRB, Washington, DC; Transportation Research Board: Washington, DC, 2016
- E.R. Green, K.R. Agent, E. Lammers. "Development of an Improved Method for Determining Advisory Speeds on Horizontal Curves", 96th Annual TRB, Washington, DC; Transportation Research Board: Washington, DC, 2017