ESTIMATION OF CRASH MODIFICATION FUNCTIONS USING SITE-LEVEL INFORMATION FROM RESULTS OF EMPIRICAL BAYES BEFORE-AFTER EVALUATIONS

FINAL REPORT

SOUTHEASTERN TRANSPORTATION CENTER

RAGHAVAN SRINIVASAN AND BO LAN

JANUARY 2016

US DEPARTMENT OF TRANSPORTATION GRANT DTRT13-G-UTC34

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Title		5. Report Date January 2016 6. Source Organization Code Budget
7. Author(s) Srinivasan, Raghavan; Lan, Bo		8. Source Organization Report No. STC-2015-##-XX
9. Performing Organization Name and Address Southeastern Transportation Center UT Center for Transportation Research 309 Conference Center Building Knoxville TN 37996-4133		10. Work Unit No. (TRAIS) 11. Contract or Grant No. DTRT13-G-UTC34
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary of Transportation–Research 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Final Report: September 2014 – December 2015 14. Sponsoring Agency Code USDOT/OST-R/STC
15. Supplementary Notes:		
16. Abstract This study investigates the performance of three different forms of crash modification functions (CMFunctions). Two traditional forms (normal and lognormal) are compared with a new negative binomial regression approach. Data from the results of a before-after empirical Bayes before-after evaluation from North Carolina were used for this investigation. The treatment was the introduction of traffic signals at locations that were controlled by stop signs of minor roads. Unlike the traditional approach of using the CMF from a site (or group of sites) as a dependent variable, this study investigated the use of numerator of the CMF as the dependent variable and denominator of the CMF as an offset in the estimation of CMFunctions using negative binomial regression. The overall conclusion is that the use of negative binomial regression for estimating CMFunctions has a lot of potential and should be investigated in future research.		
17. Key Words Normal regression, Lognormal regression, negative binomial regression, weighted regression, crash modification factor, crash modification function, traffic signal, empirical bayes	18. Distribution Statement Unrestricted; Document is available to the public through the National Technical Information Service; Springfield, VT.	

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

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

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Until recently, most CMFs from before-after evaluations were provided as point estimates representing the average effect of a particular treatment. In some cases, disaggregate analysis is conducted and CMF estimates are provided for different categories of an independent variable (e.g., CMF estimates are provided for low and high AADT categories). With point estimates, the practitioner is generally forced to use the same CMF value for a particular treatment. To overcome this problem, *crash modification functions* (CMFunctions) have been proposed.

The main goal of this study was to investigate different model forms for estimating CMFunctions using data from the results of a before-after EB evaluation. Three different model forms were explored including two traditional approaches, normal regression (model form 1) and lognormal regression (model form 2), and a new negative binomial regression approach (model form 3). With the traditional approaches, the dependent variable is the CMF for a particular site (or group of sites), and sites are usually grouped (or aggregated) in order to obtain a stable estimate of the CMF and the standard error of the CMF. With the new negative binomial regression approach, the numerator of the CMF is used as the dependent variable and the denominator of the CMF is used as an offset. The negative binomial regression approach does not require the aggregating of data from individual sites, and could provide more insights that may be lost due to the aggregation.

The project team sought data from multiple states in order to compare the performance of these different types of CMFunctions. Finally, data from the results of a before-after evaluation conducted for North Carolina Department of Transportation were used for comparing the results from the three different approaches for estimating CMFunctions. The treatment was the implementation of traffic signals at intersections that were controlled by stop signs on the minor roads.

First, the data were aggregated and CMFunctions were estimated using the three model forms. For the first two model forms, CMFunctions were estimated with and without weights. With the aggregated data, the results from model form 3 compared quite favorably with that of the traditional model forms 1 and 2, especially for the CMFunction that was estimated for injury and fatal crashes. Then, CMFunctions based on model form 3 were estimated using the original site-level results from the before-after evaluation (i.e., without aggregation). The CMFunctions with the disaggregate data included independent variables that were not significant in the models based on the aggregated data, indicating the value of using model form 3 to estimate CMFunctions using disaggregate data.

Matching funds for this study were provided by National Cooperative Highway Research Program (NCHRP) through Project 17-63 (*Guidance for the Development and Application of Crash Modification Factors*). The authors thank NCHRP for their support. The authors also thank the members of the NCHRP Project 17-63 for their support throughout this study.

BACKGROUND

A crash modification factor (CMF) is an estimate of the change in crashes expected after implementation of a countermeasure. Practitioners can use the CMF in quantifying safety in many ways including as part of the roadway management process, roadway safety audits, alternatives development and analysis, and design decisions and exceptions (FHWA, 2014). There are many ways to estimate the CMF associated with an engineering improvement. The methods for estimating CMFs can be divided into two broad categories: cross-sectional and before-after. Before-after studies include "all techniques by which one may study the safety effect of some change that has been implemented on a group of entities (road sections, intersections, drivers, vehicles, neighborhoods, etc.)" (Hauer, 1997, p. 2). On the other hand, cross-sectional studies include those where "one is comparing the safety of one group of entities having some common feature (say, STOP controlled intersections) to the safety of a different group of entities not having that feature (say, YIELD controlled intersections), in order to assess the safety effect of that feature (STOP versus YIELD signs)" (Hauer, 1997, p. 2, 3).

Many safety researchers feel that CMFs developed using cross-sectional studies may not always be reliable because cross-sectional models rarely represent causal relationships. The issues associated with the CMFs derived from cross-sectional models are discussed in some detail in Gross et al., (2010) and Carter et al., (2012). There is some consensus in the safety research community that properly designed before-after studies provide more reliable estimates of before-after studies. In before-after studies, the CMF is estimated based on two parameters: (1) crashes that occurred at the treated sites after the treatment is implemented, and (2) an estimate of the crashes that would have occurred during the same 'after' period had the treatment not been implemented, and the variance of this estimate. Often, sites are not selected for treatment at random; practitioners usually select high crash locations for treatment. This non-random selection can potentially lead to bias due to regression to the mean (RTM). Using the empirical Bayes before-after method has now been accepted as one way of addressing the potential bias due to RTM. Since before-after evaluations are based on information from sites that were treated in some fashion, the sample size for before-after evaluations usually tend to be smaller than the sample that is used in cross-sectional studies.

Until recently, most CMFs from before-after evaluations were provided as point estimates representing the average effect of a particular treatment. In some cases, disaggregate analysis is conducted and CMF estimates are provided for different categories of an independent variable (e.g., CMF estimates are provided for low and high AADT categories). With point estimates, the practitioner is generally forced to use the same CMF value for a particular treatment. To overcome this problem, *crash modification functions* (CMFunctions) have been proposed. Unlike CMFs, CMFunctions are equations that represent the safety effect of a treatment as a function of site characteristics. CMFunctions are not necessarily a new concept. In fact, some CMFunctions have been included in the 1st edition of the Highway Safety Manual (HSM) (AASHTO, 2010). However, most of the CMFunctions in the HSM are based on cross-sectional models.

The objective of this effort is to examine different ways of estimating CMFunctions based on information available from the results of an empirical Bayes before-after evaluation. The following section provides an overview of CMFunctions that have been estimated in the recent past. This is followed by a discussion of the different methods for estimating CMFunctions that will be examined in this study. The data used for this investigation is then presented. Finally, the results and conclusions are presented .

RECENT RESEARCH ON CRASH MODIFICATION FUNCTIONS AND IMPLICATIONS FOR THIS STUDY

Summary of Recent Research

Dr. Rune Elvik was one of the earliest to advocate the need for CMFunctions in highway safety. In 2005, Dr. Elvik published a paper providing guidance for conducting meta-analysis (Elvik, 2005a). Meta-analysis is essentially the approach to combine the results from multiple studies. This is typically done by weighting the safety estimate from each study based on the inverse of the variance of the estimate. In Elvik, (2005a), there is also a discussion about the use of meta-regression, which represents the safety effect as a function that includes site characteristics; this meta-regression is done based on the results from the different studies. In other words, meta-regression results in a CMFunction. Subsequent papers from Dr. Elvik provided CMFunctions for relationship between speed and crashes, speed enforcement, and bypass roads (e.g., Elvik, 2005b; Elvik, 2009; Elvik, 2011). His most recent paper (Elvik, 2015), provides methodological guidelines for developing crash modification functions.

Dr. Elvik advocates that different functional and model forms should be explored when CMFunctions are estimated. He also suggests that the variance of the individual CMF estimates should be considered in appropriately weighting each estimate.

Most of the CMFunctions estimated by Dr. Elvik are based on 'aggregate' data, i.e., each data point is a CMF from a particular study usually based on data from many sites. These CMFunctions are indeed very useful to study the effect of certain treatment characteristics. However, in order to determine the effect of site characteristics such as AADT, CMFunctions would need to be estimated using more 'disaggregate' data. A few recent studies have estimated CMFunctions using disaggregate data. A brief overview of such studies is provided below.

De Pauw et al., (2014), developed CMFs for an intersection black spot treatment program based on a before-after EB evaluation. The results from the EB evaluation were then used to estimate a CMFunction. Maximum likelihood linear regression was used with the natural logarithm of the CMF for each site as the dependent variable. The variance of the CMF estimate was not used in the development of the CMFunction.

Similar to the work De Pauw et al., (2014), Juneyoung Park and colleagues recently published four papers that used empirical Bayes before-after methods to develop CMFs and then use the results of the

CMFs to estimate CMFunctions. Park et al., (2014), estimated CMFunctions for determining the safety effect of shoulder rumble strips and widening of shoulders on rural multilane roadways. Park and Abdel-Aty (2015) followed a somewhat similar approach in estimating CMFunctions for shoulder rumble strips and shoulder widening on rural two lane roadways. Park et al., (2015a), estimated CMFunctions for adding bike lanes for urban arterials. Park et al., (2015b), estimated CMFunctions for widening urban roadways. In these papers, Park et al. explored different functional forms for the CMFunctions and in the process also investigated both simple and complex CMFunctions. Some of the papers also considered the possibility that the safety effect of a treatment may change over time. However, it is not clear if variance of the individual CMFs were considered in the estimation of the CMFunctions.

Sacchi et al., (2014, 2015) used full Bayes (FB) before-after intervention models to estimate the CMFs for individual sites and use these results to estimate CMFunctions. The natural logarithm of the CMF was the dependent variable. The variance of the CMF was explicitly considered in the CMF estimation (the variance of the CMF was estimated using MCMC simulation). An important focus of these two studies was to account for the possibility that the safety effect of a treatment may change over time.

Implications for this Study

It is clear that further work is needed in the area of CMFunctions especially for the development of such functions using disaggregate data from the results of a before-after EB evaluation. It is unclear if the recent studies conducted by Park et al. make use of the variance of the CMF estimates (as recommended in the series of studies by Elvik), since that issue is not specifically discussed in the paper.

STATISTICAL MODELING APPROACHES FOR ESTIMATING CRASH MODIFICATION FUNCTIONS

With empirical Bayes before-after studies, the equations for the CMF and the standard error of the CMF are the following (Hauer, 1997):

$$CMF_i^* = \frac{\lambda_i}{\pi_i} \quad \text{.....(1)}$$

$$CMF_i = \frac{\frac{\lambda_i}{\pi_i}}{1+\frac{Var(\pi_i)}{\pi_i^2}} \quad \text{.....(2)}$$

$$Var(CMF_i) = \frac{CMF_i^2\left[\frac{Var(\lambda_i)}{\lambda_i^2}+\frac{Var(\pi_i)}{\pi_i^2}\right]}{\left[1+\frac{Var(\pi_i)}{\pi_i^2}\right]^2} \quad \text{.....(3)}$$

Where, CMF_i^* is the biased estimate of the CMF for a particular site i, CMF_i is the unbiased estimate of the CMF, λ_i is the actual number of crashes in the after period, and π_i is the expected number of crashes in the after period had the treatment not been implemented. The unbiased estimate is different from the biased estimate because the expected value of the ratio of two random numbers is

not the same as ratio of their expected values [i.e., if A and B are two random numbers, $E(\frac{A}{B}) \neq E(A)$/E(B)]. *Var* represents the variance of these parameters.

The traditional approach for estimating CMFunctions includes the use of the CMF value as the dependent variable and site/treatment characteristics as independent variables. One way to express this is as well follows:

CMF = f(site characteristics)…….(4)

Where, f represents a generic function.

This CMFunction could then be estimated as a regression equation. Based on Elvik's recommendation, variance of the CMF needs to be considered in this estimation. The inverse of the variance is typically introduced as a weight in a weighted regression model. In other words, for an observation (or site) whose CMF is CMF_i with a variance of $Var(CMF_i)$, the weight will be $1/Var(CMF_i)$. For linear regression, this would be appropriate.

Some recent studies have recommended the use of a different model form such as a lognormal model that would ensure the predicted CMF from a CMFunction would always be greater than zero. In the case of the log-normal model, Bonneson (2015) showed that the appropriate weight for a weighted log-normal regression model would instead be $[CMF_i/Var(CMF_i)]$ (this is because, based on equation 3, the $Var(CMF_i)$ is not independent of CMF_i, i.e., lower CMFs values would tend to have lower variances as well).

For either the normal regression or lognormal regression models with weights, reliable estimates of CMFs and their variances are needed. In order to have reliable estimates of these parameters, sites with similar characteristics are often combined. However, this aggregation can lead to loss of useful information.

In this study, a new approach is proposed in order to address the possible disadvantage with grouping of sites. For this approach, equation 2 is rewritten as follows:

$$CMF_i = \frac{\lambda_i}{\mu_i\left(1+\frac{Var(\pi_i)}{\pi_i^2}\right)} \quad \text{……..(5)}$$

Following equations 1, 4, and 5,

$$CMF_i^* = \frac{\lambda_i}{\pi_i} = f(site\ charactersitics) \text{……..(6)}$$

$$CMF_i = \frac{\lambda_i}{\mu_i\left(1+\frac{Var(\pi_i)}{\pi_i^2}\right)} = f(site\ characteristics) \quad \text{……..(7)}$$

Equations 6 and 7 can be rewritten as follows:

$\lambda_i = \pi_i \times f(site\ characteristics)$ ……(8)

$\lambda_i = \pi_i \left(1 + \frac{Var(\pi_i)}{\pi_i^2}\right) \times f(site\ characteristics)$ ……(9)

Written in the form of equations 8 and 9, it is possible to estimate this model as a count data model with λ as the dependent variable. Based on equation 9, the offset will be $\pi \left(1 + \frac{Var(\pi)}{\pi^2}\right)$, and based on equation 8, the offset will be just π (a somewhat similar approach was investigated by Bonneson and Pratt, 2008, but for deriving CMFs from cross-sectional regression, not before-after evaluation). The site characteristics will serve as independent variables. Negative binomial regression is usually the most appropriate option since there crash data are typically overdispersed. Statistically, it is unclear if the offset should be just π or $\pi \left(1 + \frac{Var(\pi)}{\pi^2}\right)$ since the $\left(1 + \frac{Var(\pi)}{\pi^2}\right)$ term is introduced in the denominator in equation 2 to account for the bias when two random variables are divided (Hauer, 1997). Another issue with this approach is that the offset is not a fixed measured value, but estimated as part of the EB procedure with a variance. There has been some limited research on the implications of errors/variance in the independent variables, but further research is needed, possibly using simulation (e.g., see Weed and Barros, 1987).

In summary, three model forms were explored in this study:

- Model Form 1. Linear regression with CMF as the dependent variable. For the weighted option, the inverse of variance of the CMF was used as the weight.
  The loglikelihood (LL) for linear regression is the following:
  $LL = -\frac{1}{2}\left[\frac{w_i(y_i-\mu_i)^2}{\phi} + log\left(\frac{\phi}{w_i}\right) + log2\pi\right]$……(10)

- Model Form 2. Log normal regression where log(CMF) is the dependent variable. For the weighted option, as discussed earlier, the ratio of the CMF to its variance was included as the weight. The LL for lognormal regression is the following:

  $LL = -\frac{1}{2}\left[\frac{w_i(\log(y_i)-\mu_i)^2}{\phi} + log\left(\frac{\phi}{w_i}\right) + log2\pi\right]$……(11)

- Model Form 3. Negative binomial regression with the observed crashes in the after period as the dependent variable. The offset based on equation 9. Future research could consider including offsets based on equation 8 as well. The LL for negative binomial regression is as follows:
  $LL = y_i log\left(\frac{k\mu_i}{w_i}\right) - \left(y_i + \frac{w_i}{k}\right) log\left(1 + \frac{k\mu_i}{w_i}\right) + log\left[\frac{G\left(y_i + \frac{w_i}{k}\right)}{G\left(\frac{w_i}{k}\right)G(y_i+1)}\right]$……(12)

The functional form for the negative binomial regression model was the typical log-linear form (investigation of other forms could be a topic for future research).

In equations 10, 11, and 12, y_i refer to the observed values of the dependent variable, μ_i is the predicted value, w_i is the weight, ϕ is the variance parameter to be estimated, and π is a constant. In Model 3, *k* refers to the overdispersion parameter, and *G* refers to the gamma function. The weights for model forms 1 and 2 were discussed earlier. Based on equation 3, the weights for the CMF for a particular site cannot be estimated if the number of crashes in the after period $(i.e., \lambda_i)$ is zero for that site. This is one of the reasons model forms 1 and 2 are usually implemented with aggregated data. On the other hand, model form 3 can be implemented with both aggregate and disaggregate data.

At this time, it is not clear if a weight is needed for model form 3. In traditional SPFs that are estimated for roadway segments, Hauer (2001) indicates that if weight is not used (i.e., if w = 1, for all the sites), then shorter sections have an inordinate influence on the results, and suggests that segment length be used as the weight. However, this is a different context where the negative binomial regression model is used for estimating CMFunctions rather than predicting crash frequency as a function of site characteristics. Future research could investigate the appropriate use of weights in this context.

DATA

In order to obtain data sets to explore the development of CMFunctions, states that attended in the FHWA CMF low cost pooled fund TAC meeting in June 2014 were contacted (over 35 states participate in this annual meeting in Washington, DC). The intent was to obtain data from at least one common treatment that would provide sufficient data to investigate the different options for estimating CMFunctions. At that time, HSRC was completing a study for NCDOT to evaluate the safety of left turn lanes when stop controlled intersections were converted to signalized intersections. So, HSRC staff spoke to the other states about the possibility of data on stop to signal conversions in their states. During the meeting, four states (Ohio, Utah, Kentucky, and California), agreed to review their project files to investigate the possibility of providing us with data on stop to signal conversions. Following the meeting, Ohio and Kentucky indicated that they do not consistently keep records of such conversions. For Utah, Professor Grant Schultz from Brigham Young University provided data from a before-after full Bayes evaluation of stop to signal conversions. However, the intersections used in this evaluation did not include information on minor road AADT. Since a primary reason to implement stop to signal conversions is to reduce angle crashes, and angle crashes are a function of both major and minor road traffic volumes, the HSRC team felt that a data set without minor road AADT would not be very useful in this study. In early Fall 2015, California provided location information for four locations where stop to signal conversions were made. Since the sample from California was limited and the data from Utah did not include information on minor road AADT, the data from North Carolina were used in this study.

North Carolina Data

The data from North Carolina was based on a recent evaluation conducted by Srinivasan et al., (2014) for the North Carolina Department of Transportation. The evaluation was based on 117 intersections where traffic signals were installed; all these intersections were controlled by stop signs on the minor road approaches before the signals were implemented. None of these 117 intersections had any left

turn lanes before signalization. During signalization, at least one left turn lane was added at 67 of these intersections; at the remaining 50 intersections, left turn lanes were not added. Among the 67 intersections where at least one left turn lane was added, 19 were 3-leg intersections and 48 were 4-leg intersections. The summary statistics and the CMFs from the study are provided in the following sections.

Summary Statistics

Srinivasan et al., (2014) conducted an empirical Bayes before-after evaluation. Tables 1 and 2 provide summary statistics for the 117 treatment sites that were used in the evaluation. Separate tables are provided for 3 and 4 leg intersections. Two of the treatment intersections had no crashes in the before period. Among the 67 intersections where at least one left turn was added, one had negative offset left turn lanes. None of them had positive offsets.

Srinivasan et al., (2014) evaluated five crash types: Total, Injury and fatal, Rear end, Type 1 frontal impact, and Type 2 frontal impact crashes. Frontal impact crashes (type 1) included the following crash types:

- Left turn same roadway
- Left turn different roadway
- Angle

Frontal impact crashes (type 2) included the following crash types:

- Left turn same roadway
- Left turn different roadway
- Angle
- Right turn same roadway
- Right turn different roadway
- Sideswipe opposite direction
- Head-on

Crash Modification Factors

Following is a summary of the crash modification factors that were estimated based on the 117 intersections that were included in the evaluation (further information about these CMFs including CMFs for intersections without and without left turn lanes is available in Srinivasan et al, 2014):

- Total crashes: CMF (S.E.) = 0.590 (0.019)
- Injury and fatal crashes: CMF (S.E.) = 0.541 (0.027)
- Rear end crashes: CMF (S.E.) = 0.906 (0.052)
- Frontal impacts crashes (Type 1): CMF (S.E.) = 0.401 (0.020)
- Frontal impact crashes (Type 2): CMF (S.E.) = 0.440 (0.020)

Table 1: Summary statistics for 3 leg treatment sites (36 intersections)

Variable	Signalization without addition of left turn lanes (17 sites)			Signalization with addition of at least one left turn lane (19 sites)		
	Min.	Max.	Mean	Min.	Max.	Mean
Years before	5	5	5	5	5	5
Years after	4	5	4.88	3	5	4.79
Total Crashes/site-year before	0	4.4	2.38	0.6	10.4	3.94
Total Crashes/site-year after	0	5.4	1.86	0.6	7.4	2.55
Injury & Fatal Crashes/site-year before (KABC)	0	2.2	1.05	0.2	4.2	1.8
Injury & Fatal Crashes/site-year after (KABC)	0	1.8	0.76	0	2.6	0.81
Rear End Crashes/site-year before	0	1.4	0.64	0	5	1.61
Rear End Crashes/site-year after	0	3.2	1.01	0.2	3.2	1.16
Type 1 Frontal Impact Crashes/site-year before	0	1.8	1.01	0.4	5.8	1.65
Type 1 Frontal Impact Crashes/site-year after	0	1.2	0.43	0	3.8	0.82
Type 2 Frontal Impact Crashes/site-year before	0	2	1.12	0.6	6	1.77
Type 2 Frontal Impact Crashes/site-year after	0	1.2	0.51	0.2	4.4	0.98
Major road AADT before	3475	14539	8150	2981	15107	9518
Major road AADT after	3907	18025	8307	3870	18248	10820
Minor road AADT before	986	5871	3671	1852	13880	5686
Minor road AADT after	972	6829	3777	3104	13880	6255
Intersection AADT before	6130	16336	11821	8341	25421	15204
Intersection AADT after	6110	20247	12084	8880	32129	17075

Table 2: Summary statistics for 4 leg treatment sites (81 intersections)

Variable	Signalization without addition of left turn lanes (33 sites)			Signalization with addition of at least one left turn lane (48 sites)		
	Min.	Max.	Mean	Min.	Max.	Mean
Years before	2	5	4.79	4	5	4.96
Years after	2	5	4.76	1	5	4.75
Total Crashes/site-year before	0.2	8.6	4.41	0	10.2	4.6
Total Crashes/site-year after	0	6.6	2.64	0	7.4	2.78
Injury & Fatal Crashes/site-year before (KABC)	0	4.6	2.33	0	6	2.42
Injury & Fatal Crashes/site-year after (KABC)	0	2.6	1.19	0	4	1.13
Rear End Crashes/site-year before	0	2	0.59	0	3	0.95
Rear End Crashes/site-year after	0	2.4	0.93	0	4	1
Type 1 Frontal Impact Crashes/site-year before	0.2	7.2	3.22	0	8.2	3.07
Type 1 Frontal Impact Crashes/site-year after	0	2.8	1.25	0	4.6	1.13
Type 2 Frontal Impact Crashes/site-year before	0.2	7.4	3.35	0	8.2	3.2
Type 2 Frontal Impact Crashes/site-year after	0	3	1.32	0	5	1.38
Major road AADT before	2480	14805	5947	1360	14309	7869
Major road AADT after	2680	17566	6729	1467	15500	9241
Minor road AADT before	746	5463	2823	1036	8884	3633
Minor road AADT after	1014	5803	3295	1063	8537	4360
Intersection AADT before	4624	17412	8770	5325	18906	11502
Intersection AADT after	4394	19573	10023	5770	22392	13601

CRASH MODIFICATION FUNCTIONS

The models for the CMFunctions were estimated using PROC GLIMMIX in SAS for total and injury and fatal crashes. Since model form 2 is lognormal and model form 3 is negative binomial with a log-link, the predicted value from the model forms 2 and 3 can be obtained by taking the exponent of the results provided by SAS, i.e., the predicted value is the following:

$$\exp\{\alpha + \beta_1 * X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + ...\},$$

where α is the intercept, the *X*s are the independent variables, and the β s are the coefficients of the independent variables.

On the other hand, the predicted value from model form 1 is just the following:

$$(\alpha + \beta_1 * X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + ...).$$

Aggregation

In order to estimate CMFunctions using model forms 1 and 2, the data from the results of the before-after EB method were aggregated. As mentioned earlier, the intent of the aggregation was to have a sufficient sample of sites in each group so that a reliable CMF and the standard error of the CMF can be estimated for each group. Different options were explored. Starting with the 117 treatment sites, the final grouping was based on the following variables:

- Major road AADT before treatment (3 categories): 0 – 5,000, 5,001 – 10,000, >10,000
- Speed limit on the major road (2 categories): Less than or equal to 50 mph, > 50 mph
- Number of legs (2 categories): 3 leg and 4 leg
- Left turn added during treatment (2 categories): Yes, No

With this grouping, there is a possibility of a maximum of 24 groups (i.e., 3 X 2 X 2 X 2). However, two of these groups did not have any sites. So, the information from the 117 treatment sites was collapsed into 22 groups.

Measures of Comparison

In order to compare the performance of the different model forms, various performance measures (or goodness of fit) were investigated, including log-likelihood, Akaike's information criterion (AIC), and Bayesian information Criterion (BIC). However, since the model forms are different, it was unclear if these GOF based on the likelihood were the most appropriate. Hence, other measures were considered including mean absolute deviation (MAD) and mean square error (MSE). In addition, there was a need to determine how the well a CMFunction without covariates (i.e., with just an intercept term) predicted the CMF estimated using the EB method; this could be called as *boundary condition check*.

Results

Results are provided below for both the boundary condition check and multivariable CMFunctions to investigate the effect of site characteristics on the CMF.

Boundary condition check based on intercept only models

The results for the boundary condition check (based on a CMFunction with just an intercept term) are presented in Table 3 along with the estimated CMF from EB before-after evaluation for total and injury and fatal crashes. Ideally, the predicted CMF should be very close to the estimated CMF from the EB before-after evaluation (the estimated CMF was 0.590 for total crashes and 0.541 for injury and fatal crashes). The absolute value of the difference between the predicted CMF and estimated CMF is also provided.

Based on Table 3, it is clear that for total crashes, model form 2 provides the CMF predictions that are closest to the actual CMF value. This is followed by model form 3 and then model form 1. The prediction from model form 3 for the disaggregate data is quite close to the prediction from model form2. For injury and fatal crashes, the prediction from model form 2 for weighted regression and the prediction from model form 3 for the disaggregate data are closest to the actual CMF value. For both

crash types, the prediction from model form 3 for the grouped data performs better than the predictions from model form 1.

Multivariable CMFunctions

Following the review of the boundary condition, multivariable CMFunctions were estimated for these different model forms. The following independent variables (site characteristics) were included in the models.

- Average major road AADT in the before period
- Average minor road AADT in the before period
- Average total intersection AADT in the before period
- Average EB estimate of the expected crashes in the before period
- Number legs (3 legs versus 4 legs)
- Whether at least one left turn lane was added during signalization (Yes or No)
- Speed limit on the major road (Less than or equal to 50 mph, > 50 mph)

Table 3: CMF predictions based on intercept only CMFunction

Actual and Predicted CMFs	Model Form	Aggregate or Disaggregate; Weighted/Unweighted; Observations	Total Crashes		Injury and Fatal Crashes	
			CMF value	Difference from actual CMF (absolute value)	CMF value	Difference from actual CMF (absolute value)
Actual CMF from EB before-after evaluation			0.590	----	0.541	----
Predicted CMF	Model Form 1	Aggregated; Unweighted; 22 observations	0.615	0.025	0.625	0.084
Predicted CMF	Model Form 1	Aggregated; Weighted; 22 observations	0.539	0.051	0.464	0.077
Predicted CMF	Model Form 2	Aggregated; Unweighted; 22 observations	0.587	0.003	0.573	0.032
Predicted CMF	Model Form 2	Aggregated; Weighted; 22 observations	0.595	0.005	0.552	0.011
Predicted CMF	Model Form 3	Aggregated; Unweighted; 22 observations	0.612	0.022	0.581	0.040
Predicted CMF	Model Form 3	Disaggregated; Unweighted; 117 observations	0.598	0.008	0.529	0.012

For the models with the aggregated data, the AADT and the EB estimate values were averaged for a particular group. In the model development, the independent variables that were not statistically significant at the 0.05 level were removed in a stepwise manner. The final model only included independent variables that were statistically significant at the 0.05 level. The results of the grouped models (for all three model forms) are presented separately from the results of the disaggregate data (just for model form 3), since the GOF statistics such as MAD and MSE cannot be compared for aggregate and disaggregate data.

Results for CMFunctions from aggregate data

Table 4 shows the CMFunctions estimated using aggregate data for total crashes and Table 5 shows the CMFunctions estimated using aggregate data for injury and fatal crashes. The parameter estimates are shown along with the standard errors and GOF statistics. For the total crash CMFunctions, the MSE values are very close to each other, but the MAD values are slightly better for model form 1. The GOF for model form 3 and model form 2 with weighted regression are very close to each other. In the case of the injury and fatal crashes, model form 3 and model form 2 with weighted regression are clearly better than the other model forms. This is an important finding since not all crashes result in an injury or fatality. In this data set, injury and fatal crashes represent about 50% of total crashes before the intersections were signalized (see Tables 1 and 2). This finding along with the results from Table 3 could imply that when CMFunctions are estimated for crash types that are less frequent, model forms 2 and 3 may be more reliable. Overall, the results indicate that the following:

- CMFs seem to increase with increase in AADT values
- CMFs are higher for intersections where left turn lanes were not added (this is consistent with the results reported in Srinivasan et al., 2014)
- CMFs are higher for intersections where the major speed limit is higher than 50 mph

More importantly, model form 3 seems a reasonable substitute for the traditional model forms 1 and 2.

Results for CMFunctions from disaggregate data

Table 6 shows the results for ungrouped data based on model form 3. Unlike the models based on aggregated data, the number of legs is statistically significant for the total crash models, and implies that the CMF is lower at 3 leg intersections (i.e., the treatment is more effective at 3 leg intersections compared to 4 leg intersections). In addition, an increase in EB expected crashes per year in the before period is associated with a decrease in the CMF, i.e., the treatment is more effective at sites with higher expected crashes per year in the before period. These two variables (number of legs and expected crashes in the before period) were not significant in any of the models using the grouped data. This indicates the value in using the disaggregate data in estimating the CMFunctions. Model form 3 allows estimation of CMFunctions with disaggregate data.

Table 4: CMFunctions for Total Crashes (Aggregated Data)

Variable	Model 1 (unweighted) Normal	Model 1 (weighted) Normal	Model 2 (unweighted) Lognormal	Model 2 (weighted) Lognormal	Model 3 Negative binomial
	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)
Intercept	-2.9040 (0.9603)	-1.2934 (0.5068)	-6.2440 (1.4114)	-4.3143 (0.9852)	-4.4405 (0.8815)
ln(Major road AADT)		0.1950 (0.0560)		0.3978 (0.1083)	0.4121 (0.0969)
Minor road AADT					
ln(Intersection AADT)	0.3597 (0.1017)		0.5840 (0.1494)		
Legs = 3					
Legs = 4 (reference level)					
No turn lane added	0.1871 (0.0660)	0.1472 (0.0517)	0.3146 (0.0971)	0.2802 (0.0875)	0.2755 (0.0844)
Turn lane added (reference level)					
Speed limit > 50 mph	0.1351 (0.0647)	0.1377 (0.0473)	0.2018 (0.0951)	0.2558 (0.0833)	0.2469 (0.0828)
Speed limit $\leq$ 50 mph (reference level)					
k/scale	0.0227	1.5489	0.0491	1.5596	0.0170
MAD	0.105	0.105	0.106	0.110	0.109
MSE	0.019	0.020	0.019	0.020	0.020

Table 5: CMFunctions for Injury and Fatal Crashes (Aggregated Data)

Variable	Model 1 (unweighted) Normal	Model 1 (weighted) Normal	Model 2 (unweighted) Lognormal	Model 2 (weighted) Lognormal	Model 3 Negative binomial
	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)
Intercept				-5.6344 (2.1442)	-5.9765 (2.1284)
ln(Major road AADT)	0.0705 (0.0062)	0.0525 (0.0039)			
Minor road AADT					
ln(Intersection AADT)			-0.0593 (0.0101)	0.5044 (0.2263)	0.5531 (0.2255)
Legs = 3					
Legs = 4 (reference level)					
No turn lane added				0.4473 (0.1353)	0.4914 (0.1422)
Turn lane added (reference level)					
Speed limit > 50 mph				0.2652 (0.1255)	
Speed limit $\leq$ 50 mph (reference level)					
k/scale	0.0677	2.1076	0.1939	1.5709	0.0458
MAD	0.211	0.233	0.214	0.153	0.152
MSE	0.065	0.090	0.072	0.049	0.050

Table 6: CMFunctions for disaggregate data (model form 3)

Variable	Total Crashes	Injury and Fatal Crashes
	Estimate (S.E.)	Estimate (S.E.)
Intercept	-4.1053 (0.9731)	-3.2146 (1.1656)
ln(Major road AADT)	0.4059 (0.1092)	0.2875 (0.1276)
Minor road AADT/1000	0.0722 (0.0246)	
ln(Intersection AADT)		
EB expected crashes per year	-0.1143	-0.1445 (0.0524)
Legs = 3	-0.2813 (0.1205)	
Legs = 4 (reference level)		
No turn lane added	0.2543 (0.0940)	0.3447 (0.1145)
Turn lane added (reference level)		
Speed limit > 50 mph	0.2238 (0.0878)	0.2923 (0.1110)
Speed limit $\leq$ 50 mph (reference level)		
k	0.1245	0.1195
MAD	0.244	0.291
MSE	0.114	0.154

SUMMARY AND CONCLUSIONS

The main goal of this study was to investigate different model forms for estimating CMFunctions using the results from a before-after EB evaluation. Three different model forms were explored including two traditional approaches, normal regression (model form 1) and lognormal regression (model form 2), and a new negative binomial regression approach (model form 3). With the traditional approaches, the dependent variable is the CMF for a particular site (or group of sites), and sites are usually grouped (or aggregated) in order to obtain a stable estimate of the CMF and the standard error of the CMF. With the new negative binomial regression approach, the numerator of the CMF is used as the dependent variable and the denominator of the CMF is used as an offset. The negative binomial regression

approach does not require the aggregating of data, and could provide more insights that may be lost due to the aggregation.

The project team sought data from multiple states in order to compare the performance of these different types of CMFunctions. Finally, data from the results of a before-after evaluation conducted for North Carolina Department of Transportation were used for comparing the results from the three different approaches for estimating CMFunctions. The treatment was the implementation of traffic signals at intersections that were controlled by stop signs on the minor roads.

First, the data were aggregated and all CMFunctions were estimated using the three model forms. For the first two model forms, CMFunctions were estimated with and without weights. With the aggregated data, the results from model form 3 compare quite favorably with that of the traditional model forms 1 and 2. Then, CMFunctions based on model form 3 were estimated using the original results from the before-after evaluation (i.e., without aggregation). The models using disaggregate data included independent variables that were not significant in the models based on the aggregated data, indicating the value of using model form 3 to estimate CMFunctions using disaggregate data.

REFERENCES

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

Bonneson, J. (2015), *Local adjustment of CMFs based on crash distribution*, Working Paper 3, NCHRP Project 17-63.

Bonneson, J., and Pratt, M.P. (2008), Procedure for developing accident modification factors from cross-sectional data, *Transportation Research Record* 2083, pp. 40-48.

Carter, D., Srinivasan, R., Gross, F., and Council, F. (2012), *Recommended protocols for developing crash modification factors*, NCHRP Project 20-07 (Task 314), National Cooperative Highway Research Program, Washington, D.C.

De Pauw, E., Daniels, E., Brijs, T., Hermans, E., and Wets, G. (2014), Safety effects of an extensive black spot treatment programme in Flanders-Belgium, *Accident Analysis and Prevention*, Vol. 66, pp. 72-79.

Elvik, R. (2005a), Introductory guide to systematic reviews and meta-analysis, *Transportation Research Record* 1908, pp. 230–235.

Elvik, R. (2005b), Speed and road safety: synthesis of evidence from evaluation studies, *Transportation Research Record* 1908, pp.59–69.

Elvik, R. (2009), Developing accident modification functions: exploratory study, *Transportation Research Record* 2103, pp. 18–24.

Elvik, R. (2011), Developing an accident modification function for speed enforcement, *Safety Science*, Vol. 49, 920–925.

Elvik, R. (2015), Methodological guidelines for developing accident modification functions, *Accident Analysis and Prevention*, Vol. 80, pp. 26-36.

FHWA (2014), *Crash Modification Factors in Practice*, Report FHWA-SA-13-017, http://safety.fhwa.dot.gov/tools/crf/resources/cmfs/docs/product_summary_final.pdf.

Gross, F., Persaud, B., and Lyon, C. (2010), *A guide to developing quality crash modification factors*, Report FHWA-SA-10-032, Federal Highway Administration.

Hauer, E. (1997), *Observational Before-After Studies Road Safety*, Pergamon Press.

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

Park, J., Abdel-Aty, M., and Lee, C. (2014), Exploration and comparison of crash modification factors for multiple treatments on rural multilane roadways, *Accident Analysis and Prevention*, Vol. 70, pp. 167-177.

Park, J., Abdel-Aty, M., Lee, J., and Lee, C. (2015a), Developing crash modification functions to assess safety effects of adding bike lanes for urban arterials with different roadway and socio-economic characteristics, *Accident Analysis and Prevention*, Vol. 74, pp. 179-191.

Park, J., and Abdel-Aty, M. (2015), Development of adjustment functions to assess combined safety effects of multiple treatments on rural two-lane roadways, *Accident Analysis and Prevention*, Vol. 75, pp. 310-319.

Park, J., Abdel-Aty, M., Wang, J.-H., and Lee, C. (2015b), Assessment of safety effects for widening urban roadways in developing crash modification functions using nonlinearizing link functions, *Accident Analysis and Prevention*, Vol. 79, pp. 80-87.

Sacchi, E., Sayed, T., and El-Basyouny, K. (2014), Collision modification functions: Incorporating changes over time, *Accident Analysis and Prevention*, Vol. 70, pp. 46-54.

Sacchi, E., Sayed, T., and Osama, A. (2015), Developing crash modification functions for pedestrian signal improvement, *Accident Analysis and Prevention*, Vol. 83, pp. 47-56.

Srinivasan, R., Lan, B., and Carter, D. (2014), Overdispersion in modelling accidents on road sections and in empirical bayes estimation, Report FHWA/NC/2013-11, North Carolina Department of Transportation.

Weed, R.M. and Barros, R.T. (1987), Demonstration of regression analysis with error in the independent variable, *Transportation Research Record* 1111, TRB, Washington, DC.