Improving functional form in cross-sectional regression studies to capture the non-linear safety effects of roadway attributes - Freeway median width case study

(Accepted manuscript for article: Jafari Anarkooli A., Persaud B. and C. Lyon. Improving functional form in cross-sectional regression studies to capture the non-linear safety effects of roadway attributes - Freeway median width case study. Accident Analysis and Prevention 156, p. 106130, June 2021. https://doi.org/10.1016/j.aap.2021.106130)

ABSTRACT

Crash modification factors (CMFs) for several roadway attributes are based on cross-sectional regression models, in the main because of the lack of data for the preferred observational before-after study. In developing these models, little attention has been paid to those functional forms that reflect the reality that CMFs should not be single-valued, as most available ones are, but should vary with application circumstance. Using a full Bayesian Markov Chain Monte Carlo (MCMC) approach, this study aimed to improve the functional forms used to derive CMFs in cross-sectional regression models, with a focus on capturing the variability inherent in crash modification functions (CMFunctions). The estimated CMFunction for target crashes for freeway median width, used for a case study, indicates that the approach is capable of developing a function that can capture the logical reality that the CMF for a given change in a feature's value depends not only on the amount of the change but also on the original value. The results highlight the importance of using the functional forms that can capture non-linear effects of road attributes for CMF estimation in cross-sectional models. The case study provides credible CMFs for assessing the safety implications of decisions on freeway median width that could be used in improving current design practice.

INTRODUCTION

The methods used for crash modification factor (CMF) estimation can be generally classified as either before-after or cross-sectional studies. The before–after study compares the safety of a treatment after implementation with what would have been expected without it. On the other hand, the cross-sectional study is based on a single time period, such that the ratio of expected crash frequencies for sites with and without a feature (or a change in its value) is assumed to be the safety impact of that feature (or of that change in its value).

There is a general consensus among researchers that well-designed empirical Bayes (EB) before–after studies provide the most reliable estimations (Persaud and Lyon, 2007; Hauer, 1997). However, the before-after design can be challenging when sample sizes are inadequate, where there is site selection bias leading to a regression-to-the-mean effect, or where more than one change has been implemented at treatment sites (Wu et al., 2015). The fundamental issue with the cross-sectional design is that the comparison is between two distinct groups of sites. As such, as noted by Gross et al. (2010), the observed difference in crash experience can be due to factors other than the feature of interest. Where factors such as traffic volume or geometric characteristics are known, they can be controlled for by estimating a multiple variable regression model and inferring the CMF for a feature from its coefficient. However, such a CMF can also reflect unknown, or known but unmeasured, factors that are not included in the model, as well as correlation with one or more of the other variables in the model. For these reasons, caution needs to be exercised in making inferences about CMFs derived from cross-sectional designs, as Gross et al. note in recommending that the quality of CMFs so derived be assessed using a number of criteria. These include:

- Does the direction of effect (i.e., expected decrease or increase) in crashes meet expectations?
- Does the magnitude of the effect seem reasonable?
- Are the parameters of the model estimated with statistical significance?
- Do different cross-section studies come to similar conclusions?
- Do different before-after studies come to similar conclusions?

Regarding the last criterion, it has been suggested that cross-sectional models can yield varying results across studies (Hauer, 2013). However, several comparative studies have supported the applicability of this method in that the results are consistent with the EB before-after method. For example, Wood et al. (2015) obtained similar CMFs for Safety Edge paving from the two methods. As another example, Wu et al. (2015) investigated the issue by assessing CMFs for lane width, curve density, and pavement friction. Their results suggested that CMFs derived from the cross-sectional method is unbiased when all factors affecting traffic safety are identical in all segments, except those of interest. They concluded that the accuracy of the estimated CMFs can be acceptable as long as major contributing factors are not omitted from the cross-sectional regression models.

Cross-sectional studies are particularly useful, and are actually necessary, when the number of sites where the treatment is implemented is limited. This is commonly the case for the treatments that are on a stretch of highway (e.g., shoulder width, median width, etc.) and where there are many variations in the extent of the treatment (e.g., the several possible values for the amount of widening for a shoulder or median). For instance, as mentioned in Gross et al. (2010), there may be few instances that shoulder width has been actually widened from 4 ft. to 6 ft., which makes it impossible to draw reliable conclusions about the safety effects of this specific shoulder widening option. However, there are likely sufficient sites with shoulder widths of 4 ft. and 6 ft. to make a cross-sectional model a practical choice for estimating a CMF for shoulder widening from 4 ft. to 6 ft.

Despite the necessity of deriving CMFs from cross-sectional regression models, the main body of literature in developing SPFs has been focused on models intended for crash prediction and on the strengths and weaknesses of the various methodological approaches (Lord and Mannering, 2010, Yannis et al., 2017). As such, limited attention has been paid to the consideration of functional forms (Kononov et al., 2011) that would better facilitate CMF estimation. If the functional form intended for such purposes is inappropriate, the coefficients obtained in the regression models have no clear meaning, which, in turn, may lead to the incorrect estimation of CMFs (Hauer, 2004).

To develop a cross-sectional regression model for deriving CMFs, the common approach is estimating a generalized linear model (GLM) (McCullagh, 2018) that assumes a log-linear relationship between crash counts (dependent variable) and site traits (independent variables) (Gross et al., 2010). However, when this functional form is used for CMF estimation, the same amount of a change in the value of a feature of interest will always produce a fixed change in safety, regardless of the original value of the feature.

To illustrate, consider the typical SPF for road segments that uses the following commonly used functional form:

$$N_i = L_i \times \text{Exp}(\beta_0) \times AADT_i^{\beta_1} \times \text{Exp}(\sum_{j=2}^{n} \beta_j x_{ij}) \quad (1)$$

where N_i is the expected number of crashes per unit of time for segment i, L_i is the segment length, x_{ij} are the factors that influence the probability of crashes, and β_j are the regression coefficients for these factors. Assuming that the SPF in Equation (1) includes median width as one of the x_j, then the safety effects of changing the width of a median from the width of w_1 to w_2 is obtained from Equation (2) which shows, illogically, that the estimated CMF is only dependent on the difference between the two widths but not on the values of the widths themselves.

$$\text{CMF} = \frac{L \times \text{Exp}(\beta_0) \times AADT^{\beta_1} \times \text{Exp}(\beta_2 w_2)}{L \times \text{Exp}(\beta_0) \times AADT^{\beta_1} \times \text{Exp}(\beta_2 w_1)} = \text{Exp}(\beta_2(w_2 - w_1)) \quad (2)$$

The functional form above, suggesting a linear effect for the safety effect of a treatment, has been widely used in deriving many CMFs, including several in the Highway Safety Manual (HSM) (AASHTO, 2010). However, it stands to reason that the safety effects of many roadway attributes are not linear. The primary objective of this study is to investigate functional forms in cross-sectional studies to derive CMFs such that safety effects implied by variable coefficients are logically non-linear. In doing so, full Bayesian Markov Chain Monte Carlo (MCMC) methods using, importantly, the freely available software WinBUGS (Lunn et al., 2000) are employed for the effect on target crashes of changing freeway median width as a case study.

At first glance, the choice of case study may seem inappropriate since, typically, an existing median is rarely widened in practice for the safety improvement purposes. Nevertheless, in accord with a fundamental application of the methods in the Highway Safety Manual (AASHTO, 2010) CMFs for median width are needed for assessing the safety implications of choices for the value for this feature at the design stage.

The rest of the paper is organized as follows. The next section summarizes the literature relevant to functional form investigations in road safety. The methodology and the data used are described in the third section, followed by sections that present and discuss the modeling results. The final section summarizes the findings and makes suggestions for future research.

LITERATURE REVIEW

Functional Form for Capturing Non-linear Effects

Thanks to the progress in statistical methodologies, researchers have been able to extract more accurate and useful information from crash data sources. However, many of the previous efforts have focused on the statistical distribution of crash data (e.g., negative binomial, Zero-Inflated Poisson, Conwey-Maxwell), with insufficient attention given to the functional forms capturing the non-linear effects of variables on safety.

Hauer et al. (2004) emphasized the importance of functional form in developing a model to estimate crash frequency on urban four-lane undivided roadways. Their results reveal that some variables that are on a stretch of road have non-linear effects on safety. For example, a "U-shape" function was proposed to capture the effect of degree of curve, which was associated with an improvement in safety up to a certain point and an increased risk of crashes for larger values.

Some studies have used neural network models to deal with the non-linearity that may exist between contributing factors and crashes. For example, Xie et al. (2007) developed a Bayesian neural network (BNN) model to predict crashes using data from Texas frontage roads. They conducted sensitivity analyses on the developed BNN model for two sites, and observed that, for instance, lane width had an "inverse U-shape" relation with crash counts at one site. And Kononov et al. (2011) also used neural networks (NNs) to explore the underlying relationship between crashes and other variables for urban freeway segments. The commonly used GLM with an NB error structure was then developed to compare the results with the NNs estimations. The results indicate that a sigmoid shape functional form may be a more reasonable representation of the relationship between traffic flow and crash occurrence on urban freeways.

Another approach to deal with the non-linearity was proposed by Lao et al. (2014) who developed a generalized non-linear model (GNM) that relaxes the assumption of linear relationships in GLMs using piecewise functions. A comparison between GNM and GLM estimations was conducted on rear-end crashes and it was shown that GNMs outperformed GLMs. Moreover, more variables were found significant in GNMs compared to GLMs. In a related effort, Park and Abdel-Aty (2015) investigated the safety impacts of roadside treatments using GLM, GNM, and multivariate adaptive regression splines (MARS) model. It was found that MARS provided better results than the GNMs and the GLMs, this being in accord with the reality that roadside treatments had non-linear effects on crash risk.

Median Width and Safety

There is a considerable body of research that investigated the contribution of median characteristics to the risk of crashes (Donnell and Mason Jr, 2006, Russo and Savolainen, 2018). For example, Donnell and Mason Jr (2006) found that median barrier crash frequency decreases when the median barrier offset from the left-edge of the travel way is increased. Also, many of the previous attempts have examined how installation a particular type of barrier, such as cable median barrier, affects the frequency and severity of crashes. As a case in point, cable median barriers have been generally found to reduce severe crashes (Alluri et al., 2012, Chimba et al., 2017), while they may increase the odds of property damage only and less severe crashes.

The literature regarding how crash experience may vary with median width is inconclusive. Hauer (2000) has summarized the previous knowledge on safety benefits of median width, concluding that the effect of median width on total crashes is questionable. His review suggested that there is a negative association between cross-median crashes and median width and that widening a median may lead to an increase in median-related crashes up to 30 ft. width, with a decreasing trend beyond that.

Results from an NCHRP study (Stamatiadis and Council, 2009) that used a national database for a 12-year period covering 2,387 miles suggest that median width is a significant factor only for multiple vehicle crashes. It was found that for every 1 ft. increase in median width there will be a 1% reduction in multiple vehicle crashes. Notably, this effect is the same regardless of the median width.

To quantify the safety effects of median width for multilane rural highways, the Highway Safety Manual (HSM) (AASHTO, 2010) adopted the crash modification factors (CMF) from Harkey et al. (2008), whose cross-sectional study developed negative binomial (NB) regression models to estimate parameters used to infer CMFs. Their results indicate that both total crashes and cross-median crashes will reduce by increasing median width. However, the estimated CMFs are only dependent on the difference between the two widths, and not on the values of the widths themselves, which seems illogical.

More recently, an NCHRP study (Bonneson et al., 2012) has developed CMFs for freeway crash prediction in the HSM using a non-linear regression procedure (NLMIXED) in the SAS software (Wolfinger, 1999). "Predictive models that are non-linear and discontinuous" were proposed for single vehicle and multiple vehicle crashes. Although their models consider the original width of median, the CMF estimated from these models are only marginally sensitive to the value of median width. For instance, the CMF for multiple vehicle crashes for changing the width from 40 ft. to 60 ft. is 0.938, which is almost the same as the CMF of 0.942 for the same amount of change from 60 ft. to 80 ft. It should be noted that the CMFs estimated pertained to all single and multiple vehicle crashes regardless of whether these crashes were associated with the median.

In sum, the literature confirms the intuitive expectations for non-linear effects of some road attributes on crash frequency. Although there are several research studies aiming to address this issue, there is no consensus on which functional form performs the best for different treatments. This paper presents an alternative to the modeling methodology used in Bonneson et al., (2012) for developing the HSM CMFs. As noted above, CMFs were developed only for single and multiple vehicle crashes and, likely as a result, they showed little sensitivity to original median width. For this study, the modeling is performed using the public domain WinBUGS software (Spiegelhalter et al., 2003) (in contrast to SAS used for Bonneson et al. (2012)) to apply full Bayes (FB) Markov Chain Monte Carlo (MCMC) estimation techniques. This approach, too, allows a functional form can potentially capture the effects of the original median width when changing the width. Another point of departure is the focus on median-related crashes (i.e. ROR crashes that occurred on the left-side of the freeway segment) that are more directly targeted by changes in median width. The results from the FB-MCMC are compared to those from conventional negative binomial GLMs.

Data Summary

The data pertained to freeway segments in Ontario, Canada, and were provided by the Ministry of Transportation of Ontario (MTO). The focus is on the freeway segments with traversable (non-barrier) medians (including the inside shoulders) wider than 10 m. Moreover, it was assumed when the median width is larger than 45 m., the two travel directions serve as completely separate roads, so the few segments with such median widths were excluded. In addition, a few segments had missing values for one or more key attributes; they were also excluded from the analysis. The final database included 315 freeway segments that have the total length of 1,463 km.

The database consisted of two separate datasets, one for crashes and one for roadway attributes. The crash dataset included 3 years (2015-2017) of collision records along with detailed information on each collision, including the time and location of the crash, impact type, initial impact location, severity level, and driver and vehicle information. Median-related crashes were considered to be those crashes with initial impact location either on the left shoulder or beyond it. The roadway attributes dataset consists of the elements of an individual segment. For each year, these two datasets were merged on the basis of crash location, which is identified by a specific LHRS (Linear Highway Referencing System) number. This referencing system uses unique milepost identifiers to refer to the specific location within the road network. When LHRS is combined with a milepost offset, it is possible to reference any position along the road. Table 1 presents the data summary of the freeway segments studied, while Figure 1 shows the distribution of median width in the dataset. As seen, the 10 to 15 m. and 35 to 40 m. ranges have the most and the least frequency in the dataset, respectively.

Table 1. Descriptive statistics of the data

Item	Mean	Min	Max	St. Dev
Total Crashes	12.737	0	93	11.285
Fatal-Injury Crashes	2.854	0	19	2.849
Median Width (m)	16.531	5.5	40	9.122
Left Shoulder (m)	1.541	0	6	1.210
Median and Left Shoulder (m)	19.874	10	43.6	8.863
Average AADT	43595	10008	50365	11606
Median Type	Grass = 231;	Other types = 84		
Number of Lanes	4-lane= 230;	More than 4-lane = 85		

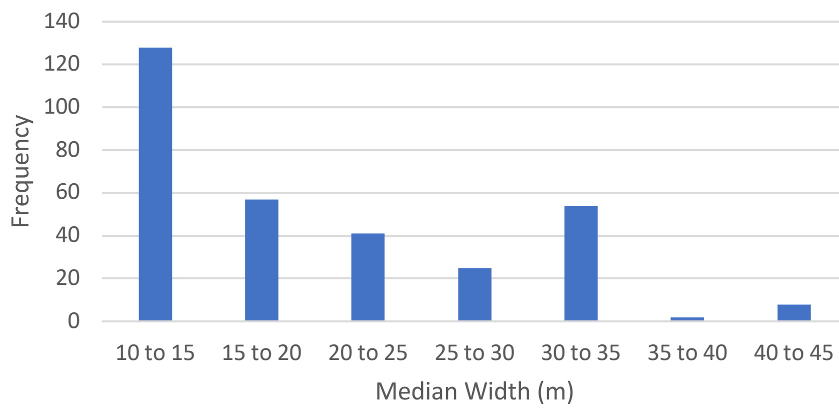


Figure 1. Frequency of segments with different ranges of median width

METHODOLOGY

Negative Binomial (NB) Model

Due to the specific properties of crash counts, including discreteness and nonnegativity, the Poisson and NB regression models have been recognized for some time as appropriate for modeling such data (Abdel-Aty and Radwan, 2000, Caliendo et al., 2007, Lord et al., 2005). In a Poisson regression, the probability of segment i having n_i crashes can be obtained as follows:

$$\mathrm{P}(n_i) = Exp(-\lambda_i)\lambda_i^{\,n_i}/n_i! \qquad (3)$$

where λ_i is the Poisson parameter which equals to the expected number of crashes for segment i, $\mathrm{E}[n_i]$. To ensure that the λ_i holds a non-negative value, the obvious choice is exponential form (Lord and Washington, 2018), so that:

$$\lambda_i = Exp(\beta X_i) \qquad (4)$$

where X_i is the vector of covariates and β denotes the vector of the coefficients to be estimated. In the frequentist approach, the model is estimated through maximum likelihood function given by:

$$L(\beta) = \prod_i \frac{Exp[-Exp(\beta X_i)][Exp(\beta X_i)]^{n_i}}{n_i!} \qquad (5)$$

The Poisson model restricts the mean, E$[n_i]$, to be equal to the variance of the distribution, var$[n_i]$. However, in crash data this restriction rarely if ever holds. The specification of an NB distribution addresses this restriction by adding an error term, ζ, to Equation (4):

$$log\lambda_i = \beta X_i + \zeta_i \qquad (6)$$

where Exp(ζ_i) is gamma-distributed which has a mean of 1 and variance of γ, an additional estimable parameter. The variance in the NB model is calculated as:

$$var[n_i] = E[n_i][1 + \gamma E[n_i]] \qquad (7)$$

Negative binomial models based on the GLM approach were estimated in this study for comparison with the full Bayesian non-linear models that were the primary focus.

Full Bayesian Model

A full Bayesian (FB) approach was the principal focus of this study. Bayesian models integrate Bayes' theorem with classical statistical models (Gelman et al., 2014). There are two main advantages in Bayesian models that make them appealing for examining complex events such as crash occurrence. First, Bayesian models take previous knowledge into consideration by allowing the use of prior information about the parameter coefficients. Therefore, the posterior estimate of the parameter value will depend on both prior information and the current data. The inference about the posterior distribution of the model $\pi(\theta|y)$ is obtained as:

$$\pi(\theta|y) = \frac{L(y|\theta)\, \pi(\theta)}{\int L(y|\theta)\, \pi(\theta) d\theta} \qquad (8)$$

where θ is the vector of parameters, y is the vector of observed data, $L(y|\theta)$ is the likelihood function, $\pi(\theta)$ is the posterior distribution of θ, $\int L(y|\theta)\, \pi(\theta) d\theta$ is the marginal distribution of data y.

WinBUGS software was employed to implement Markov Chain Monte Carlo (MCMC) algorithms, which are suited to a wide range of target distributions for analyzing complex models. In this, each parameter is assigned a distribution with a mean and a variance. Diffuse prior distributions were used to let the model potentially include a wide range of values for the parameters (Gelman 2006). It should also be noted that several reasonable mean values for the prior distributions were used to see how much the parameter estimation depends on the chosen priors. In the end, it was concluded that the parameter estimates mostly depend on the data used and that the prior has little impact.

The second advantage of FB methods in the context of this study pertains to the flexibility of functional form that can be used in Bayesian models. This overcomes the key limitation of the conventional generalized linear modeling (GLM) noted earlier in which CMFs estimated from model coefficients are only dependent on the difference in the value of a variable of interest and not on the actual values. To estimate more complex functions, e.g., those that cannot be linearized and estimated with GLM, the full Bayesian approach takes advantage of MCMC sampling techniques (Carlin and Chib, 1995). MCMC simulations are generated in a way that the stationary distribution of the Markov chain is the posterior distribution of interest. In this, a sample from the posterior distribution of interest is provided, which allows estimates of the posterior distributions to be produced using kernel density estimates and summary statistics of interest, such as posterior means, medians, and credible intervals (Ntzoufras, 2011).

It should be noted in passing that a full Bayesian approach has been also employed in before-after studies as an alternative deriving CMFs from an EB before-after study (Persaud et al., 2010). These two applications are fundamentally different, in the main because the FB method uses a specified prior whereas the EB approach allows the prior to be estimated through the use of data (Gross et al., 2010).

RESULTS

GLM Results

Conventional GLMs were developed for comparison purposes. As noted earlier, while the estimations based on GLM method allow the CMF for the changes in median width to depend on the amount of change, the implied CMFs for increasing median width by a given amount do not depend on the original median width. The estimated models were of the form:

$$No.of\ Crashes = Length \times Year \times Exp(\beta_0) \times AADT^{\beta_1} \times Exp(\beta_2 \times Sh\_width) \times Exp(\beta_3 \times Med\_width) \quad (9)$$

where,

Length= length of the segment (km)

Year= number of years (which is equal to 3)

AADT= annual average daily traffic

Sh_width= shoulder width (m)

Med_width= median width (m)

β_0, β_1, β_2, β_3 = model parameters to be estimated

The total crash models (all severities combined) were developed for multiple vehicle, single vehicle, and median-related crashes. The parameter estimates obtained using the SAS software package (SAS Institute Inc., 2004) are shown in Table 2. Regarding the median-related crashes model, it is seen that all the parameter estimates are statistically significant at the 5% level ($P<0.05$). However, the significance of the parameter estimate for median width is not as high for the multiple vehicle and single vehicle models.

Table 2. Parameter estimates using conventional GLM approach

Coefficient	Median-related		Single vehicle		Multiple vehicle	
	Estimate (Std. Error)	Pr > ChiSq	Estimate (Std. Error)	Pr > ChiSq	Estimate (Std. Error)	Pr > ChiSq
β_0	-8.152 (0.704)	<0.001	-5.84 (0.541)	<0.001	-17.963 (0.693)	<0.001
β_1	0.881 (0.051)	<0.001	0.762 (0.039)	<0.001	1.889 (0.048)	<0.001
β_2	-0.424 (0.112)	<0.001	-0.369 (0.089)	<0.001	-0.393 (0.109)	<0.001
β_3	-0.014 (0.005)	0.004	-0.003 (0.004)	0.342	-0.008 (0.004)	0.081
Dispersion	0.339 (0.037)		0.249 (0.023)		0.378 (0.037)	

Non-linear Model (NLM) Results

As mentioned in Hauer (2004), to select the functional form "the modeller has no guidance from theory, not even from dimensional analysis. Nor is there much guidance in the data, so much so that a large number of mathematical functions could be chosen to fit the same data".

With these considerations in mind, the function was aimed to be generally of an exponential form and to potentially handle rapid and non-uniform changes, should such effects be revealed in the data. Based on these objectives, different functional forms were considered using the FB MCMC approach, and, finally, the following power function was selected as the best one for inferring about the CMFs for median width:

$$No.of\ Crashes = Length \times Year \times Exp(a) \times AADT^{b} \times Exp(c \times Sh\_width) \times Exp(Exp(d \times Med\_width)) \qquad (10)$$

where,
a, b, c, and d are the coefficients to be estimated.

The prior distributions were assumed to be uninformative (also known as "vague" or "diffuse" priors); this is common practice in the road safety domain since reliable prior information is not usually available (Lan et al., 2009, El-Basyouny and Sayed, 2010). Uninformative priors imply that the assumed variance for the parameter is large in order to reflect a balance among a wide range of possible outcomes (Gelman, 2006). That said, different diffuse prior distributions were tested, including Normal, Weibul, and log-normal distributions. In the end, the diffuse Normally distributed priors (dnorm (0, 10)) provided the most stable parameter estimates. In addition, for the over-dispersion coefficient an inverse-gamma distribution was assumed with parameters 1 and 1 (dgamma (1,1)). The MCMC sampling procedure was undertaken for 100,000 iterations to calculate the posterior estimates of the parameters. The first 10,000 iterations were not considered for the parameter estimation and were discarded as burn-in. To check for the convergence of the parameters, plots for trace and autocorrelation were examined.

Separate models for total and fatal/injury crashes were developed for the three different types of crashes mentioned earlier. Tables 3-5 present the parameter estimates for the FB-MCMC models and their associated statistics at 95% credible intervals.

Table 3. MCMC statistics for multiple vehicle crashes

Parameter	Model	Mean	St.Dev	MC error	2.50%	Median	97.50%
a	Total FI	-17.87 (-18.3)	0.588 (0.758)	0.033 (0.051)	-18.97 (-19.45)	-17.86 (-18.44)	-16.76 (-16.51)
b	Total FI	1.869 (1.721)	0.049 (0.053)	0.003 (0.003)	1.769 (1.596)	1.869 (1.729)	1.961 (1.81)
c	Total FI	-0.413 (-0.295)	0.093 (0.115)	0.005 (0.007)	-0.598 (-0.524)	-0.413 (-0.296)	-0.237 (-0.075)
d	Total FI	-0.349 (-0.383)	0.17 (0.173)	0.003 (0.004)	-0.758 (-0.788)	-0.316 (-0.356)	-0.108 (-0.115)
Dispersion Parameter	Total FI	2.557 (2.875)	0.249 (0.371)	0.003 (0.008)	2.101 (2.214)	2.547 (2.854)	3.077 (3.668)

Table 4. MCMC statistics for single vehicle crashes

Parameter	Model	Mean	St.Dev	MC error	2.50%	Median	97.50%
a	Total	-5.856	0.44	0.024	-6.643	-5.897	-4.848
	FI	(-8.606)	(0.661)	(0.037)	(-9.96)	(-8.575)	(-7.414)
b	Total	0.753	0.035	0.002	0.678	0.756	0.816
	FI	(0.85)	(0.047)	(0.003)	(0.765)	(0.847)	(0.942)
c	Total	-0.365	0.087	0.004	-0.531	-0.366	-0.194
	FI	(-0.395)	(0.117)	(0.006)	(-0.606)	(-0.401)	(-0.168)
d	Total	-0.275	0.131	0.002	-0.64	-0.237	-0.13
	FI	(-0.301)	(0.112)	(0.002)	(-0.567)	(-0.28)	(-0.136)
Dispersion Parameter	Total	3.95	0.358	0.002	3.286	3.936	4.69
	FI	(3.41)	(0.448)	(0.004)	(2.628)	(3.376)	(4.38)

Table 5. MCMC statistics for median-related crashes

Parameter	Model	Mean	St.Dev	MC error	2.50%	Median	97.50%
a	Total	-8.444	0.628	0.036	-9.693	-8.416	-7.297
	FI	(-10.21)	(0.823)	(0.046)	(-11.97)	(-10.13)	(-8.702)
b	Total	0.871	0.049	0.003	0.78	0.87	0.966
	FI	(0.932)	(0.06)	(0.003)	(0.815)	(0.931)	(1.057)
c	Total	-0.445	0.101	0.005	-0.646	-0.443	-0.247
	FI	(-0.547)	(0.149)	(0.008)	(-0.842)	(-0.551)	(-0.252)
d	Total	-0.112	0.054	0.002	-0.24	-0.105	-0.03
	FI	(-0.235)	(0.167)	(0.005)	(-0.664)	(-0.187)	(-0.029)
Dispersion Parameter	Total	2.91	0.319	0.003	2.335	2.892	3.584
	FI	(2.643)	(0.524)	(0.007)	(1.807)	(2.578)	(3.837)

The results show that all parameters used are significant at the 95% credible interval in all of the six developed models. The uncertainty of the parameter estimation was evaluated by checking the range of the credible intervals, with a wide credible interval implying that the uncertainty of the parameter estimation is large. Looking at the upper and lower bounds of credible intervals, it can be seen that the ranges of variations are reasonable.

Comparative Assessment of Model Fits

The quality of fit for the functional form proposed was investigated with the cumulative residual (CURE) method described in Hauer and Bamfo (1997). To generate a CURE plot for the feature of interest, sites are first sorted in ascending order by the feature. For each site, the residual (the difference between observed and predicted values) is calculated, and then the CUREs are obtained and plotted for each value of the feature. Since crashes are random by nature, the CURE line also represents a so-called random walk (Kononov et al., 2011). A CURE plot that oscillates around the abscissa implies that the functional form fits well in all ranges of the variable of interest. On the other hand, a steady increase (decrease) in the CURE line, within a specific range of the variable, implies that the model is inclined to predict less (or more) crashes than have been observed, for that particular range of the variable. Also, if the CURE line frequently goes beyond the two standard deviations, it either shows the presence of outliers or signifies an ill-fitting model.

Figures 2 & 3 depict the CURE plots for total median-related crashes obtained from the non-linear model (NLM) estimated with the FB MCMC approach and the GLM, respectively. Since the functional forms for both methods are exponential, the predictions, and subsequently the plotted CURE lines, follow an essentially similar pattern. As seen in Figure 2, the plotted line pertaining to the proposed NLM oscillates well around the abscissa and lies within the two standard deviation lines, so it can be concluded that, for different ranges of median width, the non-linear functional form fits the data well. However, the CURE plot obtained from the GLM method indicates that the line departs from the two standard deviation line. Looking more closely at Figure 3, it is seen that there is an increasing trend in the CURE line from, say, 25 m onwards, implying that fewer crashes have been predicted than were observed in this range. This translates to an overestimation of the safety effects of median width for wider medians. This finding confirms the intuitive expectation that the safety effects of changing median width should be estimated in such a way that the estimated CMFs not only depend on the amount of change, but also on the original width.

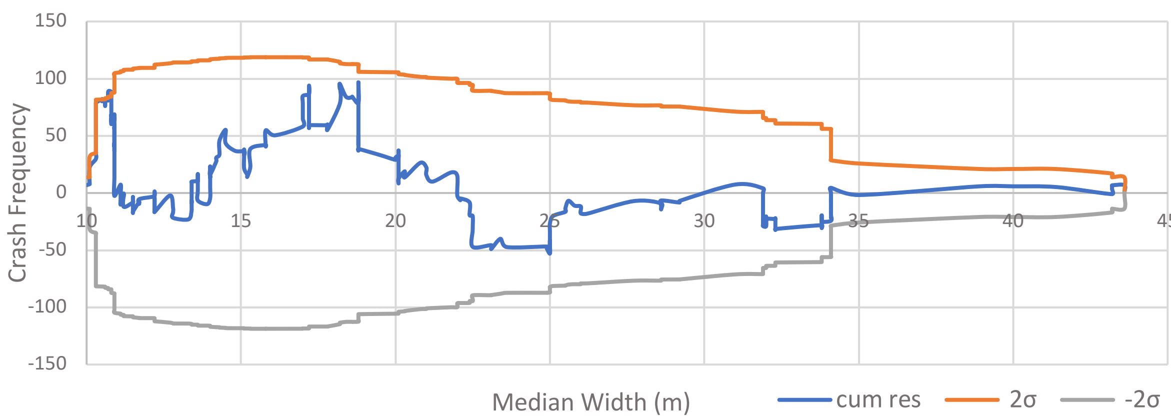


Figure 2. CURE plot for total median-related crashes using the non-linear model

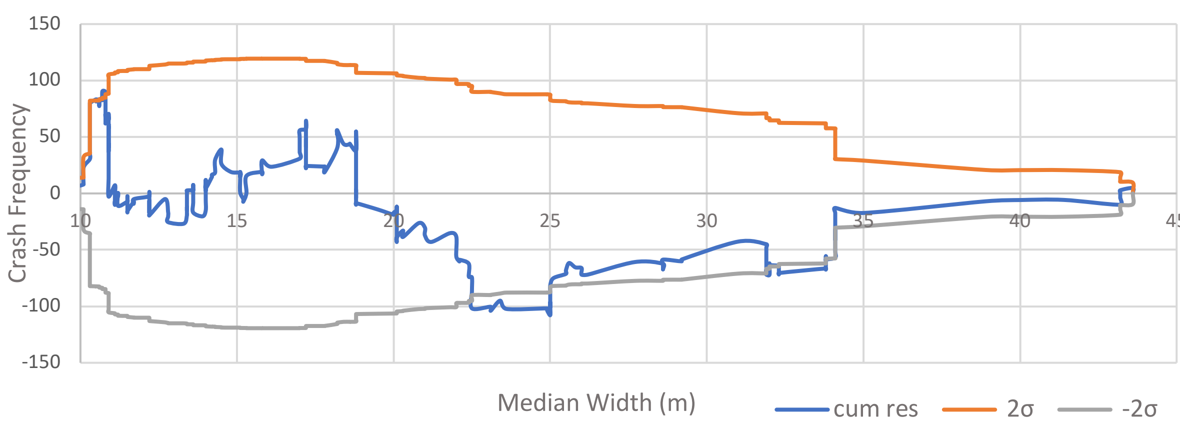


Figure 3. CURE plot for total median-related crashes using the GLM method

CMFunction Estimation and Comparison

For the GLM approach, based on Equation (9), the CMF for increasing the width of a median from Med_width_1 to Med_width_2 can be computed using the following CMFunction:

$$CMF = Exp(\beta_3(Med\_width_2 - Med\_width_1)) \quad (11)$$

For the proposed non-linear approach, based on Equation (10), the CMFunction for increasing the width of a median from Med_width_1 to Med_width_2 can be estimated as follows:

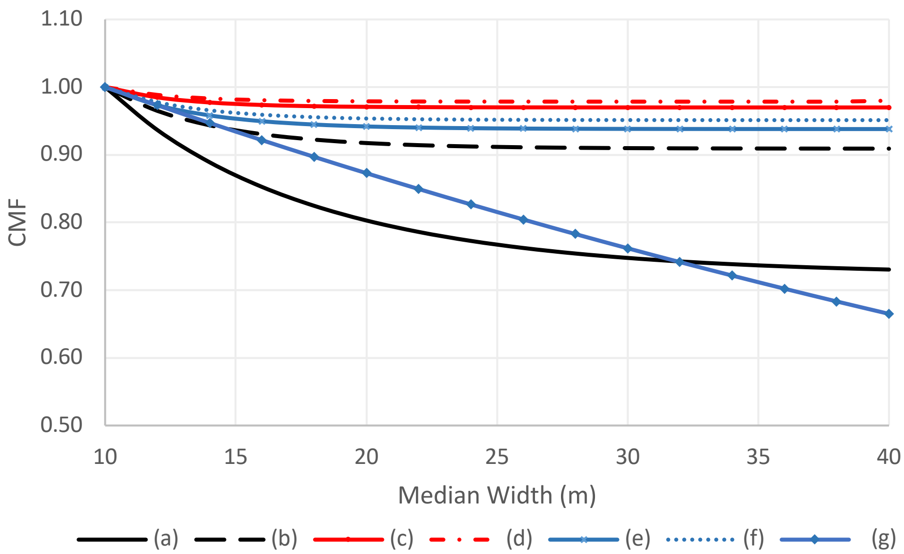
$$CMF = Exp(Exp(d \times Med\_width_2) - Exp(d \times Med\_width_1)) \quad (12)$$

To obtain the standard error of the estimates, the variance of CMF at a specific point, $var(CMF)$, can be calculated using a so-called delta method as follows (Rice, 2007, Park et al., 2016):

$$var(CMF) \approx \frac{\partial G(\hat{\beta})}{\partial \hat{\beta}} \times Cov(\beta) \times t\left(\frac{\partial G(\hat{\beta})}{\partial \hat{\beta}}\right) \quad (13)$$

where $\frac{\partial G(\hat{\beta})}{\partial \hat{\beta}}$ is a row vector of partial derivatives of $G(\beta)$ at point $\hat{\beta}$, $Cov(\beta)$ is a variance-covariance matrix of β, and $t\left(\frac{\partial G(\hat{\beta})}{\partial \hat{\beta}}\right)$ is the transpose of the function in the parenthesis.

CMFs for several changes in median width from a base condition of 10 m median width are shown for illustration in Figure 4 for the six models developed. The CMFunctions based on the non-linear model show that while the effect of changing median width is quite pronounced for median-related crashes, in particular for total median-related crashes, its impact on all multiple and single vehicle crashes is marginal, similar to what was found in the NCHRP study by Bonneson et al. (2012). The CMFunction for total median-related crashes using GLM is shown for comparison purposes. As seen, the GLM CMFunction implies that the same amount of a change in median width will always bring a fixed change in safety, regardless of the original width. On the other hand, the concave shape of the curve in the non-linear model suggests the safety effect of a change in width is more pronounced for the narrower medians compared to the same change in wide medians.



(a): total median-related, (b): FI median-related, (c): total multiple vehicle, (d): FI multiple vehicle, (e): total single vehicle, (f): FI single vehicle, (g): total median-related by GLM.

Figure 4. Plot of variation of CMF based on median width

DISCUSSION

The results of this study highlight the importance of considering median-related crashes when estimating the safety effects of increasing freeway median width where there is no barrier. These results may be best compared with those in the recent NCHRP report (Bonneson et al., 2012) in which the safety effect of increasing median width has been investigated for single vehicle and multiple vehicle crashes. The CMF for all crashes in that study comes from the summation of single vehicle and multiple vehicle crashes CMFs, weighted by the number of these crashes. The current study corroborates those findings in terms of the direction and magnitude of the safety effects of median width. More specifically, the NCHRP report suggests that although median widening on freeway segments with no barrier is negatively associated with the risk of crashes, the benefits are marginal. For instance, the effect of changing median width from 40 ft (12.2 m) to 80 ft (24.4 m) on all FI crashes is about 5%, which is comparable to the average of 3% reduction in multiple vehicle and single vehicle crashes in the present study. The main point of departure in the present study has been the focus on the median-related crashes. Notably, the results reveal that increasing the width of medians can be highly safety-effective for the median-related crashes targeted by this strategy. Importantly, it was also found while the safety benefits of small increases in width are quite pronounced for narrower medians, widening medians larger than, say, 25 m. by these same amounts may not provide any substantial reductions in crashes. To illustrate, Table 6 provides illustrative CMF estimates from Equation 2 for increasing median width by 5 m. and 10 m. steps for total median-related crashes. It is seen that the CMFs based on the proposed non-linear model depend on, and are quite sensitive to the original median width, logically decreasing with increasing original width. For instance, increasing the width by 5 m. from 10 m. to 15 m., yields a 13.1% reduction in crashes, while the same increase from 25 m. to 30 m. will decrease crashes by only 2.6%. These reductions with the same original width but with a 10 m. increase in width increase to 19.7% and 4%, respectively. It can also be seen in Table 6 that the CMFs based on the GLMs are independent of the original width, which, as noted earlier, seems illogical.

Table 6: CMFs (and standard errors) for median-related crashes for increasing median width by 5 m. and 10 m.

Original median width (m)	5-meter increase		10-meter increase	
	CMF from MCMC	CMF from GLM	CMF from MCMC	CMF from GLM
10	0.869(0.021)		0.803(0.049)	
15	0.923(0.033)		0.882(0.061)	
20	0.955(0.031)		0.931(0.054)	
25	0.974(0.025)	0.934(0.020)	0.960(0.042)	0.872(0.041)
30	0.985(0.018)		0.977(0.031)	
35	0.992(0.013)		0.987(0.021)	
40	0.995(0.008)		0.992(0.014)	

The results of the proposed non-linear model are in line with our recent research study (Persaud et al., 2020) in which increasing the length of passing relief lanes was found to have a non-linear effect on crash frequency. This suggests that the proposed approach could be potentially applied for different roadway attributes with different degrees of non-linearity.

SUMMARY AND CONCLUSIONS

CMFs for changing the width of a freeway median in a design process are based on cross-sectional regression models, in the main because of the lack of data for the preferred observational before-after study. Using a FB-MCMC approach with freely available software, to estimate non-linear models, crash modification functions (CMFunctions) were developed for estimating the effects of changing median width on all single and multiple vehicle crashes, and on median-related (i.e., left side run-off road) target crashes. It was found that the suggested power function can potentially capture the variation of the safety effects of changing median width on crashes. The estimated CMFunction for median-related crashes indicates that the CMF for a given change in width depends on, and is quite sensitive to the original median width, logically decreasing with increasing original width. These CMFunctions, in particular that for median-related crashes, complement those developed for the Highway Safety Manual (HSM) but only for single vehicle and multiple vehicle crashes The HSM functions also indicate that the effect of a change in median width depends on the original width, albeit with less sensitivity.

Further research with detailed data is warranted to confirm and extend the paper's findings. Importantly, given that the study is of necessity based on cross-sectional analysis, rather the preferred before-after approach, confidence in applying the results could be increased by performing similar studies on other datasets with a view to corroborating the results. The methodology is recommended for developing CMFunctions for other roadway features, especially those for which the preferred observational before-after study is not possible.

ACKNOWLEDGMENTS

The authors thank the Ministry of Transportation, Ontario for providing the data. The research was funded by a Discovery grant from the Natural Sciences and Engineering Council of Canada (ApplID RGPIN-2017-04457) and a grant from the Ministry of Transportation, Ontario's Highway Infrastructure Innovation Funding Program. Special thanks to Dr. Dominique Lord for referring the authors to his paper on estimating the standard deviation of CMFs/CMFunctions.

REFERENCES

AASHTO 2010. Highway safety manual, 1st edition, AASHTO Washington, DC.
ABDEL-ATY, M. A. & RADWAN, A. E. 2000. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention, 32**,** 633-642.
ALLURI, P., HALEEM, K. & GAN, A. 2012. In-service performance evaluation (ISPE) for G4 (1S) type of strong-post W-Beam guardrail system and cable median Barrier: volume i. Final Report FDOT BDK80 977-19, State of Florida Department of Transportation, Tallahassee, FL.
BONNESON, J. A., GEEDIPALLY, S. & PRATT, M. P. 2012. Safety prediction methodology and analysis tool for freeways and interchanges,. Received Online, October 2019, from: www.onlinepubs.trb.org/onlinepubs/nchrp/docs/NCHRP17-45_FR.pdf.
CALIENDO, C., GUIDA, M. & PARISI, A. 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention, 39**,** 657-670.
CARLIN, B. P. & CHIB, S. 1995. Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Methodological), 57**,** 473-484.
CHIMBA, D., RUHAZWE, E., ALLEN, S. & WATERS, J. 2017. Digesting the safety effectiveness of cable barrier systems by numbers. Transportation Research Part A: Policy and Practice, 95**,** 227-237.
DONNELL, E. T. & MASON JR, J. M. 2006. Predicting the frequency of median barrier crashes on Pennsylvania interstate highways. Accident Analysis & Prevention, 38**,** 590-599.
EL-BASYOUNY, K. & SAYED, T., 2010. Full Bayes approach to before-and-after safety evaluation with matched comparisons: case study of stop-sign in-fill program. Transportation Research Record, 2148(1), pp.1-8.
GELMAN, A., HWANG, J. & VEHTARI, A. 2014. Understanding predictive information criteria for Bayesian models. Statistics and computing, 24**,** 997-1016.
GROSS, F., PERSAUD, B. N. & LYON, C. 2010. A guide to developing quality crash modification factors. United States. Federal Highway Administration. Office of Safety.
HARKEY, D.L., SRINIVASAN, R. BAEK, J. COUNCIL, F. ECCLES, K. LEFLER, N. GROSS, F. PERSAUD, B. LYON, C. HAUER, E. & BONNESON, J., 2008. National Cooperative Highway Research Report 617: Accident Modification Factors for Traffic Engineering and ITS Improvements, NCHRP, Transportation Research Board, Washington, DC.
HAUER, E. 1997. Observational Before-After Studies in Road Safety. Pergamon Press, Elsevier Science Ltd., Oxford, England.
HAUER, E. 2000. The Median and Safety. Transportation Research Board, NCHRP 17-19(4).

HAUER, E. 2004. Statistical road safety modeling. Transportation Research Record, 1897**,** 81-87.
HAUER, E., COUNCIL, F.M. AND MOHAMMEDSHAH, Y., 2004. Safety models for urban four-lane undivided road segments. Transportation Research Record, 1897(1), pp.96-105.
HAUER, E., 2013. Even perfect regressions may not tell the effect of interventions. Transportation Research Board 92nd Annual Meeting, Washington D.C. (No. 13-4477).
HAUER, E. & BAMFO, J. Two tools for finding what function links the dependent variable to the explanatory variables. Proceedings of the ICTCT 1997 Conference, Lund, Sweden, 1997.
KONONOV, J., LYON, C. & ALLERY, B. K. 2011. Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. Transportation Research Record, 2236, 11-19.
LAN, B., PERSAUD, B., LYON, C. & BHIM, R., 2009. Validation of a full Bayes methodology for observational before–after road safety studies and application to evaluation of rural signal conversions. Accident Analysis & Prevention, 41(3), pp.574-580.
LAO, Y., ZHANG, G., WANG, Y. & MILTON, J. 2014. Generalized nonlinear models for rear-end crash risk analysis. Accident Analysis & Prevention, 62, 9-16.
LORD, D. & MANNERING, F. 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: policy and practice, 44, 291-305.
LORD, D. & WASHINGTON, S. 2018. Safe mobility: Challenges, methodology and solutions, Emerald Publishing.
LORD, D., WASHINGTON, S. P. & IVAN, J. N. 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention, 37**,** 35-46.
LUNN, D. J., THOMAS, A., BEST, N. & SPIEGELHALTER, D. 2000. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and computing, 10**,** 325-337.
MCCULLAGH, P. 2018. Generalized linear models, Routledge.
NTZOUFRAS, I. 2011. Bayesian modeling using WinBUGS, John Wiley & Sons.
PARK, J. & ABDEL-ATY, M. 2015. Assessing the safety effects of multiple roadside treatments using parametric and nonparametric approaches. Accident Analysis & Prevention, 83**,** 203-213.
PARK, B.J., LORD, D. & WU, L., 2016. Finite mixture modeling approach for developing crash modification factors in highway safety analysis. Accident Analysis & Prevention, 97, pp.274-287.
PERSAUD, B. & LYON, C., 2007. Empirical Bayes before–after safety studies: Lessons learned from two decades of experience and future directions. Accident Analysis & Prevention, 39(3), pp.546-555.
PERSAUD, B., JAFARI ANARKOOLI, A., ALMASI, S. & LYON, C., 2020. Crash Modification Functions for Passing Relief Lanes on Two-Lane Rural Roads. Transportation Research Record, Accepted for Publication.
PERSAUD, B., LAN, B., LYON, C. & BHIM, R., 2010. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. Accident Analysis & Prevention, 42(1), pp.38-43.

RICE, J., 2007. Mathematical Statistics and Data Analysis, 3rd edition. Duxbury advanced series, USA.

RUSSO, B. J. & SAVOLAINEN, P. T. 2018. A comparison of freeway median crash frequency, severity, and barrier strike outcomes by median barrier type. Accident Analysis & Prevention, 117**,** 216-224.

SAS INSTITUTE INC. 2004. SAS® Qualification Tools User's Guide. Version 9.1.2. SAS Institute Inc., Cary, NC, USA.

SPIEGELHALTER, D., THOMAS, A., BEST, N. & LUNN, D. 2003. WinBUGS user manual. Citeseer.

STAMATIADIS, N. & COUNCIL, N. R. 2009. Impact of shoulder width and median width on safety, National Academies Press.

WOLFINGER, R.D., 1999, April. Fitting nonlinear mixed models with the new NLMIXED procedure. In Proceedings of the 24th Annual SAS Users Group International Conference (SUGI 24) (pp. 278-284).

WOOD, J.S., DONNELL, E.T. & PORTER, R.J., 2015. Comparison of safety effect estimates obtained from empirical Bayes before–after study, propensity scores-potential outcomes framework, and regression model with cross-sectional data. Accident Analysis & Prevention, 75, pp.144-154.

WU, L., LORD, D. & ZOU, Y., 2015. Validation of crash modification factors derived from cross-sectional studies with regression models. Transportation Research Record, 2514(1), pp.88-96.

XIE, Y., LORD, D. & ZHANG, Y. 2007. Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. Accident Analysis & Prevention, 39**,** 922-933.

YANNIS, G., DRAGOMANOVITS, A., LAIOU, A., LA TORRE, F., DOMENICHINI, L., RICHTER, T., RUHL, S., GRAHAM, D. & KARATHODOROU, N. Road traffic accident prediction modelling: a literature review. Proceedings of the Institution of Civil Engineers-Transport, 2017. Thomas Telford Ltd, 245-254.