Development of Accident Modification Factors for Rural Frontage Road Segments in Texas Using Generalized Additive Models

Xiugang Li[1]; Dominique Lord, M.ASCE[2]; and Yunlong Zhang, M.ASCE[3]

Abstract: The objective of this study consists of assessing the application of generalized additive models (GAMs) for estimating accident modification factors (AMFs). GAMs are a new type of models that have been recently introduced by the statistical community for modeling observed data. These models offer more flexible functional forms than traditional generalized linear models and allow for more adaptable variable interactions. As recently documented in the literature, variable interactions should be included in the development of AMFs. To accomplish the study objective, AMFs were derived from GAMs using data collected on rural frontage roads in Texas. The AMFs were then compared to the AMFs produced from a previous study using the same data set. The results of the study show that AMFs produced from GAMs are more flexible to characterize the safety effect of simultaneous changes in geometric and operational features (or variable interactions) than when independent AMFs are applied together. The results also show that GAMs indicated a nonlinear relationship between crash risk and changes in lane and shoulder widths for frontage roads in Texas.

DOI: 10.1061/(ASCE)TE.1943-5436.0000202

CE Database subject headings: Regression models; Traffic accidents; Rural areas; Texas.

Author keywords: Regression models; Accident modification factors; Generalized additive models; Frontage roads; Texas.

Introduction

The development and use of accident modification factors (AMFs) in highway safety has gained a lot of popularity over the last few years (see Shen and Gan 2003; Hughes et al. 2005; Lord and Bonneson 2006; Bahar et al. 2007; Elvik 2009). AMFs are multiplicative factors used for adjusting accident frequency estimated from baseline models to quantify changes in geometric design and traffic operational features (Hughes et al. 2005). An AMF greater than 1.0 represents the situation where the change is associated with more crashes while an AMF less than 1.0 indicates a change with fewer crashes. The AMF can be represented by a single value that describes average conditions (Hughes et al. 2005) or as a function linking crash risk and variables, such as traffic flow, lane, or shoulder widths (Elvik 2009). Eq. (1) illustrates how AMFs are applied

$$\mu_{\text{final}} = \mu_{\text{baseline}} \times \text{AMF}_1 \times \cdots \times \text{AMF}_k \quad (1)$$

where μ_{final}=final predicted number of crashes per unit of time; μ_{baseline}=baseline predicted number of crashes per unit of time (usually via a regression model); and $\text{AMF}_1 \times \cdots \times \text{AMF}_k$=accident modification factors assumed to be independent.

In Eq. (1), baseline models represent estimated regression models using data that meet specific nominal conditions, such as 12-ft lane and 8-ft shoulder widths for two-lane rural highway segments or no turning lanes at intersections. These conditions usually reflect design or traffic operational variables most commonly used by state transportation agencies (defined as state DOTs). Consequently, baseline models typically only include traffic flow as covariates (e.g., $\mu=\beta_0 F_{\text{major}}^{\beta_1} F_{\text{minor}}^{\beta_2}$ for intersection or $\mu=\beta_0 L F_L^{\beta_1}$ for highway segments, where $F_{\text{major}}, F_{\text{minor}}, F_L$=entering flows for the major and minor approaches at intersections and on segments and L=length of the segment).

Various methods have been proposed to estimate AMFs. The most popular ones include the before-after study, regression based models and the cross-sectional study (see E. Hauer and B. N. Persaud, unpublished report, 1996; Harwood et al. 2000; Washington et al. 2005; Gross and Jovanis 2007a; Fitzpatrick et al. 2008; Bonneson and Pratt 2008; Gross et al. 2009). Over the last few years, a selected number of researchers have found several issues with these methods (Shen and Gan 2003; Bonneson and Lord 2005; Gross and Jovanis 2007a). From those noted in the literature, two important limitations have been identified. First, notwithstanding the method used in their development, each AMF is usually assumed to be independent, which means that each design and operational element is analyzed by itself without considering the influence of other design or operational features. In practice, AMFs may not be completely independent, since changes in geometric design or operational characteristics on highways are not done independently (e.g., lane and shoulder width may be changed simultaneously) and the combination of these changes can influence crash risk differently than if they are estimated separately. As a matter of fact, Bonneson et al. (2007) and Gross et al. (2009) have both argued that the interaction be-

[1]Transportation Analyst, Oregon Dept. of Transportation, 555 13th St. NE Ste 2, Salem, OR 97301. E-mail: xiugang.li@odot.state.or.us

[2]Associate Professor, Texas A&M Univ., CE/TTI 301-A, 3136 TAMU, College Station, TX 77843 (corresponding author). E-mail: d-lord@ttimail.tamu.edu

[3]Associate Professor, Texas A&M Univ., CE/TTI 301-G, 3136 TAMU, College Station, TX 77843. E-mail: yzhang@civil.tamu.edu

Note. This manuscript was submitted on February 25, 2009; approved on July 6, 2010; published online on December 15, 2010. Discussion period open until June 1, 2011; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering***, Vol. 137, No. 1, January 1, 2011. ©ASCE, ISSN 0733-947X/2011/1-74–83/$25.00.

tween design features should be included in the development of AMFs.

Second, most of the methods used for estimating AMFs assume a linear (or exponential) relationship between safety and changes in design or operational features (this only applies to accident modification functions). Recently, a few researchers have noted that design elements, such as shoulder or lane width could follow a U-shaped relationship with safety, where the crash risk could be higher both for narrow and wide widths [E. Hauer, "Shoulder width, shoulder paving and safety," unpublished manuscript prepared for the Federal Highway Administration, Toronto, 2000 (http://ca.geocities.com/hauer@rogers.com/Pubs/Shoulderwidth.pdf); Xie et al. 2007; Li et al. 2008]. On the other hand, others have found sinusoidal relationships between crash risk and lane and shoulder widths (Hauer 2004; Gross and Jovanis 2007a,b). Given the limitations described above, as well as the contradictory results with regards to the function linking crashes to design features, there is a need to determine whether a new method could be used for estimating AMFs that specifically include variable interactions in the development of AMFs, and examine the characteristic of the nonlinear relationship between safety and changes in these features.

The objective of this study consists of assessing the application of generalized additive models (GAMs) for estimating AMFs. This work expands on the work done by Lord and Bonneson (2007) on AMF development for rural frontage roads in Texas. GAMs are a new type of models that have been recently introduced in the statistical community to model observed data (Hastie and Tibshirani 1990; Wood 2006, 2003). These models offer more flexible functional forms than traditional generalized linear models (GLMs) and allow for more adaptable variable interactions. [Note: variable interactions in GLMs are usually used to determine whether or not the variables are different from each other (e.g., different slopes or intercept). However, even when they are used as such, all the variables are assumed to be independent of each other]. Xie and Zhang (2008), who first introduced GAMs for predicting highway crash frequency, showed that GAMs provided better nonlinear approximation abilities than GLMs while retaining the basic framework of GLMs. Furthermore, GAMs can still generate statistically interpretable results, similar to GLMs. To accomplish the study objective, AMFs produced from GAMs were estimated using data collected on rural frontage roads in Texas. The AMFs were then compared to the AMFs documented in Lord and Bonneson (2007).

This paper is divided into eight sections. The second section presents a literature review on existing methods used for estimating AMFs. The third section describes the summary statistics of the data. The fourth section provides details about the characteristics of GLMs and GAMs. The fifth section explains the statistical analysis procedure to develop the GAMs. The sixth section describes the frontage road AMFs derived from the data and GAMs. The seventh section documents the comparison between the safety performance function (SPF) developed in this work, that is the relationship between the number of crashes and traffic flow, and the one documented in Lord and Bonneson (2007). The last section summarizes the work carried out in this research and provides ideas for further work.

Background

AMFs can be estimated using various statistical methods. The four most common methods that have been documented in the literature are briefly described below.

The first method is based on the before-after study framework. This method consists of estimating the safety effects of changes in geometric design features, traffic operations, or other characteristics by examining the increase or reduction in crash counts between the before and after periods. Three techniques have been proposed for this kind of study: (1) the simple or naïve before-after study; (2) the before-after study with a control group; and (3) the before-after study using the empirical Bayes (EB) method. These techniques, including their limitations, have been well documented by others and are not described here (Hauer 1997; Persaud et al. 2001; Ye and Lord 2009). With the before-after study, the AMF can only take a single value rather than a function.

The second method consists of estimating AMFs using the coefficients of regression models. This method has been used by Lord and Bonneson (2007) and Washington et al. (2005) for estimating AMFs for rural frontage roads in Texas and rural intersections in various states, respectively. The AMFs are estimated the following way:

$$\mathrm{AMF}_k = e^{(\beta_k \times [x_k - \bar{x}_k])} \tag{2}$$

where x_k=range of values or a specific value investigated (e.g., lane width, shoulder width, etc.) for AMF_k; $\bar{x}_k$=baseline conditions or average conditions for the variable k (when needed or available); and β_k=regression coefficient associated with the variable k (estimated from data).

This method provides a simple way to estimate the effects of changes in geometric design features [note: each regression coefficient is associated with one geometric design feature in the model and Eq. (2) is used to estimate one AMF per design feature]. However, although the variables of the original regression model are assumed to be independent, they may still be correlated (or not truly independent), which could affect the model's coefficients. If these coefficients are biased because of correlation, there is no point in using Eq. (2) to develop AMFs. The variance inflation factor can be used for detecting correlated variables, but this procedure only flags extreme cases of correlation (Myers 2000). With this method, the AMF can only follow an exponential relationship or function.

For the third method, AMFs are estimated using baseline models and applying them to data that do not meet the nominal conditions. This method has been proposed by Washington et al. (2005), who have recalibrated models for estimating the safety performance of rural signalized and unsignalized intersections. For this method, the baseline model is first applied to sites not meeting all of the baseline conditions; then, the predicted and observed values per year are compared with each other, and a simple linear relationship between these two values is estimated via a regression model to determine whether or not AMFs could be produced from its coefficients. The linear equation is given by the following:

$$Y_i - \mu_i = \gamma_1 x_1 + \cdots + \gamma_k x_k \tag{3}$$

where μ_i=mean number of crashes for site i per year estimated by the baseline model; Y_i=observed number of crashes for site i per year; x_k=series of baseline variables (each site not meeting one or more of these variables); and γ_k=regression coefficients estimated from the data (note: original data set less the data used for estimating the baseline model). The AMFs are estimated using the following relationship when the coefficients are found to be statistically significant (e.g., 5% or 10% level):

J. Transp. Eng., 2011, 137(1): 74-83

$$\mathrm{AMF}_k = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} Y_i - \gamma_k} \quad (4)$$

where AMF_k=AMF for the variable k; N=number of observations in the sample in which the model in Eq. (1) was applied; and γ_k=coefficients estimated from data. This method only provides a single value for the AMF, similar to the first method.

The fourth method consists of estimating the AMF using a cross-sectional study. In this method, sites with different characteristics are directly compared with each other. In the literature, two approaches have been proposed under this method. For the first approach, Gross and Jovanis (2007a,b) proposed the use of a case-control design (often referred to as a cohort study in epidemiology). The objective of this approach is to estimate crash risk using odds ratio for different geometric design characteristics. They applied their approach to estimate the safety effects of lane and shoulder widths located on rural two-lane highways in Pennsylvania. They found a sinusoidal relationship between lane and shoulder widths, which may be counterintuitive since very narrow widths were found to be almost as safe as widths meeting the nominal conditions (Hauer 2000, unpublished; Li et al. 2008). Nonetheless, one advantage of the case-control design is that it does not have an inherent assumption about the functional form between the geometric feature under investigation and changes in safety.

The second approach was proposed by Bonneson and Pratt (2008) who estimated AMFs using a match-paired study design. In their approach, sites with similar traits, but having a different characteristic for the AMF under investigation, are compared directly with each other (i.e., same segment length, flow, lane width, etc.). Regression models are estimated using each paired comparison and AMFs are produced from these models to capture this difference. In their example application, they developed AMFs for highway curvature located on rural two-lane highways in Texas. Thus, the paired sites were located adjacent to each other, where the study site was a horizontal curve while the matched (or controlled) site was a straight tangent located within 0.1 mi from the curved section. This approach has the advantages of controlling for confounding factors, since the sites are adjacent to each other. On the other hand, this approach can be difficult to implement, since finding matched pairs with the exact same characteristics (other than the one used for developing the AMF) may not always be feasible (Fitzpatrick et al. 2009). Furthermore, the relationship between safety and changes in design features can only follow an exponential function.

Although some are still at an experimental stage, a few researchers have proposed other methods for estimating AMFs. For instance, Xie et al. (2007) applied Bayesian neural networks for estimating predictive models and AMFs also for rural frontage roads in Texas. They used a subset of the data used in this research. The results showed that AMFs for lane and shoulder widths were also nonlinear and followed a U-shaped relationship, similar to the relationship described below (narrow and very wide widths experienced more crashes). Expanding on their work, Li et al. (2008) used support vector machine models to the same data set and found similar results. The next section describes the characteristics of the data.

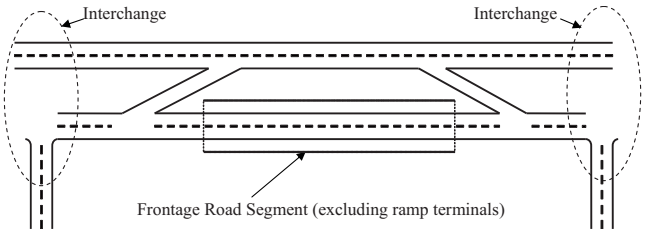


Fig. 1. Frontage road segment for analysis (Lord and Bonneson 2007, with permission from Transportation Research Board)

Data

The data used for developing the GAMs were collected at 123 segments on rural frontage roads in Texas. This is the same data set used by Lord and Bonneson (2007). This data set included both one-way and two-way frontage road segments, as shown in Fig. 1. Lord and Bonneson (2007) reported that the ramp terminal area was not included in the selected segments (defined as segment proper), since they have distinct operational characteristics. The segments were located along four Texas highway corridors: (1) I-35 between Georgetown, Texas and the splitting point between I-35E and I-35W north of Waco, Texas; (2) I-10 between Glidden and Brookshire, Texas; (3) I-45 between Willis and Centerville, Texas; and (4) SR-6 to SR-190 near Bryan and College Station, Texas.

Crash data for each frontage road segment were extracted from the Texas Department of Public Safety electronic databases for the years 1997–2001 (Lord and Bonneson 2007). At the time Lord and Bonneson (2007) performed their study (in 2005), the most recent data available were in 2001. Only "segment-related" crashes were used in this study in order to eliminate the influence of intersections and ramp terminals. Tables 1 and 2 summarize the characteristics related to the geometric design and traffic operational features of frontage-road segments and the crash data, respectively. During the 5-year period, 186 and 124 injury crashes occurred on the frontage road segments. The levels of severity included fatal (K), incapacitating injury (A), nonincapacitating injury (B), and possible injury (C) and property damage only (O). Other data that were collected in this study include segment length, ADT, lane width, and right-shoulder width. More detailed descriptions about the data collection process can be found in Lord and Bonneson (2007). In this research, the data were grouped for KABCO (or defined as total) and KABC crashes (injury).

Characteristics of Models

This section briefly describes the fundamental characteristics of GLMs and GAMs. The GLMs are described here to better explain the differences with GAMs.

Generalized Linear Models

The number of crashes at the ith rural frontage road segment Y_i can be assumed to follow a negative binomial (NB) distribution. A typical NB regression model is usually characterized the following way (Miaou 1994):

$$\Pr(Y_i = y_i) = \frac{\Gamma(y_i + \rho)}{\Gamma(y_i + 1)\Gamma(\rho)} \left(\frac{\mu_i}{\mu_i + \rho}\right)^{y_i} \left(\frac{\rho}{\mu_i + \rho}\right)^{\rho} \quad (5)$$

J. Transp. Eng., 2011, 137(1): 74-83

Table 1. Frontage-Road Segment Physical and Crash Characteristics

Highway	Operation	Number of segments	ADT, vpd			Segments with edge delineation (%)		Segment length (mi)			Total (KABCO)[a] crash frequency in 5 years	Injury (KABC) crash frequency in 5 years
			Avg.	Min.	Max.	Left	Right	Avg.	Min.	Max.		
SH-6/	Two-way	11	2,360	110	6,168	73	73	2.00	1.06	2.66	19	11
SH-190	One-way	20	2,550	140	5,270	100	10	1.12	0.78	1.89	33	23
I-10	Two-way	16	675	168	1,585	25	25	2.74	1.36	5.34	15	8
I-35	Two-way	57	575	125	2,199	33	33	1.97	0.69	3.76	72	44
	One-way	6	790	361	1,046	100	100	1.93	1.13	2.64	9	8
I-45	Two-way	10	1,990	218	1,988	50	40	2.40	0.79	4.21	21	17
	One-way	3	4,470	3,093	5,766	100	100	1.26	1.00	1.77	17	13
Summary	Two-way	94	2,385	110	6,188	38	37	2.15	0.69	5.34	127	80
	One-way	29	940	140	5,766	100	38	1.30	0.78	2.64	59	44
	Overall	123	1,230	110	6,168	53	37	1.92	0.69	4.21	186	124

[a]K=fatal; A=injury Type A (incapacitating injury); B=injury Type B (nonincapacitating injury); C=injury Type C (possible injury); and O=PDO.

$$\text{Expectation of } Y_i \text{ is } \mu_i = g(\mathbf{x}_i) \tag{6}$$

$$\text{Variance of } Y_i \text{ is } \text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\rho} \tag{7}$$

where Y_i=dependent random variable following a NB distribution with the inverse dispersion parameter ρ; y_i=number of crash collected on segment i; $\mathbf{x}_i$=vector representing the crash related variables for segment i; and $g(\mathbf{x}_i)$=functional form of the regression model.

It can be shown that the NB distributed Y_i can be derived as a gamma mixture of Poisson distribution (Cameron and Trivedi 1998). If $\rho \to \infty$, the distribution reverts back to a Poisson distribution. In this case, a Poisson regression model would be more suitable than a NB regression model. That is, $Y_i \sim \text{Poisson}(\mu_i)$ and $\mu_i = g(\mathbf{x}_i)$.

An important aspect related to the development of predictive models is the selection of the functional form, $g(\mathbf{x}_i)$, linking the dependent variable to the covariates of the model. As discussed by Xie et al. (2007), the functional form is usually determined empirically and is often influenced by transportation safety analyst's experience. For both NB and Poisson GLMs, Lord and Bonneson (2007) selected the following functional form:

$$\mu_i = g(\mathbf{x}_i) = \beta_0 L_i F_i^{\beta_1} \exp\left(\sum_{k=2}^{n} \beta_k x_{ik}\right) \tag{8}$$

where μ_i=estimated number of crashes per year for segment i; F_i=average daily traffic (ADT) for segment i, vehicles per day (veh/day); L_i=length of segment i, mile; x_{ik}=kth explanatory variable for segment i; and $\beta_0, \beta_1, \ldots, \beta_k$=coefficients to be estimated. Eq. (9) below is obtained by taking logarithm on both side of Eq. (8)

$$\ln(\mu_i) = \ln(\beta_0) + \ln(L_i) + \beta_1 \ln(F_i) + \sum_{k=2}^{n} \beta_k x_{ik} \tag{9}$$

The terms $\ln(\beta_0)$ and $\ln(L_i)$=intercept and the offset with respect to segment length, respectively. In Eqs. (8) and (9), μ_i is reasonably assumed to increase in direct proportion to the increase of L_i (Lord and Bonneson 2007). However, with the limitation of GLM, $\ln(\mu_i)$ is assumed to have linear relationship with $\ln(F_i)$ and x_{ik}, which might be nonlinear (other than logarithmic when it is transformed back). As discussed below, the GAM could offer another choice with more flexible options for modeling.

Generalized Additive Model

Compared to the GLM shown in Eq. (9), GAMs provide a more flexible functional form, $g(\mathbf{x}_i)$, which involves smooth functions for the covariates of the model. A potential functional form is illustrated in Eq. (10)

$$\ln(\mu_i) = \ln(\beta_0) + \ln(L_i) + s_1(F_i) + \sum_{k=2}^{n} s_k(x_{ik}) \tag{10}$$

where $s_1, s_2, \ldots, s_k$=univariate smooth functions. The smooth function can be used for different combinations of covariates. For instance, $s(F_i, x_{i,2}, x_{i,3}, \ldots, x_{i,k})$ is a smooth function that includes all explanatory variables. The smooth function represents a more flexible relationship between $\ln(\mu_i)$ and the covariates, and is not limited to the linear or logarithm relationship, as it is defined for GLMs.

The theoretical development of GAMs has been documented in Hastie and Tibshirani (1990) and Wood (2006). Xie and Zhang (2008) introduced GAMs with smooth function bases of cubic regression splines to predict the crash frequency at signalized

Table 2. Lane Width and Pavement Shoulder Width of Frontage-Road Segment Data

Operation	Lane width (ft)		Paved right-shoulder width (ft)		Paved left-shoulder width (ft)	
	Two-way	One-way	Two-way	One-way	Two-way	One-way
Avg.	10.5	11.7	1.3	1.3	1.1	2.4
Min	9	10	0	0	0	0
Max.	13	13	9	8	9	7

J. Transp. Eng., 2011, 137(1): 74-83

intersections. The bases for using cubic regression splines are only useful for representing smooth functions for one variable. In this study, the bases known as thin plate regression splines were adopted. With these bases, the smooth function can be used for grouping (or smoothing) multiple variables together. The following paragraphs in this section briefly describe the characteristics of thin plate regression splines (Wood 2003, 2006). Suppose we have N observations $(y_i, \mathbf{x}_i)$ for $i=1,2,\ldots,N$, and need to estimate the smooth function as

$$y_i = s(\mathbf{x}_i) + \varepsilon_i \tag{11}$$

Wood (2003, 2006) showed that thin plate spline smoothing estimates $s(.)$ can be calculated using the following:

$$\min_{\boldsymbol{\delta},\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\delta}^{\mathbf{T}}\mathbf{E}\boldsymbol{\delta} \text{ subject to } \mathbf{T}^{\mathbf{T}}\boldsymbol{\delta} = \mathbf{0} \tag{12}$$

where $\boldsymbol{\delta}, \boldsymbol{\alpha}$=vectors of coefficients to be estimated; $\mathbf{y}$=vector of y_i data; λ=smoothing parameter; $\mathbf{T}$=matrix of $T_{ik}=\phi_k(\mathbf{x}_i)$, functions $\phi_k(.)$ are linearly independent polynomials spanning the space of polynomials in d dimension R^d, and $m=\Sigma_{l=1}^{d} v_l$, where v_l is the lth element of the base; $\mathbf{E}$=matrix of $E_{ik} \equiv \eta_{md}(\|\mathbf{x}_i - \mathbf{x}_k\|)$; and

$$\eta_{md}(r) = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d}\log(r), & d \text{ even} \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d}, & d \text{ odd} \end{cases}$$

By letting $\mathbf{E}=\mathbf{U}\mathbf{D}\mathbf{U}^{\mathbf{T}}$ be the eigendecomposition of $\mathbf{E}$ and $\boldsymbol{\delta}=\mathbf{U}_k\boldsymbol{\delta}_k$, Eq. (12) becomes (see Wood 2003, 2006)

$$\min_{\boldsymbol{\delta}_k,\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\boldsymbol{\delta}_k - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\delta}_k^{\mathbf{T}}\mathbf{D}_k\boldsymbol{\delta}_k \text{ s.t. } \mathbf{T}^{\mathbf{T}}\mathbf{U}_k\boldsymbol{\delta}_k = \mathbf{0} \tag{13}$$

where $\mathbf{U}_k$=matrix consisting of the 1st k columns of $\mathbf{U}$, and $\mathbf{D}_k$=top right $k \times k$ submatrix of $\mathbf{D}$.

It should be pointed out that the smooth function with thin plate regression splines is isotropic and good for smoothing variables with the same unit. The curious reader is referred to Wood (2003, 2006) for additional details on how to derive thin plate regression splines.

Although GAMs are more flexible than GLMs, they are still subjected to a few limitations. First, because GAMs include more parameters, the estimation of the coefficients could become very complex, especially when the default values (in the statistical software package) are not used. Second, since GAMs use spline functions, the estimated coefficients may not be clearly presented or defined. Third, the modeling results between GAMs and GLMs are likely to be similar if the covariates are "truly" independent and the dependent variable has a linear or exponential relationship with the covariates. In this case, there is no advantage in using GAMs over GLMs, and it is possible that GAMs will over fit the data.

Model Development

In this paper, the GAMs were estimated using the multiple smoothing parameter estimation by generalized cross validation (MGCV) package in the software R (Wood 2008). The MGCV package provides smooth function with thin plate regression splines and offers a great tool for estimating GAMs. Since the same data set as in Lord and Bonneson (2007) was used, the results for the GLMs were taken directly from their paper. Nonetheless, to validate the results documented in their paper, GLMs were estimated using the MASS statistical package (Ripley 2008). The same models were reproduced.

Table 3. Correlation Matrix for the Covariates

Variables[a]	F	LW	SW	EM$\times I_2$
F	1	−0.28	0.03	−0.19
LW		1	−0.46	0.32
SW			1	0.28
EM$\times I_2$				1

[a]F=ADT (veh/day); LW=lane width, ft; SW=combined shoulder width, ft; EM=presence of edge marking; and I_2=indicator variable (1=two-way, 0=one-way).

It should be pointed out that estimating GAMs is a little more complex than estimating GLMs because GAMs include more parameters. For example, the parameter γ has influence on the order of the regression splines of GAMs. To simplify the model development, the default values in R were used. For readers not familiar with the work of Lord and Bonneson (2007), the original functional form was described as follows:

$$\ln(\mu_i) = \ln(\beta_0) + \ln(L_i O_i) + \beta_1 \ln(F_i) + \beta_2 \mathrm{LW}_i + \beta_3 \mathrm{SW}_i + \beta_4 \mathrm{EM}_i \times I_2^i \tag{14}$$

where μ_i=estimated number of crashes for segment i; O_i=number of years during which the crash data were collected; F_i=average daily traffic (ADT) for segment i, veh/day; L_i=length of segment i, mile; LW_i=lane width of segment i, ft; SW_i=combined shoulder width (left+right shoulders) of segment i, ft; EM_i=presence of pavement edge markings: left edge marking (0.5=yes, 0=no)+right edge marking (0.5=yes, 0=no); and I_{i2}=indicator variable (=1.0 for two-way operation, 0.0 for one-way operation).

Table 3 shows the correlation matrix of covariates for the GLM model. This table shows that lane and shoulder widths are not independent, but they are not highly correlated either. Thus, the GAM may not overfit the data. For the GAMs, the following functional form, which includes the smooth terms for each explanatory variable, was initially evaluated:

$$\ln(\mu_i) = \ln(\beta_0) + \ln(L_i O_i) + s_1(F_i) + s_2(\mathrm{LW}_i) + s_3(\mathrm{SW}_i) + s_4(\mathrm{EM}_i \times I_2^i) \tag{15}$$

where $s(.)$=smooth function with thin plate regression splines. Since the coefficients s_2 and s_3 were not significant and the interaction between the key variables was an important study objective, the following functional form was used:

$$\ln(\mu_i) = \ln(\beta_0) + \ln(L_i O_i) + s_1(F_i) + s_5(\mathrm{LW}_i, \mathrm{SW}_i) + s_4(\mathrm{EM}_i \times I_2^i) \tag{16}$$

where $s(.)$=smooth function with thin plate regression splines. In this model, the degrees of freedoms for the terms $s_1(F_i), s_4(\mathrm{EM}_i \times I_2^i)$ were made equal to 1 (since modeling output showed that the degrees of freedom were close to 1). This means that in the model, $s_1(F_i)$ could be replaced with variable F_i, as well as $s_4(\mathrm{EM}_i \times I_2^i)$ with variable $\mathrm{EM}_i \times I_2^i$.

The modeling results using the functional form described in Eq. (16) are shown in Eqs. (17*a*) and (17*b*) for total crashes and injury crashes, respectively. More details are summarized in Table 4. Note that the estimated smooth function $s(\mathrm{LW}_i, \mathrm{SW}_i)$ cannot be presented using a single value. The values estimated from the functions show the contribution of each term to the

Table 4. Frontage-Road Segment Safety Performance Function for Total Crashes and Injury Crashes, Poisson Regression (GAM)

Model variable	Coefficient value (standard error)	Statistic (t or F)	p-value
Model for injury crashes (KABC)			
Intercept[$\ln(\beta_0)$]	−7.3587 (0.7612)	$t=-9.667$	2×10^{-16}
$\ln(\text{ADT})(\beta_1)$	0.7754 (0.1065)	$t=7.280$	5.02×10^{-11}
EDGETWOWAY[a] (β_2)	−0.5947 (0.2549)	$t=-2.333$	0.0214
Smooth function of lane width and combined shoulder width[b]	Estimated degree of freedom=2.594	$F=2.124$	0.056
Model for total crashes (KABCO)			
Intercept[$\ln(\beta_0)$]	−6.0167 (0.6035)	$t=-9.970$	2×10^{-16}
$\ln(\text{ADT})(\beta_1)$	0.6411 (0.0858)	$t=7.476$	1.83×10^{-11}
EDGETWOWAY[a] (β_2)	−0.5185 (0.2027)	$t=-2.558$	0.0119
Smooth function of lane width and combined shoulder width[b]	Estimated degree of freedom=2	$F=4.405$	0.0144

[a]EDGETWOWAY=edge marking presence×two-way operation ($EM\times I_2$); edge marking presence=lane edge marking (0.5=yes, 0=no)+right edge marking (0.5=yes, 0=no); two-way operation (1=yes, 0=no).

[b]Combined shoulder width=paved left shoulder width+paved right shoulder width (ft).

estimated crash number, and can be used to produce AMFs, shown in Tables 5 and 6, and explained in the next section

$$\text{KABCO:}\mu_i = 0.002\,438 \times L_i O_i F_i^{0.6411} e^{[s(\text{LW}_i,\text{SW}_i)-0.5185\text{EM}_i\times I_2^i]} \quad (17a)$$

$$\text{KABC:}\mu_i = 0.000\,637 \times L_i O_i F_i^{0.7754} e^{[s(\text{LW}_i,\text{SW}_i)-0.5947\text{EM}_i\times I_2^i]} \quad (17b)$$

We compared the goodness-of-fit performances of the GLM and GAM models, using the Akaike's information criterion (Akaike 1974), the mean abstract error and mean squared error (Oh et al. 2003). All three measures showed that the GAM and GLM models performed equally.

Description of AMFs

Four AMFs were derived from the GAMs, two each for KABCO and KABC models. AMFs were developed using the coefficients shown in Eqs. (17*a*) and (17*b*). The AMFs produced from the KABCO model were also compared with the AMFs developed in Lord and Bonneson (2007).

AMF for Lane Width and Shoulder Width

Since AMFs have never been previously estimated using GAMs, there is no existing formula that can be used to compute the AMF for the interaction between lane and shoulder widths. For this interaction, the AMF was estimated using Eq. (18). The baseline conditions reflected 12-ft lane width, and an average-shoulder

Table 5. AMF Values for Lane Width and Shoulder Width for Total Crashes

	LW[a]								
ASW[b]	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0
0.0	1.854	1.688	1.536	1.398	1.273	1.158	1.054	0.960	0.874
0.5	1.821	1.658	1.509	1.374	1.250	1.138	1.036	0.943	0.858
1.0	1.790	1.629	1.483	1.350	1.228	1.118	1.018	0.926	0.843
1.5	1.758	1.600	1.457	1.326	1.207	1.099	1.000	0.910	0.829
2.0	1.727	1.572	1.431	1.303	1.186	1.079	0.982	0.894	0.814
2.5	1.697	1.545	1.406	1.280	1.165	1.060	0.965	0.879	0.800
3.0	1.667	1.518	1.382	1.258	1.145	1.042	0.948	0.863	0.786
3.5	1.638	1.491	1.357	1.236	1.125	1.024	0.932	0.848	0.772
4.0	1.610	1.465	1.334	1.214	1.105	1.006	0.915	0.833	0.758
4.5	1.581	1.439	1.310	1.193	1.086	0.988	0.899	0.819	0.745
5.0	1.554	1.414	1.287	1.172	1.067	0.971	0.884	0.804	0.732
5.5	1.527	1.389	1.265	1.151	1.048	0.954	0.868	0.790	0.719
6.0	1.500	1.365	1.243	1.131	1.030	0.937	0.853	0.776	0.707
6.5	1.474	1.341	1.221	1.111	1.012	0.921	0.838	0.763	0.694
7.0	1.448	1.318	1.199	1.092	0.994	0.905	0.823	0.749	0.682
7.5	1.422	1.295	1.178	1.073	0.976	0.889	0.809	0.736	0.670
8.0	1.397	1.272	1.158	1.054	0.959	0.873	0.795	0.723	0.659
8.5	1.373	1.250	1.138	1.035	0.943	0.858	0.781	0.711	0.647

[a]LW=lane width (ft).

[b]ASW=average shoulder width (ft) (both sides).

Table 6. AMF Values for Lane Width and Shoulder Width for Injury Crashes

	LW[a]								
ASW[b]	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0
0.0	2.046	1.844	1.660	1.493	1.341	1.205	1.083	0.973	0.876
0.5	1.995	1.798	1.618	1.455	1.308	1.175	1.056	0.949	0.854
1.0	1.943	1.750	1.575	1.417	1.273	1.144	1.028	0.925	0.833
1.5	1.890	1.702	1.532	1.377	1.238	1.112	1.000	0.900	0.811
2.0	1.836	1.653	1.487	1.337	1.202	1.080	0.972	0.875	0.789
2.5	1.782	1.603	1.442	1.296	1.165	1.048	0.944	0.851	0.768
3.0	1.729	1.554	1.397	1.256	1.129	1.017	0.916	0.827	0.748
3.5	1.676	1.506	1.353	1.216	1.095	0.986	0.890	0.805	0.729
4.0	1.627	1.461	1.312	1.180	1.062	0.958	0.866	0.784	0.711
4.5	1.581	1.419	1.274	1.146	1.032	0.932	0.844	0.765	0.695
5.0	1.540	1.382	1.241	1.116	1.007	0.910	0.825	0.749	0.682
5.5	1.505	1.350	1.212	1.091	0.985	0.892	0.809	0.736	0.671
6.0	1.476	1.324	1.189	1.071	0.968	0.878	0.798	0.726	0.662
6.5	1.453	1.304	1.172	1.057	0.956	0.868	0.789	0.719	0.657
7.0	1.437	1.290	1.161	1.048	0.949	0.862	0.785	0.715	0.653
7.5	1.426	1.282	1.155	1.043	0.946	0.859	0.783	0.714	0.652
8.0	1.421	1.279	1.153	1.043	0.946	0.860	0.783	0.714	0.652
8.5	1.421	1.280	1.156	1.047	0.950	0.864	0.786	0.717	0.654

[a]LW=lane width (ft).

[b]ASW=average shoulder width (ft) (both sides).

width equal to 1.5 ft [(left shoulder and right shoulder widths)/2] of the original sample, as described in Lord and Bonneson (2007). In Eq. (18), the AMF is equal to 1 for the baseline condition. It should be noted that $AMF_{LW,SW}$ in Eq. (18) is equal to $\mu(LW_i, ASW_i)/\mu(LW_i=12, ASW_i=1.5)$ computed with Eqs. (17*a*) and (17*b*), respectively

$$AMF_{LW,SW} = e^{s(LW_i,ASW_i)-s(12,1.5)} \tag{18}$$

where ASW_i=average paved shoulder width in feet for segment i. With Eq. (18), the AMF values calculated for various lane and average shoulder widths are listed in Tables 5 and 6 for KABCO and KABC, respectively. The graphical representation of these AMFs is shown in Fig. 2. Fig. 2(a) also shows the AMF developed by Lord and Bonneson (2007). They assumed the AMFs for lane and average shoulder widths to be independent. They are as follows:

$$KABCO: AMF_{LW} = e^{(-0.188\times[LW_i-12.0])} \tag{19}$$

$$KABCO: AMF_{SW} = e^{(-0.070\times[ASW_i-1.5])} \tag{20}$$

Since AMFs are multiplicative factors when they are used to predict crash frequency, Fig. 2(a) shows the values of $AMF_{LW} \times AMF_{SW}$. Then, the $AMF_{LW,SW}$ developed in this paper can be compared with $AMF_{LW} \times AMF_{SW}$ developed by Lord and Bonneson (2007). Fig. 2 illustrates that GAMs offer better abilities to approximate a nonlinear relationship and model the interaction of covariates more efficiently

- For KABCO, the decreasing rate of AMF with average shoulder width is smaller than that of the AMF developed by Lord and Bonneson (2007); and
- For KABC, when average shoulder width is larger than 6 ft, the curve becomes almost horizontal and starts increasing again around 8 ft, indicating that wider shoulder widths do not reduce crashes.

AMF for Edge Marking Presence on Two-Way Frontage Roads

The AMF derived from the variables associated with the presence of edge line delineation is the following:

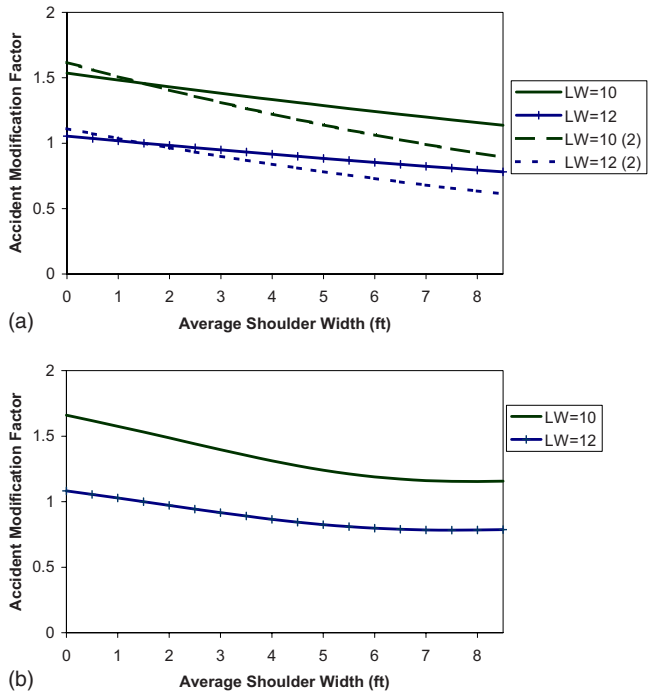
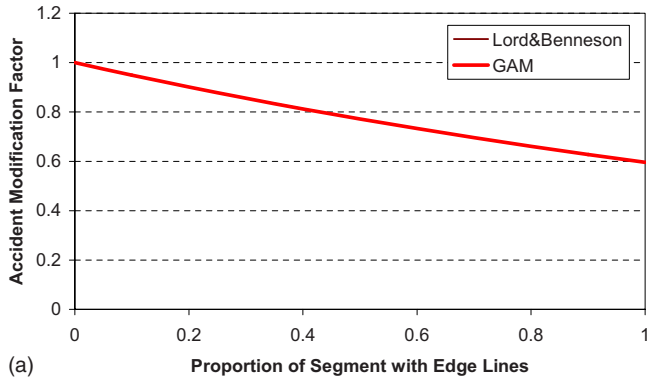


Fig. 2. AMF for lane width and average shoulder width: (a) AMF for total crashes (KABCO); (b) AMF for injury crashes (KABC) [Note: curves produced from results in Lord and Bonneson (2007) are marked with "LW=10(2)" "LW=12(2)"]



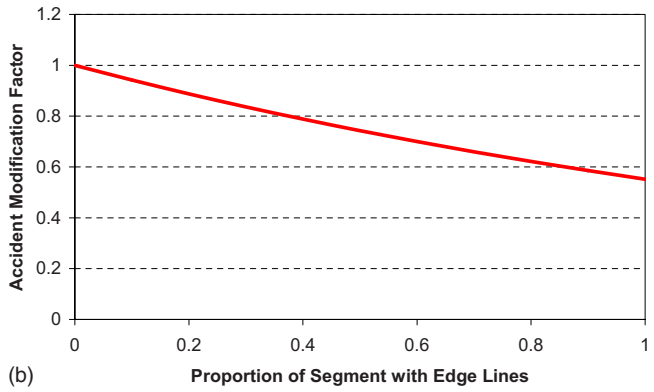


Fig. 3. AMF for edge marking presence on two-way frontage roads: (a) AMF for total crashes (KABCO); (b) AMF for injury crashes (KABC) [Note: the curves in Fig. 3(a) overlap each other]

$$\text{KABCO:AMF}_{\text{EM}} = e^{(-0.5185\text{EM}_i)} \tag{21a}$$

$$\text{KABC:AMF}_{\text{EM}} = e^{(-0.5947\text{EM}_i)} \tag{21b}$$

where EM=proportion of segments with pavement edge markings (two-way frontage road). This AMF was derived for two-way frontage roads. In Lord and Bonneson (2007), the same AMF is given by the following:

$$\text{KABCO:AMF}_{\text{EM}} = e^{(-0.518\text{EM})} \tag{22}$$

The graphical representations of these AMFs are shown in Fig. 3. For KABCO, the two AMFs are almost the same, as expected. Since these variables are used the same way for both models, the coefficient should be similar.

The curve in Fig. 3 suggests that edge markings can reduce severe crashes on two-way frontage road segments by about 40%. As indicated in Lord and Bonneson (2007), this AMF explains more than just the effect of the presence of pavement markings on crash frequency because the markings is likely to be accompanied by additional warning signs that denote two-way operations.

Safety Performance Function

Using the regression model in Eq. (17*b*), a baseline SPF linking the number of crashes to the traffic flow for KABC can be derived as follows:

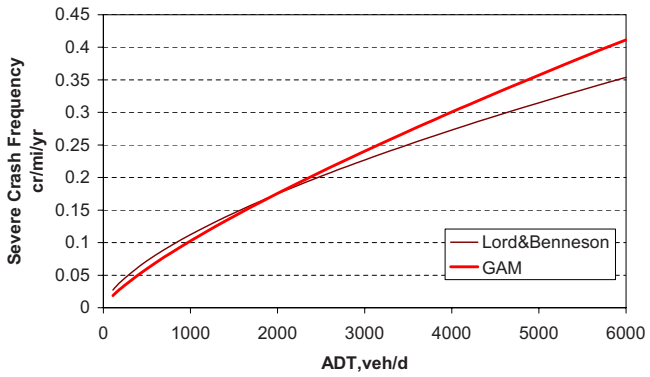


Fig. 4. Comparison of SPF developed from the GAM and SPF estimated by Lord and Bonneson (2007)

$$\begin{aligned}\mu_{\text{baseline}\_i} &= 0.000\,637 \times L_i F_i^{0.7754} e^{s(\text{LW}_i,\text{SW}_i)} \\ &= 0.000\,637 \times L_i F_i^{0.7754} e^{-0.187\,083\,3} \\ &= 0.000\,528 \times L_i F_i^{0.7754}\end{aligned} \tag{23}$$

where μ_{baseline_i}=estimated number of injury crashes per year for segment i for the base conditions: lane width equal to 12 ft, combined paved shoulder width of 3 ft (summation of both shoulder widths), and no edge markings on frontage road segments. In order to compare this curve with the one taken from Lord and Bonneson (2007), we used the same shoulder width and excluded PDO collisions; this type of collisions was removed from their study because the results were compared with another predictive model published in the literature that only included injury crashes. Thus, to be consistent with the previous work, we also removed PDO collisions. The final predicted value (say for segment i) after AMFs are applied can be computed with the following equation:

$$\mu_{\text{final}\_i} = \mu_{\text{baseline}\_i} \times \text{AMF}_{\text{LW,SW}} \times \text{AMF}_{\text{EM}} \tag{24}$$

The graphical representation of the SPF, using Eq. (23), is shown in Fig. 4, which also presents the SPF developed by Lord and Bonneson (2007). The two curves show that for ADTs greater than 2,500 veh/day, the SPF in this paper is slightly larger than the one developed by Lord and Bonneson (2007), while both curves get closer together for ADTs less than 2,500 veh/day. Overall, the difference in predicted values is not significant if one examines the 95% confidence intervals (see Lord et al. 2010), which means that both models could be used for predicting crashes. However, the analysis shows that, although both models performed equally, lane and shoulder widths should not be considered as independent.

Summary and Conclusions

The primary objective of this study consisted of describing the application of GAMs for estimating AMFs. GAMs are a new type of models that have been recently introduced by the statistical community for modeling observed data. These models offer more flexible functional forms than traditional GLMs and allow for more adaptable variable interactions. As reported in the literature, the interaction between variables should be included in the development of AMFs (Bonneson et al. 2007; Gross et al. 2009) and GAMs allow for such interaction. To accomplish the study objec-

J. Transp. Eng., 2011, 137(1): 74-83

tive, AMFs were derived from GAMs using data collected on rural frontage roads in Texas. The AMFs were then compared to the AMFs documented in a previous study performed by Lord and Bonneson (2007).

The results of the study show that AMFs produced from GAMs are more flexible to characterize the safety effect of simultaneous changes in geometric and operational features than when independent AMFs are applied together. The results also show that GAMs allow for a nonlinear relationship between crash risk and changes in roadway features. Furthermore, lane and shoulder widths were found to be not completely independent. Compared to the results documented in Lord and Bonneson (2007), the AMFs and SPF provided similar values with the exception of the following: the decreasing rate for the average shoulder width KABCO AMF is smaller; for an average shoulder width larger than 6 ft, the decreasing rate for the KABC AMF becomes horizontal and starts increasing at around 8 ft, which indicates no safety gains are obtained for wider widths; and for traffic flow greater than 2,500 veh/d, the SPF produced from GAMs predicts slightly larger values. The results presented here are different than those reported by Gross and Jovanis 2007a,b), and Gross et al. (2009) for the same variables (i.e., lane and shoulder widths, although those were for nonfrontage rural roads).

Although GAMs offered a useful and innovative approach for estimating AMFs, further work is needed on this topic. It includes applying GAMs to other data sets to confirm the results in this study and determine what types of nonlinear relationship could exist between crashes and design and operational features for other highway facilities. Finally, given the recent work on this topic, developing multivariate GAMs to predict crash frequency for different severity levels, and extract AMFs for each severity level, using the response variable as a vector should be examined.

Acknowledgments

Although the study was not funded by TxDOT, the writers wish to thank Ms. Elizabeth Hilton from TxDOT and Dr. James A. Bonneson for providing the data. The data were initially collected for TxDOT Project 0-4703 led by Dr. Bonneson. The writers would like to thank the anonymous reviewers for providing useful comments aimed at improving this paper.

References

Akaike, H. (1974). "A new look at the statistical model identification." *IEEE Trans. Autom. Control*, 19(6), 716–723.

Bahar, G., et al. (2007). "Prepare parts I and II of the highway safety manual." *Final Rep. Prepared for NCHRP Project 17-27*, iTRANS Consulting, Richmond Hill, Ont., Canada.

Bonneson, J., Lord, D., Zimmerman, K., Fitzpatrick, K., and Pratt, M. (2007). "Development of tools for evaluating the safety implications of highway design decisions." *TTI Rep. No. FHWA/TX-07/0-4703-4*, Texas Transportation Institute, College Station, Tex.

Bonneson, J., and Pratt, M. (2008). "Calibration of safety prediction models and AMFs for urban and suburban arterial street segments in Texas." *Draft Technical Memorandum No. 14*, *TxDOT Project 0-4703*, Texas Dept. of Transportation, Austin, Tex.

Bonneson, J. A., and Lord, D. (2005). "Role and application of accident modification factors in the highway design process." *TTI Rep. No. FHWA/TX-05/0-4703-2*, Texas Transportation Institute, College Station, Tex.

Cameron, A. C., and Trivedi, P. K. (1998). "Regression analysis of count data." *Econometric Society Monograph No. 30*, Cambridge University Press, New York.

Elvik, R. (2009). "Developing accident modification functions: Exploratory study." *Proc., 88th Annual Meeting of the Transportation Research Board*, No. 09-0299, Washington, D.C.

Fitzpatrick, K., Lord, D., and Park, B.-J. (2008). "Accident modification factors for medians on freeways and multilane highways." *Transp. Res. Rec.*, 2083, 62–71.

Fitzpatrick, K., Lord, D., and Park, B.-J. (2009). "Horizontal curve accident modification factors with consideration of driveway density on rural, four-lane highways in Texas." *Proc., 88th Annual Meeting of the Transportation Research Board*, No. 09-0204, Washington, D.C.

Gross, F., and Jovanis, P. P. (2007a). "Estimation of the safety effectiveness of lane and shoulder width: Case-control approach." *J. Transp. Eng.*, 133(6), 362–369.

Gross, F., and Jovanis, P. P. (2007b). "Estimation of safety effectiveness of changes in shoulder width with case control and cohort methods." *Transp. Res. Rec.*, 2019, 237–245.

Gross, F., Jovanis, P. P., and Eccles, K. A. (2009). "Safety effectiveness of lane and shoulder width combinations on rural, two-lane, undivided roads." *Proc., 88th Annual Meeting of the Transportation Research Board*, No. 09-1294, Washington, D.C.

Harwood, D. W., Council, F. M., Hauer, E., Hughes, W. E., and Vogt, A. (2000). "Prediction of the expected safety performance of rural two-lane highways." *Rep. No. FHWA-RD-99-207*, Federal Highway Administration, Washington, D.C.

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized additive models*, Chapman and Hall, New York.

Hauer, E. (1997). *Observational before-after studies in road safety*, Pergamon, Elsevier Science, Oxford, U.K.

Hauer, E. (2004). "Statistical road safety modeling." *Transp. Res. Rec.*, 1897, 81–87.

Hughes, W., Eccles, K., Harwood, D., Potts, I., and Hauer, E. (2005). "Development of a highway safety manual. Appendix C: Highway safety manual prototype chapter: Two-lane highways." *NCHRP Web Document 62 [Project 17-18(4)]*, ⟨http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w62.pdf⟩ (July 2008).

Li, X., Lord, D., Zhang, Y., and Xie, Y. (2008). "Predicting motor vehicle crashes using support vector machine models." *Accid. Anal Prev.*, 40(4), 1611–1618.

Lord, D., and Bonneson, J. A. (2006). "Role and application of accident modification factors (AMFs) within the highway design process." *Transp. Res. Rec.*, 1961, 65–73.

Lord, D., and Bonneson, J. A. (2007). "Development of accident modification factors for rural frontage road segments in Texas." *Transp. Res. Rec.*, 2023, 20–27.

Lord, D., Kuo, P.-F., and Geedipally, S. R. (2010). "Comparing the application of the product of baseline models and accident modification factors and models with covariates: Predicted mean values and variance." *Transp. Res. Rec.*, in press.

Miaou, S. P. (1994). "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions." *Accid. Anal Prev.*, 26(4), 471–482.

Myers, R. H. (2000). *Classical and modern regression with applications*, 3rd Ed., Duxbury, Belmont, Calif.

Oh, J., Lyon, C., Washington, S., Persaud, B., and Bard, J. (2003). "Validation of FHWA crash models for rural intersections: Lessons learned." *Transp. Res. Rec.*, 1840, 41–49.

Persaud, B. N., Retting, R., Garder, P., and Lord, D. (2001). "Observational before-after study of U.S. roundabout conversions using the empirical Bayes method." *Transp. Res. Rec.*, 1751, 1–8.

Ripley, B. (2008). "The VR package (version 7.2–42)." ⟨http://cran.r-project.org/web/packages/VR/VR.pdf⟩ (May 22, 2008).

Shen, J., and Gan, A. (2003). "Development of crash reduction factors: Methods, problems and research needs." *Transp. Res. Rec.*, 1840, 50–56.

Washington, S., Persaud, B., Lyon, C., and Oh, J. (2005). "Validation of accident models for intersections." *FHWA-RD-03-037*, Federal Highway Administration, Washington, D.C.

Wood, S. (2008). "The MGCV package (version 1.3–31)." ⟨http://cran.r-project.org/web/packages/mgcv/mgcv.pdf⟩ (May 22, 2008).

Wood, S. N. (2003). "Thin plate regression splines." *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, 65, 95–114.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*, Chapman and Hall/CRC, Boca Raton, Fla.

Xie, Y., Lord, D., and Zhang, Y. (2007). "Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis." *Accid. Anal Prev.*, 39(5), 922–933.

Xie, Y., and Zhang, Y. (2008). "Crash frequency analysis with generalized additive models." *Transp. Res. Rec.*, 2061, 39–45.

Ye, Z., and Lord, D. (2009). "Estimating the variance in before-after studies." *J. Safety Res.*, 40(4), 257–263.