Hierarchical Bayesian Estimation of Safety Performance Functions for Two-Lane Highways Using Markov Chain Monte Carlo Modeling

Xiao Qin[1]; John N. Ivan[2]; Nalini Ravishanker[3]; and Junfeng Liu[4]

Abstract: A critical part of any risk assessment is identifying how to represent exposure to the risk involved. Recent research shows that the relationship between crash count and traffic volume is nonlinear; consequently, a simple crash rate computed as the ratio of crash count to volume is not suitable for comparing the safety of sites with different traffic volumes. To solve this problem, we describe a new approach for relating traffic volume and crash incidence. Specifically, we disaggregate crashes into four types: (1) single-vehicle, (2) multivehicle same direction, (3) multivehicle opposite direction, and (4) multivehicle intersecting, and then define candidate exposure measures for each (as a function of site traffic volumes) that we hypothesize will be linear with respect to each crash type. This article describes investigation using crash and physical characteristics data for highway segments from Michigan, California, Washington, and Illinois obtained from the Highway Safety Information System. We have used a hierarchical Bayesian framework to fit zero-inflated-Poisson regression models for predicting counts for each of the above crash types as a function of the daily volume, segment length, speed limit and lane/shoulder width using Markov Chain Monte Carlo methods. We found that the relationship between crashes and the daily volume is nonlinear and varies by crash type, and is significantly different from the relationship between crashes and segment length for all crash types. Significant differences in exposure functions by crash type are proven using analysis of variance and Tukey tests.

DOI: 10.1061/(ASCE)0733-947X(2005)131:5(345)

CE Database subject headings: Traffic accidents; Algorithms; Highways; Rural areas; Traffic safety; Risk management; Michigan; California; Washington.

Introduction

A critical part of any risk assessment is identifying the appropriate measure of exposure to the risk in question. It is useful to cast the crash prediction problem by defining number of crashes to be the product of exposure to the risk and the risk of a crash in which vehicles may be involved. However, only the number of crashes is an observable value, as neither risk nor exposure to risk is self-explanatory, and each is dependent on how the other is defined. Therefore, evaluating safety between different travel modes, or comparing site crash risks is somewhat arbitrary and varies by how one defines exposure.

This concept of exposure was introduced to highway safety analysis through the definition of the crash rate—the number of crashes divided by the number of million vehicle-miles traveled (VMT). The development of this approach was important for identifying truly hazardous locations, as opposed to locations with high volumes but a low number of crashes per vehicle. However, a shortcoming of this approach is that it assumes a linear relationship between the number of crashes and the VMT. In fact, this relationship may not be simply linear, and arbitrarily assuming linearity may weaken the accuracy of the crash count prediction. As an alternative, Hauer introduced the concept of safety performance functions to generalize the representation of exposure as a non-linear function of the traffic volume (Hauer 1995); this concept is the basis for the research described in this paper.

Logically, exposure can be thought of as a statistical measure that provides information on the extent of travelers' vulnerability to the surrounding crash risk environment. In other words, the more a traveler is exposed to crash risks, the greater is the possibility for his involvement in a crash. For a single entity, the amount of exposure can be measured by either the length of time exposed, or by the length of the trip. For a highway segment, the exposure is closely associated with the number of entities exposed during a time period, usually 1 year, and the distance over which an entity is exposed to the crash risks. Hence, annual average daily traffic (AADT) and segment length are regarded as components of exposure in crash prediction.

Moreover, different crash types have different risks. For single vehicle crashes, the risk is related to the probability for any given vehicle to run off the road, collide with roadside objects or roll over. For multiple vehicle crashes, however, the probability of the

[1] Associate Researcher, Dept. of Civil and Environmental Engineering, Univ. of Wisconsin-Madison, Madison, WI 53706. E-mail: xqin@engr.wisc.edu; formerly, ITS/Safety Engineer, Maricopa Association of Governments, 302 North 1st Ave., Ste. 300, Phoenix, AZ 85003.

[2] Associate Professor, Dept. of Civil & Environmental Engineering, Univ. of Connecticut, Unit 2037, Storrs, CT 06269-2037. E-mail: john.ivan@engr.uconn.edu

[3] Professor, Dept of Statistics, Univ. of Connecticut, Unit 4120, Storrs, CT 06269. E-mail: nalini.ravishanker@uconn.edu

[4] Research Assistant, Dept. of Statistics, Univ. of Connecticut, Unit 4120, Storrs, CT 06269. E-mail: junfeng.liu@uconn.edu

Note. Discussion open until October 1, 2005. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on September 29, 2003; approved on October 20, 2004. This paper is part of the ***Journal of Transportation Engineering***, Vol. 131, No. 5, May 1, 2005. ©ASCE, ISSN 0733-947X/2005/5-345–351/$25.00.

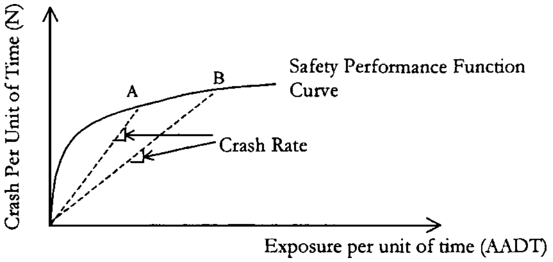
Crash Per Unit of Time (N)
A
B
Safety Performance Function Curve
Crash Rate
Exposure per unit of time (AADT)

Fig. 1. Example safety performance function

collision depends on the occasions when vehicles cross paths, meet or follow one another. For example, given the same AADT, the risk for a same-direction vehicle crash may be different from that for an opposite-direction crash; this is a result of the varying number of vehicles in each direction. Therefore, the exposure measure is necessarily dependent on crash type and an analysis of variance (ANOVA) test is proposed with the null hypothesis that the exposure function is the same for all the crash types. Furthermore, a Tukey test will be employed to investigate the variation within the crash types.

The problem of accurately modeling the true risk of crashes occurring at a highway location is further exacerbated by overdispersion and the regression-to-mean effect. This effect has been well documented by many, especially Hauer, who derived an empirical Bayesian approach for estimating the true mean crash rate for a location (Hauer 1997). Tunaru improved upon Hauer's method, using a hierarchical Bayesian generalized linear modeling approach for multiple crash responses at a location (Tunaru 1999). These two landmark research efforts provide welcome examples of how to more accurately evaluate highway safety, yet they both still compute crash rates using VMT.

Extending these ideas, this paper describes investigation into using hierarchical Bayesian modeling to estimate safety performance functions that best represent the risk of exposure to four different types of highway crash: (1) single vehicle, (2) multivehicle same direction, (3) multivehicle intersecting direction, and (4) multivehicle opposite direction. To focus the study, analysis is confined to two-lane rural highway segments in four United States states: Michigan, California, Washington, and Illinois. Moreover, different prediction model forms are tested with time effects, the interactions between time and AADT, the interaction between time and segment length and the linear relationship between exposure measures and segment length. The models are estimated separately for each state to show both model consistency and differences among states using an ANOVA test with the null hypothesis that for the same crash type, there is no regional (state) effect on exposure functions.

Safety Study Design

Fig. 1 depicts a hypothetical nonlinear safety performance function for a highway segment. For any point on the safety performance function curve, the crash rate (N/AADT), is defined as the slope of the line joining the origin to that point (as indicated). Therefore, if the safety performance function is not a straight line, the crash rate varies with the amount of the exposure. For example, the number of crashes at Point B is greater than at Point A, but the crash rate at Point B, conversely, is smaller than at Point A, because the slope of the line joining it to the origin is less steep. From the point of view of highway safety engineers, this crash rate change due only to a change in exposure should not be regarded as an improvement in the site safety because there is no change in the physical characteristics of the site.

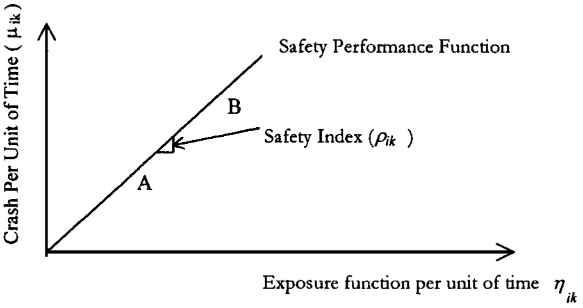
Crash Per Unit of Time (μ_{ik})
Safety Performance Function
B
Safety Index (ρ_{ik})
A
Exposure function per unit of time η_{ik}

Fig. 2. Normalized safety performance function

Instead, it is helpful to define a new exposure function to transform the traffic volume into an exposure measure $f(V)$, yielding a linear relationship between crashes and exposure. After implementing the new coordinate, the safety performance function becomes linear and each point on the line has the same slope, representing a normalized crash rate that is constant for all levels of exposure at the same location. This newly defined crash propensity, or safety index, is therefore more meaningful for making comparisons among different entities with different exposures and safety performance functions.

In keeping with the nonlinear relationship for the safety performance function, we use an exponential form for estimating safety indices for each crash type ρ_{ik} as follows:

$$\mu_{ik} = \eta_{ik}\rho_{ik} \tag{1}$$

where μ_{ik}=expected number of crashes for crash type k at segment i; η_{ik}=computed exposure function at segment i for crash type k; and ρ_{ik}=normalized crash rate for crash type k at segment i, also defined as the safety index.

Further, we define functions for η_{ik} and ρ_{ik} as follows:

$$\eta_{ik} = \eta_k(V_i, L_i) = V_i^{\alpha_{Vk}} L_i^{\alpha_{Lk}} \tag{2}$$

and

$$\rho_{ik} = \exp(\mathbf{X}_i\boldsymbol{\beta}) \tag{3}$$

where V_i=AADT at segment i; L_i=length of segment i; $\mathbf{X}_i$=vector of road characteristics for segment i; and $\boldsymbol{\beta}$=entire vector of parameters to be estimated for crash type k.

Fig. 2 shows the resulting safety performance function for crash type k at segment i as a straight line with slope equal to the safety index ρ_{ik}.

Data Collection and Processing

We gathered records of traffic counts and crash incidence (by type) from the Highway Safety Information System (HSIS) database maintained by the Federal Highway Administration (FHWA) of the United States Department of Transportation. In this study, we include data from four states: Michigan, California, Washington, and Illinois; summary statistics are listed in Table 1. As would be expected for rural road segments, more than half of the crashes are of the single-vehicle type. Opposite direction crashes are the least represented. The average speed limit is around 80 km/h (50 mph), with a maximum of 104 km/h (65 mph), indicating that there is not a great deal of variation among states.

Table 1. Variable Definitions and Summary Statistics for Four States

	Michigan			California			Washington			Illinois		
Variable[a]	Minimum	Maximum	Mean	Minimum	Maximum	Mean	Minimum	Maximum	Mean	Minimum	Maximum	Mean
SV	0	61	0.68	0	29	0.61	0	20	0.32	0	30	0.27
SD	0	23	0.15	0	22	0.2	0	10	0.1	0	28	0.08
OD	0	7	0.04	0	11	0.08	0	4	0.05	0	4	0.01
ID	0	23	0.08	0	15	0.28	0	7	0.04	0	20	0.04
L (m)	16	12,585	998	2	9,640	322	16	10,267	177	16	14,194	193
V (1000s)	0.24	40	5.45	0.05	28	3.05	0.05	24	2.39	0.26	25.7	5.45
W (m)	6	13	12	6	15	10	5	15	9	5	15	10
S (kph)	40	89	85	40	105	87	48	105	84	40	97	77

Note: SV=single vehicle crashes; SD=multivehicle same direction crashes; OD=multivehicle opposite direction crashes; ID=multivehicle intersecting direction crashes; L=segment length (m); V=annual average daily traffic (in 1000s of vehicles); W=pavement width (m); and S=speed limit.

[a] (km/h)

There is, however, considerable variation in the pavement width, although it is mostly due to the differences in shoulder width. There is a wide range in segment lengths, although most segments appear to be less than 1.6 km (1 mile) in length; consequently it is vital that the models account for differences in length from one segment to another. Moreover, the number of years for the study segments is different for each state. Michigan and California have 5 years of data from 1993 to 1997. Illinois has 3 years of data corresponding to 1991, 1992, and 1994. Washington has 4 years of data from 1993 to 1996.

The data for each state were available in two separate databases: one containing information about each crash, and the second containing information about each highway segment. Consequently, it was necessary to process the data into a different format before they could be used for estimating statistical models. This required linking the accident and highway inventory databases, filtering out observations with missing or illogical values from each database (such an AADT or segment length of 0), translating the given accident type variables into a new variable that matches our definition, and finally aggregating the multiple cases for each accident into a single case for each segment, with accident counts and AADTs by year.

Methodology

In this section, we describe the zero-inflated-Poisson (ZIP) regression model under the hierarchical Bayesian framework for the crash data. The following section presents the details of the model structure, the while the "Bayesian Approach for Inference" section describes the Bayesian approach for estimation and inference.

Zero-Inflated-Poisson Crash Prediction Model

While previous studies have provided insight into the factors determining crash frequencies, it is important to realize that traditional application of the Poisson or negative binomial distribution alone does not address the possibility that more than one underlying process may be influencing crash frequencies (Miaou and Lum 1993). For instance, if the study segments are collected randomly, a preponderance of zero-crash observations will appear in the data because crashes are rare events. This over-representation of zero-crash observations in the data may erroneously suggest overdispersion in the data even though the Poisson distribution is actually otherwise correct. To account for the large probability "spike" at zero, P_i is used to represent the additional probability of segment i to have no crashes while $1-P_i$ represents the probability that segment i follows the Poisson distribution. Assuming a Poisson distribution, the probability that a segment will have no crashes (apart from the additional spike) is $e^{-\mu_i}$. The total probability of observing zero crashes consists of mixing these two probabilities together. The entire distribution is called the ZIP distribution with the following probability density function (Lambert 1992):

$$P(N_i) = \begin{cases} P_i + (1-P_i)e^{-\mu_i} & (N_i = 0) \\ (1-P_i)\dfrac{e^{-\mu_i}\mu_i^{N_i}}{N_i!} & (\text{otherwise}) \end{cases} \tag{4}$$

where $P(N_i)$=probability that N_i crashes occur on segment i, assuming values 0,1,2,…; P_i=zero-inflated probability on segment i; and μ_i=expected number of crashes at segment i.

In the framework of generalized linear models, a link function is employed to connect the mean number of crashes with related covariates. As defined previously

$$\mu_i = \eta(V_i,L_i)e^{\mathbf{X}_i\boldsymbol{\beta}} \tag{5}$$

which implies a log-linear link function. As for the ZIP P_i at segment i, we use a logit function as follows:

$$\log it(P_i) = \log\left(\frac{P_i}{1-P_i}\right) = \mathbf{G}_i\boldsymbol{\gamma} \tag{6}$$

where $\mathbf{G}_i$=vector of covariates expected to influence the value of P_i; and $\boldsymbol{\gamma}$=vector of coefficients of the covariates $\mathbf{G}_i$.

The likelihood function of the parameters is maximized in order to estimate all the unknown coefficients given data at n sites, and is

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}|N) = \prod_{i=0}^{n}[P_i + (1-P_i)e^{-\mu_i}](1-P_i)\frac{e^{-\mu_i}\mu_i^{N_i}}{N_i!} \tag{7}$$

We use the covariate vector $\mathbf{X}_i$ and $\mathbf{G}_i$ along with the exposure function of ν and L for predicting both P and μ. Recall that each P, N, ν, and μ are indexed on k; these indices are omitted here for brevity.

Eq. (7) is not a closed form and cannot be solved easily. A solution is possible by the introduction of an auxiliary variable (latent variable) Z in the likelihood function; here

$$Z(i) = \begin{cases} 1 & \text{If } N_i \text{ is from the zero-inflated state} \\ 0 & \text{If } N_i \text{ is from the nonzero-inflated} \\ & \text{(Poisson distribution) state} \end{cases}$$

The procedure of introducing an auxiliary variable into the likelihood function is called data augmentation (Tanner 1993). All data augmentation algorithms share a common approach to problems: rather than performing a complicated maximization or simulation, one augments the observed data with latent data to simplify the calculation. Now the log-likelihood function with the complete data (N,Z) would be (Lambert 1992)

$$LL(\boldsymbol{\gamma},\boldsymbol{\beta};N,Z) = \sum_{i=1}^{n} \{Z_i\mathbf{G}_i\boldsymbol{\gamma} - \log[1+\exp(\mathbf{G}_i\boldsymbol{\gamma})]\} + \sum_{i=1}^{n} (1-Z_i) \times\{[N_i\mathbf{X}_i\boldsymbol{\beta} - \exp(\mathbf{X}_i\boldsymbol{\beta})] - \log(N_i!)\} \qquad (8)$$

Numerical maximization for obtaining the maximum likelihood estimates of the model parameters is done using software such as *SPLUS* (Mathsoft 1995).

Bayesian Approach for Inference

In this section, we describe a fully Bayesian framework for modeling and inference. The Bayesian framework enables the modeler to specify a prior distribution that represents the best guess about the parameters before incorporating information from the data. In general, given data and model parameters, the Bayesian model specification requires a likelihood function and a prior distribution, from which, by Bayes' theorem, we obtain the posterior density of the parameters given the data as being proportional to the product of the likelihood and the prior (up to a normalizing constant). The advantage of this approach over the empirical Bayesian approach is that it takes into account the uncertainty associated with the estimates of the parameters and can provide exact measures of uncertainty. The Empirical Bayesian (EB) estimate of the individual site is based on the assumption that the mean and the variance of the control group are both estimated without errors, which is not true in practice (Miaou and Lord 2003). It also facilitates extensive predictive analysis through the use of numerical summary statistics and graphical displays, such as histograms and density plots for estimated parameters and functions of these parameters.

Review of Sampling Based Bayesian Framework

In some situations, such as under a conjugate prior setup, computation of the posterior and resulting inference is easy, but in general it can be difficult, if not impossible to obtain analytical forms for the joint posterior and related quantities. Numerical integration may be less efficient and may also be difficult to implement in high dimensions. In the last decade, sampling-based methods have gained popularity for Bayesian inference in a variety of problems, their attractiveness being their conceptual simplicity and ease of implementation.

In particular, the Gibbs sampler is a Markovian updating scheme, which requires sampling from the complete conditional distributions associated with the parameter vector $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\gamma})$. A key point is that each complete conditional density is also proportional to the joint posterior. In certain cases, the form corresponds to a standard distribution, while in others it emerges only as a nonstandard, non-normalized density. The customary Gibbs sampler proceeds by making draws from the complete conditional distribution in some systematic order. The Metropolis algorithm is employed when the complete conditional distributions are not easily identified in standard forms (Tanner 1993). This algorithm creates a sequence of random points, whose distribution converges to the target posterior distribution. The final samples from the posterior are obtained after monitoring convergence.

Hierarchical Bayesian Model Fitting

As we will see, a useful offshoot of the sampling based Bayesian framework for modeling crashes is that it enables us to make inferences about functions of parameters (such as differences between parameters) effortlessly, as we describe later. We fit a hierarchical fully Bayesian model using Markov chain Monte Carlo (MCMC) algorithms; in particular, we employ the Metropolis algorithm.

Eq. (7) is used within a hierarchical Bayesian modeling framework using MCMC methods for making inferences on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The advantage of this approach over the empirical Bayesian approach is that it provides the entire posterior distributions of the model parameters, permitting a wide range of inference beyond just the first few moments. The updated uncertainty about the value of these parameters is expressed as posterior distributions as follows:

$$P(\boldsymbol{\beta},\boldsymbol{\gamma}|N,Z) \propto L(\boldsymbol{\beta},\boldsymbol{\gamma}|N,Z)L(\boldsymbol{\gamma}|Z)\phi(\boldsymbol{\beta},\boldsymbol{\gamma}) \propto L(\boldsymbol{\beta},\boldsymbol{\gamma}|N,Z)L(\boldsymbol{\gamma}|Z)\phi_1(\boldsymbol{\beta})\phi_2(\boldsymbol{\gamma}) \qquad (9)$$

where $P(\boldsymbol{\beta},\boldsymbol{\gamma}|N,Z)$=posterior distribution for all the unknown coefficients given the complete data set (N,Z); $L(\boldsymbol{\beta},\boldsymbol{\gamma}|N,Z)$=likelihood function for all the unknown coefficients given the complete data set (N,Z); the same as in Eq. (8); $L(\boldsymbol{\gamma}|Z)$=likelihood function of unknown coefficients $\boldsymbol{\gamma}$ given Z; and $\phi(\boldsymbol{\beta},\boldsymbol{\gamma})$=joint prior distribution of unknown coefficients $(\boldsymbol{\beta},\boldsymbol{\gamma})$. Because the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are independent, their joint prior distribution is the product of their individual distributions $\phi_1(\boldsymbol{\beta})$ and $\phi_2(\boldsymbol{\gamma})$, $\phi_1(\boldsymbol{\beta})$ is the prior distribution of unknown coefficients $\boldsymbol{\beta}$, and $\phi_2(\boldsymbol{\gamma})$ is the prior distribution of unknown coefficients $\boldsymbol{\gamma}$.

The Bayesian framework also provides for specifying a prior distribution that represents the best guess about the parameters. In this case, we do not have sufficient knowledge of the distribution for individual parameters on the risk factors and specify a diffuse prior distribution, which implies a vague specification for ϕ_1 and ϕ_2

$$\boldsymbol{\beta} \sim \text{normal}(\hat{\boldsymbol{\beta}},\sigma^2\mathbf{I}_{\mathbf{q}_1})$$

and (10)

$$\boldsymbol{\gamma} \sim \text{normal}(\hat{\boldsymbol{\gamma}},\sigma^2\mathbf{I}_{\mathbf{q}_2})$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$=initial values decided by experience. In this study, they are estimators from the previous ZIP process. σ^2=very large number and I_N=identity $q\times q$ matrix, and q=number of covariates.

Again, the indices are omitted for brevity. Once we obtain the posterior distributions for these parameters, we can easily construct any summary information of interest, such as the mean, median, or credible intervals. Since the posterior distribution is complicated and cannot be obtained in closed form, we use MCMC algorithms to effectively simulate from the posterior. As an example, for the coefficients vector $\boldsymbol{\beta}$, the idea of the MCMC is to simulate a random move in the space of the unknown parameters $\boldsymbol{\beta}$, which converge to a stationary distribution that is the joint posterior distribution $P(\boldsymbol{\beta}|N)$. The Metropolis algorithm is

MCMC simulation method that is useful for drawing samples from a posterior distribution that is not available in a standard form. The Metropolis algorithm creates a sequence of random points $\boldsymbol{\beta}^1,\boldsymbol{\beta}^2,\ldots,\boldsymbol{\beta}^i$ whose distribution converges to the target posterior distribution (Tanner 1993; Gelman et al. 1995).

In particular, the Metropolis algorithm proceeds as follows. The following steps have been programmed in *Visual Fortran6*, using the related IMSL library. Note that $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\gamma})$ in this narrative.

1. Initialization, $l=0$: Obtain an initial value $\boldsymbol{\theta}^0$, for which $P(\boldsymbol{\theta}|N)>0$, from an initial distribution $P_0(\boldsymbol{\theta})$. In our study, the data set is too large for *S-Plus* to handle. In order to obtain reasonable initial values, we computed the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ from a random 1/3 sample of the entire data set using *S-Plus* to fit a ZIP regression model. These estimates constitute initial values for the sampler.
2. Recall that the complete conditional distribution of $\boldsymbol{\theta}$ does not have a closed form. We generate a sample at the lth iteration using a Metropolis scheme, via a suitable proposal density. That is, at the lth iteration, we sample a candidate point $\boldsymbol{\theta}^{l*}$ from a proposal density $\phi_l(\boldsymbol{\theta}^l|\boldsymbol{\theta}^{l-1})$, which is symmetric. [Note that the proposal distribution is said to be symmetric if $\phi_l(\boldsymbol{\theta}^l|\boldsymbol{\theta}^{l-1})=\phi_l(\boldsymbol{\theta}^{l-1}|\boldsymbol{\theta}^l)$ for all $\boldsymbol{\theta}^l$, $\boldsymbol{\theta}^{l-1}$, and for all l. Our transition distribution is a multivariate normal distribution centered at $\boldsymbol{\theta}^{l-1}$, and is symmetric.]
3. The criterion for deciding whether or not to accept the candidate generate from the proposal, viz., $\boldsymbol{\theta}^l$, is based on the following procedure about the ratio of the posterior for the current $\boldsymbol{\theta}^l$ and $\boldsymbol{\theta}^{l-1}$.

$$\boldsymbol{\theta}^l=\begin{cases}\boldsymbol{\theta}^l & \text{with the probability } \min\left\{\dfrac{P(\boldsymbol{\theta}^l|N)}{P(\boldsymbol{\theta}^{l-1}|N)},1\right\}\\ \boldsymbol{\theta}^{l-1} & \text{otherwise}\end{cases}\qquad(11)$$

4. Go back to step 2, and repeat the same procedure with $l=l+1$. (Remark: if we accept the candidate value $\boldsymbol{\theta}^l$ and the Markov chain moves to it, the distribution ϕ is now centered at $\boldsymbol{\theta}^l$. Otherwise, the distribution ϕ is still centered at the previous iteration $\boldsymbol{\theta}^{l-1}$.)
5. The chain must be run for several iterations, and convergence monitored using Bayesian output analysis (BOA 2002). After many iterations, a series of values $\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\ldots,\boldsymbol{\theta}^l,\ldots$ converge to the posterior distribution in the end (BOA 2002).

Crash Prediction Modeling

The crash model includes year dummies to account for time effects on the intercept along with segment characteristics such as speed limit and highway cross-sectional width. We expect that our models will provide information about how the two contributing road characteristics, AADT, and segment length, affect exposure. We propose a prediction model via a log link for the mean number of crashes

$$\ln(\mu)=\text{intercept}+\beta_Y(D_Y)+\beta_V\ln(V)+\beta_L\ln(L)+\beta_W W+\beta_S S\qquad(12)$$

where D_Y=vector of dummy variables for year effects; V denotes the covariate AADT; L denotes the segment length; W denotes the pavement width; S denotes the speed limit; and the regression coefficients β are subscripted to be self-explanatory.

Analysis of Variance Test

In order to compare the state and crash type effects on the exposure factors AADT and segment length, a two way ANOVA design is applied to test the exponents of AADT and segment length. There are two null hypotheses: the first assumes that there is no regional (state) effect on the exposure functions; the second that there is no crash type effect on the exposure functions

$$\beta_{ij}=C+(\alpha_s)_i+(\alpha_c)_j+\varepsilon_{ij}\qquad(13)$$

where β_{ij}=estimated exponents for either AADT or segment length; C=overall mean; $(\alpha_s)_i$=regional effect (state) where $i=1,4$ representing four different states; $(\alpha_c)_j$=crash type effect where $j=1,4$ representing four different crash types; and ε_{ij}=random error following the normality and equal variance assumption.

Analysis and Results

Estimation Results

The Bayesian algorithm requires an initial value for each parameter before beginning the Metropolis sampling iterations. The criteria for selecting these initial values is arbitrary with respect to the algorithm, but the actual values used have a strong influence on the fast convergence of the iteration results. Therefore, we have chosen maximum likelihood estimators from the ZIP model as our initial values. After starting with these initial parameter estimates, the Metropolis algorithm provided posterior distributions for each parameter and each crash type. Table 2 shows the mean for each β parameter estimated by crash type and by state. Because the parameter vector of $\boldsymbol{\gamma}$ helps to predict the possibility for sites without crashes, which is an intermediate product and not of direct interest to us, these values are omitted here for brevity. As can be seen, the resulting parameter distributions vary greatly from one crash type to another, validating our hypothesis that a single response variable could lead to unreliable conclusions. A notable result is that the exponent on V, the traffic volume parameter, indeed varies markedly with crash type, and is lowest for single-vehicle (SV) crashes. This implies that the marginal SV crash rate is higher at low traffic volumes and lower at high traffic volumes; as volume increases, this crash type becomes less likely. This is a reasonable result; the more common presence of other vehicles on the road offers more opportunities for multivehicle crashes rather than single vehicle crashes. Also, the lower marginal single vehicle crash rate at high traffic volumes may be due to drivers being more attentive and cautious when more vehicles are around. For multivehicle crashes, the exponents on AADT are almost all greater than 1.0, with means of more than 1.40 for same direction (SD) crashes, slightly above 1.0 for opposite direction (OD) crashes, and around 1.0 for intersecting direction (ID) crashes. This complements the finding for single vehicle crashes: the marginal crash rate is lower at low traffic volumes and higher at high traffic volumes. In any case, it seems clear that this exponent on AADT differs substantially from one crash type to another, confirming our study design.

The nonlinear relationship between crash counts and AADT challenges the traditional way of computing crash rate which is the rate of crash frequency and AADT with the assumption that crash is linear to the AADT. However, our study shows that the nonlinear relationship may be due to many reasons such as the interactive effects between different crash types like single vehicle and multivehicle crashes; missing information such as

Table 2. Posterior Mean Parameters for Regression Model 3

	SV				OD			
Variable	MI	CA	WA	IL	MI	CA	WA	IL
Intercent	−2.294	−3.578	−5.124	−6.121	−4.544	−5.460	−6.554	−9.283
D_{Y1}	0.040	0.113	0.070	−0.083	−0.385	−0.079	−0.237	−0.196
D_{Y2}	0.182	0.107	0.073	−0.202	−0.215	−0.110	−0.261	−0.868
D_{Y3}	0.150	0.010	0.076	—	−0.077	−0.261	−0.117	—
D_{Y4}	0.090	0.029	—	—	−0.264	−0.122	—	—
Ln(V)	0.397	0.685	0.788	0.795	1.203	1.091	0.944	1.326
Ln(L)	0.908	0.838	0.716	0.699	0.780	0.851	0.681	0.806
W	0.007	−0.025	−0.015	0.004	−0.036	−0.028	−0.010	−0.020
S	0.002	−0.020	−0.008	−0.004	−0.023	−0.030	−0.005	0.005
	SD				ID			
	MI	CA	WA	IL	MI	CA	WA	IL
Intercent	−3.449	−4.542	−5.861	−7.596	−3.137	−3.802	−5.231	−6.485
D_{Y1}	−0.098	0.090	0.067	0.123	−0.257	−0.144	−0.860	−0.047
D_{Y2}	−0.029	0.098	0.067	0.099	−0.154	−0.093	−0.887	−0.145
D_{Y3}	−0.046	−0.048	0.077	—	−0.163	−0.033	−0.814	—
D_{Y4}	−0.153	0.059	—	—	−0.085	0.041	—	—
Ln(V)	1.422	1.263	1.000	1.740	1.123	0.915	0.877	0.948
Ln(L)	0.701	0.704	0.615	0.528	0.568	0.642	0.615	0.429
W	−0.023	−0.007	−0.016	−0.005	−0.030	−0.001	−0.010	0.020
S	−0.030	−0.028	−0.008	−0.018	−0.026	−0.021	−0.010	−0.018

Note: SV=single vehicle; OD=opposite direction; SD=same direction; ID=intersecting direction; MI=Michigan; CA=California; WA=Washington; and IL=Illinois.

intersecting traffic volume; the confounding factors like driver behaviors and so on. Some of the issues are out of control of the traffic agencies and researchers. Hence, successfully using the available data resource will require a unique approach for reconciling missing or unavailable factors. Using the nonlinear relationship between the crash count and volume provides meaningful output for the safety model without losing its accuracy.

The other component of exposure, segment length, also exhibits its variation by crash type, though to a lesser extent. The exponent on segment length is essentially close to 1.0 for SV crashes, an entirely intuitive result analogous to the vehicle-miles exposure measure commonly used now, while it is significantly lower than 1.0 for the multiple vehicle crashes. Finding different coefficients for AADT and segment length for each crash type shows that the two factors must be addressed separately when accounting for exposure to crashes.

In the two-way ANOVA test, in order to compare the state and crash type effects on the exposure factors AADT and segment length, a two way ANOVA design is applied to test the exponents of AADT and segment length. There are two null hypotheses: the first assumes that there is no regional (state) effect on the exposure functions; the second that there is no crash type effect on the exposure functions. In order to meet the ANOVA normality and equal variance assumption, we use a natural log transformation for the data. Note that the independence assumption of the ANOVA is not valid here; the results are to be interpreted in the nature of exploratory analysis. Table 3 displays the two-way ANOVA results for regional effect and crash type effect on the AADT in the prediction model.

The p value for the crash type factor rejects the null hypothesis that there is no crash type effect on the exponents on AADT at a 5% level of significance. However, we are not able to reject the hypothesis that there is no significant regional effect on the exponent on AADT at a 5% significance level. Although it does not indicate we can use pooled data from different states with confidence, it proves the consistency of the exposure function among different states and, if the data sample is small, using pooled state data remains an alternative to increase the sample size.

The estimates for the exposure coefficients show the different values due to the crash type. However, whether or not specific levels within crash type are significantly different from each other is still unclear until a multiple means comparison method is undertaken. A Tukey multiple means comparison was run in SAS (SAS 1990) to identify differences by crash type for the exponent on AADT. Keeping in mind that this procedure requires the same assumptions as the ANOVA procedure, we observe that significant differences exist between the exponents on volume for multivehicle opposite direction or same direction crashes and single vehicle crashes and there is no significant difference between the exponents on volume for single vehicle and multi-vehicle intersecting direction crashes (see Table 4). The reason the exponent for single vehicle crashes is not significantly different from that for intersecting direction multivehicle crashes may be due to unincluded effects in the crash prediction model for intersecting direction crashes, such as driveway or minor intersecting road volumes.

Table 3. Analysis of Variance Table for Effects on Exponents on V

Source	DF	Mean square	F value	p value
Crash type	3	0.39	7.95	0.0067
State effect	3	0.046	0.96	0.4545
Error	9	0.049	—	—
Total	15	—	—	—

Note: DF=degree of freedom.

Table 4. Tukey Test for Exponents on V and L for Different Crash Types

Crash types	V^a		L^a	
SV	A		A	
ID	A	B		B
OD		B	A	
SD		B		B

Note: SV=single vehicle; ID=intersection direction; OD=opposite direction; and SD=same direction.

[a] Level indicated by the same letter code is not significantly different from each other.

The resulting exponents on segment length rank upwardly from ID, SD, OD, to SV. Moreover, the test results show that the exponents on ID and SD differ significantly from those on SV and OD, demonstrating that there are important explanatory factors left out of the model that correlate with segment length. These findings are reasonable, because the segment length has less effect on the ID and SD crashes which are closely related to the number of driveways along the segment rather than the length of the segment, while OD and SV crashes are more likely to be related to the length of the segment, and have exponents close to 1.0. In other words, segment length may be confounded with unincluded factors such as driveways and minor intersections along the segments.

The parameters on the crash rate covariates speed limit and total pavement width do vary significantly by state. Some show a positive relationship with the number of crashes, while others show negative relationships even for the same crash type. The coefficients on speed limit are negative for most crash types. This is expected, and is usually because the roads with higher speed limits have safer designs, or because the speed limit has been reduced for road sections found to be unsafe. The positive coefficient on pavement width for single vehicle crashes in some states, however, is entirely unexpected because this coefficient is negative for almost all multivehicle crashes. Intuitively, one would expect a wider pavement to give drivers more maneuvering room and reduce the likelihood of running off road or colliding with another vehicle. A possible explanation is that the greater pavement width creates a road setting that encourages higher speeds than are actually appropriate for the actual geometric conditions, thus placing drivers in situations where their speed is too great to react safely to unexpected stimuli.

Conclusion

We describe here an investigation into the relationship between crashes and traffic volume (AADT) on rural two-lane highway segments in four states in the United States. Hierarchical Bayesian modeling was used in a zero-inflated Poisson framework to generate posterior distributions for parameter values. The findings show that the relationship between crashes and traffic volume is indeed nonlinear for each of the four crash types examined: single-vehicle, and multivehicle same direction, opposite direction, and intersecting direction. In particular, the relationship for single-vehicle crashes results in the marginal crash rate decreasing with traffic volume, while the opposite was observed for the three multivehicle crash types. The exponents for traffic volume and segment length are different, which casts doubt on the usage of vehicle miles traveled (VMT) with the same exponents on AADT and segment length. Furthermore, two-way ANOVA are tested for exponents on AADT and segment length with crash type and regional (states) factors. The ANOVA outputs demonstrate the necessity of disaggregating the crashes by type, but are not able to reject the hypothesis that each state has different exponents. In other words, there is no significant difference for exposure functions from state to state. The nonlinear relationship between crash frequency and segment length indicates that our models omit important explanatory factors that are likely to be correlated with the segment length. The differences in the exponent on segment length among crash types revealed by the ANOVA and Tukey's test suggest that these factors could be related to intersecting volumes, for example the number of driveways or minor intersections along the segment. The authors do not recommend using these exponents on segment length for predicting crashes on segments other than from the population sampled; as with any other statistical estimating exercise, findings must be recalibrated before being applied to other populations.

Related research by the authors has investigated a more disaggregate model accounting for temporal effects as well as crash type. Furthermore, the exposure factors includes hourly flow rate by direction rather than AADT. New modeling techniques are employed to explore the relationship between number of crashes and exposure function using hourly flow rate and segment length at a disaggregate level.

Acknowledgments

This research was sponsored by a Transportation Statistics Research Grant from the Bureau of Transportation Statistics, United States Department of Transportation and was performed at the Connecticut Transportation Institute in the University of Connecticut. The writers thank Yusuf Mohamedshah of Federal Highway Administration for providing them with the HSIS dataset and assisting us in interpreting it. They also gratefully acknowledge contribution to this project made by Dr. Don Tepas and Jianhe Du. Thanks also go to Dr. Diane Lambert of Bell Labs for providing ZIP model computer codes.

References

Bayesian Output Analysis (BOA) (2002). 〈http://www.public-health.uiowa.edu/boa/〉, accessed June 1, 2002.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*, Chapman & Hall/CRC, London.

Hauer, E. (1995). "On exposure and accident rate." *Traffic Eng. Control*, 36, 134–138.

Hauer, E. (1997). *Observational before–after studies in road safety*, Pergamon, New York.

Lambert, D. (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics*, 34(1), 1–14.

Mathsoft (1995). *S-plus, 1995 User's Manual, Version 3.3 for Windows*, Statistical Sciences Division, Seattle.

Miaou, S.-P., and Lord, D. (2003). "Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes." *Presented at Transportation Research Board Proc.*, Washington, D.C.

Miaou, S.-P., and Lum, H. (1993). "Modeling vehicle accidents and highway geometric design relationships." *Accid. Anal Prev.*, 25(5), 689–709.

SAS Institute, Inc. (1990). *SAS/STAT user's guide, Volume 2, Version 6*, 4th Ed. Cary, N.C.

Tanner, M. (1993). *Tools for statistical inference*, Springer, New York.

Tunaru, R. (1999). "Hierarchical Bayesian models for road accident data." *Traffic Eng. Control*, 40, 318–324.