ELSEVIER

Accident Analysis and Prevention 38 (2006) 1071–1080

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Bayesian estimation of hourly exposure functions by crash type and time of day

Xiao Qin [a,*], John N. Ivan [b,1], Nalini Ravishanker [c,2], Junfeng Liu [d,3], Donald Tepas [e,4]

[a] *Traffic Operations and Safety (TOPS) Laboratory, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, USA*
[b] *Civil & Environmental Engineering, University of Connecticut, Unit 2037, Storrs, CT 06269-2037, USA*
[c] *Department of Statistics, University of Connecticut, Unit 4120, Storrs, CT 06269-2037, USA*
[d] *Department of Statistics, West Virginia University, Hodges Hall 406, P.O. Box 6330, Morgantown, WV, 26506, USA*
[e] *Connecticut Transportation Institute, University of Connecticut, Storrs, CT 06269-2037, USA*

Received 19 August 2005; received in revised form 12 December 2005; accepted 15 April 2006

Abstract

The study describes an investigation of the relationship between crash occurrence and hourly volume counts for small samples of highway segments from two states: Michigan and Connecticut. We used a hierarchical Bayesian framework to fit binary regression models for predicting crash occurrence for each of four crash types: (1) single-vehicle, (2) multi-vehicle same direction, (3) multi-vehicle opposite direction, and (4) multi-vehicle intersecting direction, as a function of the hourly volume, segment length, speed limit and pavement width. The results reveal how the relationship between crashes and hourly volume varies by time of day, thus improving the accuracy of crash occurrence predictions. The results show that even accounting for time of day, the disaggregate exposure measure – hourly volume – is indeed non-linear for each of the four crash types. This implies that at any time of day, the crash occurrence is not proportional to the hourly volume. These findings help us to further understand the relationship between crash occurrence and hourly volume, segment length and other risk factors, and facilitate more meaningful comparisons of the safety record of seemingly similar highway locations.

Keywords: Exposure; Crash type; Time of day; Hourly volume; Metropolis algorithm; Binary response model

1. Introduction

Exposure in highway safety analyses, defined as some measure of the opportunity for crashes of a certain type in a given time in a given area, is applied to calculate crash rate – the ratio of the number of crashes to the exposure (Chapman, 1973). In this construct, only the occurrence of crashes is observable, as neither crash rate nor exposure is self-explanatory and each is dependent on how the other is defined. Hence, the quantitative definition of exposure determines the magnitude of the crash rate value.

The usual exposure measures applied to quantify the opportunity for crashes, such as Annual Average Daily Traffic (AADT), Vehicle-miles Traveled (VMT), or Number of Entering Vehicles (NEV), are aggregate quantities that do not consider temporal traffic variation. For example, the distribution of weekday to weekend traffic volume may vary from one location to another or from daytime to nighttime. To the extent that the actual hourly volume is an important factor in explaining the number of crashes, the hourly volume can accurately account for this effect in a way that AADT or other aggregate exposure measures cannot.

Since these opportunities for crashes are occasions when vehicles cross paths, follow one another, or even travel alone on a winding road, it follows that the occurrence of crashes involving a single vehicle may have different likelihood from those involving multiple vehicles, and further, even the likelihood of those involving multiple vehicles may vary according to the direction

* Corresponding author. Tel.: +1 608 262 3649; fax: +1 608 262 5199.
E-mail addresses: xqin@engr.wisc.edu (X. Qin), john.ivan@engr.uconn.edu (J.N. Ivan), Nalini.ravishanker@uconn.edu (N. Ravishanker), jfliu@stat.wvu.edu (J. Liu), tepas@uconnvm.uconn.edu (D. Tepas).
[1] Tel.: +1 860 486 0352; fax: +1 860 486 2298.
[2] Tel.: +1 860 486 4760; fax: +1 860 486 4120.
[3] Tel.: +1 304 293 3067x1059.
[4] Tel.: +1 860 486 5928; fax: +1 860 486 2399.

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.04.012

of travel of the vehicles involved (Chapman, 1973). In order to distinguish between the likelihood of different crash types occurring, it is customary to categorize crashes by the vehicle travel directions (Hauer et al., 1996; Brown, 1981).

Also, factors such as light condition, weather condition, driver characteristics and physical status are associated with exposure, since these factors are correlated with the temporal effects (Jovanis and Delleur, 1983). Intuitively, for the same amount of exposure, one might expect the incidence of crashes to be higher at night during the day due to differences in visibility, and due to human factors such as biological clock influences on driver alertness and sleepiness (Garbarino et al., 2000; Langlois et al., 1983). Similarly, one should expect peak hour crash occurrence to be different from off-peak crash occurrence because of different trip purposes (to/from work).

This study focuses on defining crash exposure measures at a more disaggregate level, specifically as a function of hourly directional traffic volume by time of day. For example, the exposure for a crash type involving multiple vehicles could be a function of the related traffic flows, while for single-vehicle crashes it may only be related to the volume through the roadway cross-section. The subject of this research is to formulate and estimate disaggregate crash prediction models of the actual hourly volume and segment length based on functions that are proportional to crash incidence, and whose parameters vary by crash type and by time of day.

2. Background

According to previous studies, the safety performance function (relationship between number of crashes and exposure) is nonlinear when AADT is applied as exposure for road segments; that is, crashes increase with the traffic volume in a non-linear fashion. Consequently, the crash rate (ratio of crashes to AADT) is not constant with respect to traffic volume even at the same location, and hence this rate should not be regarded as a measure of the site safety. The non-linear relationship between number of crashes and AADT may be due to some factors overshadowed by this aggregated exposure measure. For example, for the same level of AADT, one might expect more crashes to occur at night than during the day due to differences in visibility and human factors such as biological clock influences on driver alertness and sleepiness (Garbarino et al., 2000; Langlois et al., 1983). Wang and Ivan explored the interaction between exposure and time of day and argued that the effect of exposure may not be consistent throughout the entire day. They found that the effect of exposure is significantly different at 11 p.m.–6 a.m. than at other times (Wang and Ivan, 2000).

Logically, crashes at a specific time should relate closely to the hourly traffic volume or more accurately, to a real-time traffic volume. There is evidence that the hourly volume explains much of the variation in highway crash rates (Ivan et al., 2000). A number of studies have explored such microscopic models stratified by hourly volume. Gwynn studied the relationship of crash rate and crash involvement with hourly volume using a 3.8-mile section of U.S. Rte 22 in New Jersey, showing a U-shaped relationship between the number of crashes and hourly volume (Gwynn, 1967). Cedar explored in detail the relationship between road accidents and hourly traffic flow using power functions of hourly flow rate. The study found a negative relationship between single-vehicle crashes and hourly flow rate, but a U-shaped pattern for the total number of crashes as did Gwynn (Ceder and Liveh, 1982). Recently, Persaud found nonlinear relationships for both single-vehicle and multi-vehicle crashes using hourly volume under the different effect of day/night condition for two-lane rural roads. The observation indicates a convex relationship between single-vehicle accidents and traffic flow, but a concave relationship for the multi-vehicle accident (Persaud and Mucsi, 1995). Ivan also found a nonlinear relationship between single-vehicle crashes and the hourly volume to capacity ratio on two-lane rural road segments (Ivan et al., 2000). Similarly, Chang predicted the effects of traffic condition on safety at freeway sections using hourly volume and presented a U-shaped pattern between accident rates and the ratio of flow rate (volume) to capacity (V/C ratio) for all sections (Chang et al., 2000).

These studies draw both conflicting and consistent conclusions, indicating that there is indeed a relationship between the number of crashes and the hourly volume, though its exact form is still unknown. In addition, using only the total number of crashes ignores the differences in the relationship between the number of crashes and traffic volume by crash type. A recent study by Hauer categorized crashes by type and related them to the actual volumes to which the two colliding vehicles belonged, which shows that opportunities for the occurrence of single-vehicle and multiple vehicle crashes are different (Hauer et al., 1996). Similar findings can be found in Brown's study at a four-leg signalized intersection (Brown, 1981).

Consequently, in our study, we propose the following crash types on rural two-lane highways: (1) single-vehicle crashes (SV), (2) multi-vehicle same direction crashes (SD), (3) multi-vehicle opposite direction crashes (OD), and (4) multi-vehicle intersecting direction crashes (ID).

3. Data collection and description

Before using the hourly volume as one of the exposure components in order to estimate or predict crash occurrences, a few points need to be clarified, and the experiment should be carefully designed. Here, the time interval is an hour, therefore the corresponding number of crashes is unlikely to exceed one. Consequently, the dependent variable can be defined as crash occurrence denoted by a binary indicator assuming either zero or one.

The data for this study were collected from different agency resources in the states of Michigan and Connecticut. Hourly traffic volumes from Automatic Traffic Recorders (ATR) were requested from each state's Department of Transportation, with crash records and road segment characteristics gathered for contiguous highway segments to ensure hourly volume consistency. For Michigan, road segment geometric features and crash data were requested from Federal Highway Administration's (FHWA) Highway Safety Information System (HSIS) and hourly traffic volumes requested from Michigan Department of

Table 1
Variable definitions and summary statistics of road segments

Variables	Symbol	Michigan			Connecticut		
		Min	Max	Mean	Min	Max	Mean
Additive exposure	$v_1 + v_2$	2	1636	175	2	1678	285
Multiplicative exposure	$v_1 v_2$	1	628505	14036	1	696000	23837
Segment length	L	0.01	6	1.66	0.5	0.5	0.5
Pavement width	W	38	46	42	28	44	32
Speed limit	S	50	55	54	35	50	40

v_1 is the hourly volume in one direction of the two-lane rural highway. v_2 is the hourly volume in the opposite direction of the two-lane rural highway.

Transportation (MIDOT). The study period for Michigan runs from 1995 to 1997 and a total of 32 road segments were used. For Connecticut, we defined one-half mile (about 0.8 km) segments, each with homogeneous cross-sectional features, close to the ATR stations located on two-lane rural highways. The corresponding crash records were collected from Connecticut Department of Transportation (ConnDOT), and the geometric features were obtained by viewing the ConnDOT photolog archive, a videodisc system containing images of the entire 6300 centerline km (3900 miles) of the state-maintained highway network (TR News Research, 1995). Compared with the Michigan sample, Connecticut has a smaller sample size of 17 segments along with a longer time period from 1995 to 2000.

We have included the directional hourly volume, segment length, full roadway width and speed limit. The crash types defined by state police have been re-categorized into the four types defined earlier. As expected, there are several differences in the observed data between the two states, as displayed in Table 1.

Table 2 presents the number of records and the number of crashes of each type during each of three times of day periods. We selected time periods of 7 a.m.–3 p.m., 3 p.m.–11 p.m. and 11 p.m.–7 a.m. in order to be consistent with commonly defined work shifts (for comparison with the literature on circadian effects on sleepiness and work) and typical definitions of morning and afternoon peak periods. The different number of hours in each cell is a result of missing data. The two states display different patterns. Single-vehicle crashes are dominant in Michigan, while multiple vehicle same direction crashes are more common in Connecticut. The crashes for each state are distributed similarly in daytime and evening shifts. The midnight shift exhibits the lowest number of crashes, which may be due to the lowest traffic volumes at that time. Without the corresponding volume data, it is not possible to draw any valuable conclusions about crash risk. Detailed statistical analysis will be discussed in the next section.

4. Methodology

In this section, we describe the binary regression model for the crash data. An extensive discussion of these models is given in (McCullagh and Nelder, 1989). The following sections present the details of the model structure and describe the Bayesian approach for estimation and inference. The sampling-based Bayesian approach is attractive since it provides the user with samples generated from the posterior distribution of the parameters from which several features of interest, such as the estimated marginal posterior densities, posterior moments, robust posterior moments, quantiles, scatter plots exhibiting interesting relations between parameters, etc., may be derived in a straightforward manner. Further, model selection and prediction discussed based on the predictive density is attractive and we obtain the entire distribution of unknown responses, and not merely point predictions with the associated variances. This is not directly possible using the frequentist approaches, which usually provide only point estimates with the associated estimated standard errors for model parameters.

4.1. Binary regression model

The binary regression model is used for modeling these data and for making predictions. Let h denote the observation hour,

Table 2
Number of crashes by type and time of day

States	7 a.m.–3 p.m	3 p.m.–11 p.m.	11p.m.–7 a.m.	Total
Michigan				
Number of total records	219264	219311	216590	655165
SV	123	305	175	603
SD	28	35	6	69
OD	10	1	4	25
ID	19	18	2	39
Connecticut				
Number of total records	214867	214875	211773	641515
SV	17	17	8	42
SD	49	44	2	95
OD	11	8	1	20
ID	17	16	4	37

$t(h)$ denote our specification for the time of day corresponding to hour h, and let $N_{i,k,t(h)}$ be the binary indicator variable denoting the occurrence of crashes of type k at site i during a given hour h falling at time of day t. Let $p_{i,k,t(h)}$ denote the unknown probability of a crash of type k occurring at site i during a given hour h at time of day t. Note that each case represents 1 hour, but separate models are estimated for each of the three times of day for each state, with indicators for each year in the dataset. The hierarchical modeling framework has the following setup. A binary model, which is most suitable for predicting such a variable, has the form:

$$Pr(N_{i,k,t(h)}|p_{i,k,t(h)}) = p_{i,k,t(h)}^{N_{i,k,t(h)}}(1 - p_{i,k,t(h)})^{1-N_{i,k,t(h)}} \quad (1)$$

where $Pr(N_{i,k,t(h)})$ is the probability of observing $N_{i,k,t(h)}$.

Since the probability $p_{i,k,t(h)}$ always lies between zero and one, we use the generalized linear model (GLIM) with a logit link function in order to relate the probability of a crash to the observed covariates:

$$\text{logit}(p_{i,k,t(h)}) = \log\left(\frac{p_{i,k,t(h)}}{1 - p_{i,k,t(h)}}\right) = \log(\eta_{i,k,t(h)}) + \vec{X}_i\vec{\beta} \quad (2)$$

where

$$\eta_{i,k,t(h)} = \eta_{k,t(h)}(v_{i,t(h)}, L_i) \quad (3)$$

and $v_{i,t(h)}$ is the hourly volume by direction on segment i in a given hour h at time of day t, L_i is the length of the road segment i, $\vec{X}_i$ is the set of road characteristics for segment i (in this study, we use year, AADT (V), pavement width (W) and speed limit (S)), $\vec{\beta}$ is the vector of parameters to be estimated.

For simplicity of notation, we use $\eta_{ikt(h)}$ to represent $\eta_{ikt(h)}(v_{it(h)}, L_i)$ and hypothesize different functions for different crash types. In fact, the exponent on each function will also vary by time; however, this dimension has also been omitted here for brevity of notation.

We define two functions for relating traffic flow to crash incidence by type of crash. One is an additive function of volumes, and the other is a multiplicative function. Use of the additive function assumes that each entity (vehicle or driver) on the road segment has a potential opportunity to be in a crash, and the crashes on the two directions of a road are independent of each other. The multiplicative function, on the other hand, assumes that each vehicle on its own path has a potential probability to collide with a vehicle in the opposite direction, so that the number of meetings is proportional to the product of the flows, accounting for the directional split.

The additive exposure function is defined as

$$\eta_k = (v_1 + v_2)^{\alpha_{v_k}} L^{\alpha_{L_k}} \quad (4)$$

and the multiplicative exposure function is defined as

$$\eta_k = (v_1 v_2)^{\alpha_{v_k}} L^{\alpha} \quad (5)$$

where η_k is the exposure function for potential crash conflict type k, $k \in$K (SV, SD, OD, ID), v_1 is the hourly volume in one direction of the two-lane rural highway, v_2 is the hourly volume in the opposite direction of the two-lane rural highway, α_{v_k} and α_{L_k} are the exponents on flow rate and segment length, respectively, to be estimated for crash type k, $k \in$K (SV, SD, OD, ID).

A statistical model selection procedure enables the choice of the best function for a given data set and for given values of other parameter specifications.

4.2. Bayesian approach for inference

In this section, we describe a fully Bayesian framework for modeling and inference. 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 being proportional to the product of the likelihood and the prior (up to a normalizing constant). Given the entire posterior density distributions of the model parameters, we are able to do a wide range of inference beyond just the first few moments. It is superior to the other empirical methodologies such as empirical Bayesian (EB) method. 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.

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 the functions of parameters (such as differences between parameters) effortlessly, as we describe later. We fit the hierarchical fully Bayesian model using Markov chain Monte Carlo (MCMC) algorithms. The Gibbs sampling approach to estimating the model parameters involves sampling from the complete conditional distribution of each parameter in a systematic manner, conditional on the previous sampled values of the other parameters. Although the posterior density that results as the product of the likelihood function and the prior densities is analytically intractable, the Gibbs sampling approach is always possible, since the complete conditional densities are available, up to a normalizing constant, from the form of the product of the likelihood and the prior (Gelfand and Smith, 1990). When these conditional densities do not have standard form, as is often the case, the Metropolis-Hastings algorithm may be used to obtain realizations from a Markov chain having the required stationary distribution (Tanner, 1993; Gelman et al., 1995). The Metropolis-Hastings 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.

The likelihood function of the binary model parameters given the observed data is

$$\begin{aligned} L(\vec{\theta}|N) &= \prod_{h=1}^{m} p_{i,k,t(h)}^{N_{i,k,t(h)}}(1 - p_{i,k,t(h)})^{1-N_{i,k,t(h)}} \\ &= \prod_{h=1}^{m}(1 - p_{i,k,t(h)}) \prod_{N_{i,k,t=1}} \left(\frac{p_{i,k,t(h)}}{1 - p_{i,k,t(h)}}\right) \end{aligned} \quad (6)$$

where $\vec{\theta} = (\alpha_v, \alpha_L, \vec{\beta})$, and m is the total observation hours.

Replacing the forms from Eqs. (2), (4) and (5) for $p_{ikt(h)}$, we obtain these respective likelihood function forms for the additive and multiplicative functions

$$L(\vec{\theta}|N) = \prod_{h=1}^{m} \left(\frac{1}{1 + (v_{i,t(h)1} + v_{i,t(h)2})^{\alpha_v} L^{\alpha_L} e^{\vec{X}_i \vec{\beta}}} \right) \times \prod_{N_{i,k,t(h)=1}} ((v_{i,t(h)1} + v_{i,t(h)2})^{\alpha_v} L^{\alpha_L} e^{\vec{X}_i \vec{\beta}}) \quad (7)$$

or

$$L(\vec{\theta}|N) = \prod_{h=1}^{m} \left(\frac{1}{1 + (v_{i,t(h)1} v_{i,t(h)2})^{\alpha_v} L^{\alpha_L} e^{\vec{X}_i \vec{\beta}}} \right) \times \prod_{N_{i,k,t(h)=1}} ((v_{i,t(h)1} v_{i,t(h)2})^{\alpha_v} L^{\alpha_L} e^{\vec{X}_i \vec{\beta}}) \quad (8)$$

Numerical maximization for obtaining the maximum likelihood estimates (MLE) of the model parameters is possible using software such as SPLUS, our models are too cumbersome to make feasible.

The updated uncertainty about the value of these parameters is expressed via the posterior distribution as follows:

$$P(\vec{\theta}|N) \propto L(\vec{\theta}|N)\pi(\vec{\theta}) \quad (9)$$

where $P(\vec{\theta}|N)$ is the joint posterior distribution of $\vec{\theta}$ given the data, $L(\vec{\theta}|N)$ is the likelihood function, (see Eqs. (7) or (8), and $\pi(\vec{\theta})$ is the prior distribution for the vector of parameters.We specify a diffuse proper prior distribution for the parameter vector

$$\vec{\theta} \sim \text{Normal}(\vec{0}, \sigma^2 I_q) \quad (10)$$

where σ^2 is a large number, and I_q is an identity $q \times q$ matrix, q being the number of covariates.

4.3. Bayesian model selection

To study model selection we use the conditional predictive ordinate (CPO) which is defined as the estimate of $f(y_i|y_{-i})$ (for simplicity, we let y_i denote $N_{i,k,t(h)}$ with other subscripts suppressed), the cross validation density evaluated at the observation y_i (Gelfand et al., 1992). In comparing two models, the one with a larger CPO value is the one more likely to observe y_i.

$$\hat{f}(y_i|y_{-i}) = \left[\frac{1}{G} \sum_{g=1}^{G} \frac{1}{f(y_i|y_{-i}, \theta_g)} \right]^{-1} \quad (12)$$

where G denotes the number of samples obtained from the Gibbs sampler, θ is the vector of samples obtained from the Gibbs sampler. y_{-i} denotes all observed ys except y_i.

In the i.i.d case, $f(y_i|y_{-i}, \theta_g)$ is equal to $f(y_i|\theta_g)$. Hence, for the binary model where the parameter vector θ is identical independent distributed, the CPO has the form

$$\hat{f}(N_{i,k,t(h)}|\theta_g) = \left[\frac{1}{G} \sum_{g=1}^{G} \left(\frac{e^{\vec{X}_i \vec{\theta}_g}}{1 + e^{\vec{X}_i \vec{\theta}_g}} \right)^{N_{i,k,t(h)}} \left(\frac{1}{1 + e^{\vec{X}_i \vec{\theta}_g}} \right)^{1 - N_{i,k,t(h)}} \right]^{-1} \quad (13)$$

where $N_{i,k,t(h)}$ is the observed crash occurrence and it is either one or zero; $\vec{\theta}$ is the vector of parameters to be estimated for a vector of covariates $\vec{X}_i$.

Therefore, the ratio (or log ratio) of the two models indicates relative support of the observation y_i. If we aggregate over the number of observations and compute the product of all the cross validation predictive densities for all observations we get the product predictive density (PPD). Sometimes, one uses log(PPD) instead of PPD. We prefer the model with the larger PPD or log(PPD) value. Therefore, we get the pseudo-Bayes factor (PsBF) which is the ratio of the marginal likelihood under model 1 in the format of PPD and marginal likelihood under model 2 in the format of PPD. PsBF is suggested as an alternative criterion for selecting among competing models (Gelfand et al., 1992).

$$\text{PsBF} = \frac{\prod_j f(N_j|N_{-j}, \text{Model 1})}{\prod_j f(N_j|N_{-j}, \text{Model 2})} \quad (14)$$

5. Analysis and results

In the hierarchical Bayesian approach, coefficients for the covariates are considered to be random variables rather than fixed values as in classical statistical inference. Thus, the result is a sampled posterior distribution for each estimated parameter. The estimated coefficient means are shown in Tables 3 and 4, and present some interesting conclusions.

5.1. Directional factor of hourly volume

The models using the additive hourly volume and multiplicative hourly volume functions have similar estimates for all of the covariate parameters. The only obvious variation is the magnitude of the estimates for the exponents on the hourly traffic volume, which is reasonable because the scales are different for $v_1 + v_2$ and $v_1 v_2$. The purpose of testing exposure with both additive and multiplicative models is that the latter is presumed to include more information, such as the directional split, which may explain some of the crash variation. MCMC model selection using pseudo-Bayes factor is applied as the selection criteria. Table 5 displays the model selection procedure results for Michigan and Connecticut. From the results, it is difficult to judge which one is better, additive or multiplicative, because the model efficiency varies by crash type and time of day. In fact, according to Raftery's selection criteria, the difference between the two models is weak (PsBF of under 3.0) and no confident conclusions can be drawn on which model performs better (Raftery, 1995). Generally speaking, the estimated parameters are con-

Table 3
Posterior mean parameters for connecticut hourly binary model

Time	Covariate	Additive model				Multiplicative model			
		SV	SD	OD	ID	SV	SD	OD	ID
7 a.m.–3 p.m.	Intercept	**−6.422**	**−4.560**	**−5.753**	**−9.429**	**−6.655**	**−4.234**	**−6.215**	**−9.106**
	Year 1997	−0.374	−0.592	−1.249	−0.920	−0.431	**−0.572**	−1.212	−0.874
	Year 1998	**−0.823**	−0.686	**−0.556**	**−0.666**	**−0.804**	−0.701	**−0.574**	**−0.625**
	Year 1999	−0.789	**−1.057**	**−1.021**	−1.032	−0.853	**−1.039**	**−0.976**	−1.034
	Year 2000	−0.751	**−0.366**	**−1.108**	**−1.313**	−0.748	**−0.370**	**−1.057**	**−1.290**
	ln(V)*	−0.396	0.392	**−0.160**	**0.599**	−0.204	**0.203**	**−0.073**	**0.299**
	W	−0.084	0.178	0.073	**0.151**	−0.084	**0.179**	0.071	**0.152**
	S	0.064	**−0.288**	−0.110	**−0.188**	0.065	**−0.292**	−0.101	**−0.187**
3 p.m.–11 p.m.	Intercept	**−8.258**	**−12.067**	**−14.577**	**−9.747**	**−9.147**	**−11.669**	**−14.412**	**−9.356**
	Year 1997	−0.754	−0.629	**−0.663**	**−0.557**	−0.874	**−0.648**	**−0.633**	**−0.522**
	Year 1998	−0.839	**−0.078**	−1.325	−1.394	**−0.625**	**−0.083**	−1.293	−1.412
	Year 1999	−0.659	**−0.692**	**−1.245**	**−0.216**	**−1.034**	**−0.707**	**−1.305**	**−0.188**
	Year 2000	−0.550	−0.964	**−1.295**	**−1.155**	−1.290	−0.964	**−1.270**	**−1.144**
	Ln(V)*	−0.051	**0.795**	0.707	**0.525**	−0.299	**0.416**	0.351	**0.268**
	W	0.010	**0.144**	0.060	**0.185**	**0.152**	**0.142**	0.058	**0.193**
	S	−0.011	**−0.135**	**−0.011**	**−0.203**	−0.187	**−0.133**	**−0.001**	**−0.211**
11 p.m.–7 a.m.	Intercept	**−9.350**	**−9.849**	**−8.822**	**−9.997**	**-9.631**	**−9.777**	**−8.823**	**−9.314**
	Year 1997	−0.568	−0.490	−0.656	−0.582	−0.518	−0.571	−0.647	−0.500
	Year 1998	**−0.884**	−0.653	−0.924	−0.850	**−0.838**	−0.809	−0.887	−0.857
	Year 1999	**−0.875**	**−0.880**	**−0.983**	−0.910	**−0.967**	**−0.779**	**−1.018**	−0.838
	Year 2000	**−1.667**	**−1.682**	**−1.532**	**−1.655**	**−1.553**	**−1.773**	**−1.497**	**−1.622**
	ln(V)*	−0.004	0.044	−0.024	0.010	−0.011	0.009	0.012	0.034
	W	0.051	0.070	0.038	0.062	0.059	0.052	0.046	0.039
	S	−0.029	−0.037	−0.034	−0.024	−0.032	−0.020	−0.042	−0.026

* ln $(v_1 + v_2)$ for additive; ln $(v_1 v_2)$ for multiplicative model and boldface indicates for significance at 5%.

Table 4
Posterior mean parameters for michigan hourly binary model

Time	Covariate	Additive model				Multiplicative model			
		SV	SD	OD	ID	SV	SD	OD	ID
7 a.m.–3 p.m.	Intercept	**−13.132**	**−13.339**	**4.727**	**−8.052**	**−12.661**	**−12.553**	**5.631**	**−7.221**
	Year 1996	−0.137	−0.610	**−0.969**	**−1.017**	−0.131	−0.614	**−0.975**	**−1.015**
	Year 1997	−0.282	**−0.846**	**−1.236**	−0.234	−0.275	**−0.861**	**−1.264**	−0.239
	ln(V)*	0.197	**0.916**	0.480	**1.244**	0.076	**0.476**	0.242	**0.619**
	ln(L)	**0.131**	0.043	0.024	0.067	**0.134**	0.042	0.017	0.066
	W	−0.028	0.071	−0.143	−0.002	−0.029	0.072	−0.148	−0.005
	S	0.107	−0.058	−0.188	−0.141	0.106	−0.065	−0.194	−0.136
3 p.m.–11 p.m.	Intercept	**−10.190**	**−9.867**	**−26.094**	**−7.108**	**−10.139**	**−10.440**	**−27.152**	**−6.985**
	Year 1996	−0.057	**−0.723**	−0.831	−0.651	−0.056	**−0.715**	−0.827	−0.666
	Year 1997	−0.161	**−0.819**	−0.566	−0.774	−0.153	**−0.835**	−0.584	−0.765
	ln(V)*	−0.145	**1.276**	0.416	**0.888**	−0.076	**0.650**	0.228	**0.452**
	ln(L)	0.048	−0.216	0.006	0.075	0.050	−0.227	0.013	0.070
	W	−0.050	−0.043	−0.063	−0.053	−0.050	−0.045	−0.085	−0.066
	S	0.120	−0.065	0.319	−0.077	0.119	−0.052	0.353	−0.068
11 p.m.–7 a.m.	Intercept	**−20.036**	**−0.048**	**−19.928**	**−1.918**	**−12.086**	**−2.969**	**13.428**	**1.492**
	Year 1996	−0.228	**−1.523**	**−1.107**	**−1.410**	−0.224	**−1.539**	**−1.100**	**−1.388**
	Year 1997	−0.286	**−1.080**	**−1.228**	**−1.822**	−0.149	**−0.580**	**−0.642**	**−0.989**
	ln(V)*	**0.477**	0.494	0.474	0.205	**0.332**	0.339	0.307	0.143
	ln(L)	**0.166**	−0.063	−0.158	−0.034	**0.121**	−0.037	−0.119	−0.030
	W	0.009	0.013	−0.158	−0.043	0.013	−0.027	−0.258	−0.072
	S	**0.198**	−0.210	0.285	−0.115	**0.230**	−0.216	0.238	−0.115

* ln $(v_1 + v_2)$ for additive; ln $(v_1 v_2)$ for multiplicative model and boldface indicates for significance at 5%.

Table 5
Model selection between additive and multiplicative exposure

ln(PPD)	Additive exposure function			Multiplicative exposure function			Pseudo-Bayes factor (BF)		
	11 p.m.–7 a.m.	7 a.m.–3 p.m.	3 p.m.–11 p.m.	11p.m.–7 a.m.	7a.m.–3 p.m.	3 p.m.–11 p.m.			
Connecticut									
SV	−17.610	−25.344	−25.114	−16.771	−25.202	−25.416	0.432	0.868	1.352
SD	−17.656	−57.389	−52.339	−18.734	−57.599	−52.319	2.940	1.233	0.980
OD	−16.129	−20.008	−16.768	−17.350	−20.084	−16.680	3.388	1.079	0.915
ID	−17.535	−25.779	−24.657	−17.958	−26.018	−25.440	1.528	1.270	2.187
ln(PPD)	Additive exposure function			Multiplicative exposure function			Pseudo-Bayes factor (BF)		
	7 a.m.–3 p.m.	3 p.m.–11 p.m.	11 p.m.–7 a.m.	7 a.m.-3 p.m.	3 p.m.–11 p.m.	11 p.m.–7 a.m.			
Michigan									
SV	−1046.621	−2309.835	−1405.011	−1046.618	−2309.485	−1403.637	0.997	0.705	0.253
SD	−280.115	−326.171	−70.906	−279.754	−325.765	−70.726	0.697	0.666	0.836
OD	−113.568	−125.415	−52.567	−113.610	−125.444	−53.021	1.043	1.029	1.575
ID	−193.071	−188.387	−31.229	−192.936	−188.105	−31.370	0.874	0.755	1.152

sistent in significance, sign and magnitude, indicating that the choice of flow split factor has no significant effect on the other risk factors. The additive exposure model, which does not require directional volume (two-way), is simpler and more commonly accepted.

5.2. Crash type and time of day factors

One of the key issues of the study is to identify the crash prediction model variation by crash type and time of day. In this study, we focus on testing the exponents on hourly volume for different models by crash type and time of day under the Bayesian framework. For example, we have a random sample of the exponent on hourly volume, say α, from the marginal posterior distribution of α. We can state the hypothesis as wishing to see whether the posterior distribution of α in Model 1 is the same as that in Model 2. To do this, we take the pairwise differences between the MCMC samples from these two α distributions. If the 95% confidence interval of the distribution of the difference between two αs contains zero, the two α distributions cannot be assumed to be different. In this way, we can form a table to show the relationship between α values for different crash types and times of day.

Tables 6 and 7 respectively describe the comparison of exponents on hourly volume by crash type for Connecticut and Michigan for additive exposure model only. Each possible pairwise comparison of crash types is performed for each time of day. The differences found to be significantly different at the 5% level are listed in boldface. Several other comparisons, while not significantly different at this level, are substantially skewed to one direction or the other. We define these as marginally significant. We find the occurrence of crashes during the morning and afternoon shifts from 7 a.m. to 11 p.m. vary significantly by crash type while the variation is not distinctive at late night from 11 p.m. to 7 a.m. Moreover, the variation within the multi-vehicle crashes defined by vehicle traveling directions are not as significant as that of the single-vehicle crashes versus multi-vehicle crashes. It suggests that the distinction between single and multi-vehicle crashes could be statistically significant enough to disaggregate crashes.

5.3. Relationship between crash occurrence and hourly volume

In order to aid the understanding of the relationship between the crash occurrence and the hourly volume, plots of predicted number of crashes versus hourly volume are made according to different prediction models. For Connecticut, we predicted 1997 crash occurrence with the predominant geometric feature. Pavement width is 32 ft and speed limit is 40 mph, and plots are given in Figs. 1–3. The plots for Michigan data are omitted for brevity. Note that the variation among levels and curvatures for these plots also confirms the need to analyze crash prediction at this disaggregate level with the consideration of crash type and time of day. The figures indicate that the shape and scale of the safety performance function for hourly volume varies by crash type, implying that a single response model rather than a multiple response model would lead to unreliable conclusions.

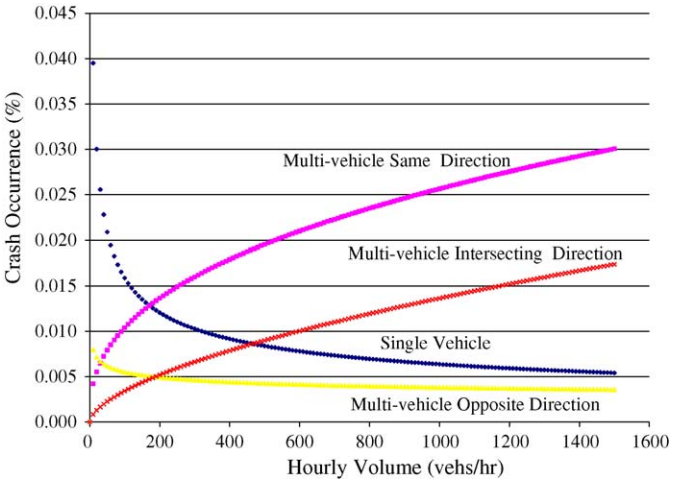

Fig. 1. Predicted number of crashes vs. flow rate (CT, 7 a.m.–3 p.m.).

Table 6
The MCMC comparison of exponent on flow rate by crash type (Connecticut)

Time Period	7am-3pm			3pm-11pm			11pm-7am		
	SV v SD	SV v OD	**SV v ID**	**SV v SD**	SV v OD	SV v ID	SV v SD	SV v OD	SV v ID
Mean	**-0.583**	-0.175	**-0.737**	**-0.626**	-0.561	-0.426	-0.048	0.021	-0.014
Std DEV	**0.206**	0.257	**0.256**	**0.217**	0.325	0.275	0.311	0.368	0.348
2.5% Quantile	**-0.966**	-0.686	**-1.219**	**-1.041**	-1.196	-0.985	-0.619	-0.603	-0.749
97.5% Quantile	**-0.171**	0.303	**-0.218**	**-0.161**	0.088	0.094	0.520	0.763	0.709
		SD v OD	SD v ID		SD v OD	SD v ID		SD v OD	SD v ID
Mean		0.408	-0.153		0.065	0.200		0.069	0.034
Std DEV		0.238	0.236		0.319	0.255		0.336	0.298
2.5% Quantile		-0.059	-0.629		-0.620	-0.327		-0.485	-0.484
97.5% Quantile		0.872	0.304		0.683	0.702		0.829	0.628
			OD v ID			OD v ID			OD v ID
Mean			**-0.562**			0.135			-0.030
Std DEV			**0.277**			0.331			0.332
2.5% Quantile			**-1.079**			-0.536			-0.621
97.5% Quantile			**-0.044**			0.791			0.582

Table 7
The MCMC comparison of exponent on flow rate by crash type (Michigan)

Time Period	7am-3pm			3pm-11pm			11pm-7am		
	SV v SD	SV v OD	**SV v ID**	**SV v SD**	SV v OD	**SV v ID**	SV v SD	SV v OD	SV v ID
Mean	**-0.439**	-0.173	**-0.638**	**-1.066**	-0.421	**-0.775**	-0.008	0.002	0.252
Std DEV	**0.183**	0.250	**0.237**	**0.200**	0.257	**0.227**	0.274	0.257	0.281
2.5% Quantile	**-0.799**	-0.681	**-1.136**	**-1.478**	-0.972	**-1.229**	-0.508	-0.510	-0.310
97.5% Quantile	**-0.094**	0.282	**-0.175**	**-0.696**	0.039	**-0.343**	0.549	0.503	0.805
		SD v OD	SD v ID		**SD v OD**	SD v ID		SD v OD	SD v ID
Mean		0.266	-0.199		**0.645**	0.291		0.005	0.245
Std DEV		0.283	0.259		**0.311**	0.261		0.375	0.403
2.5% Quantile		-0.296	-0.717		**0.053**	-0.238		-0.699	-0.542
97.5% Quantile		0.804	0.299		**1.242**	0.796		0.734	1.053
			OD v ID			OD v ID			OD v ID
Mean			-0.465			-0.354			0.321
Std DEV			0.306			0.332			0.258
2.5% Quantile			-1.034			-1.009			-0.171
97.5% Quantile			0.150			0.329			0.829

The same procedure is repeated for the differences in the exponents on hourly volume by time of day for each crash type. For both states, in at least one crash type the exponent on hourly volume varies by time of day, strongly suggesting the necessity of defining crash prediction models by time of day. The factors such as drivers' circadian rhythms, light condition, the use of alcohol or drugs, and trip purpose are closely related to time of day. Therefore, it is a reasonable alternative variable to cover their effects on the exposure parameters.

In this study, we are most concerned about the relationship between crash occurrence and exposure components of hourly volume. The exponents on hourly volume during a majority of the time periods exhibit a positive relationship for multi-vehicle crash occurrence and a negative one for single-vehicle crash occurrence (Tables 3 and 4). Moreover, the linear relationship between the occurrence of crashes and hourly volume is tested using a similar method, i.e., if the 95% credible interval derived from the posterior distribution of the exponent on hourly volume excludes 1.0, this suggests we reject the null hypothesis that the parameter is equal to 1.0. Our study indicates that even under models disaggregated by crash type, time of the day with actual hourly volume, the relationship between crash occurrence and traffic volume or segment length is not linear. Therefore, when evaluating the crash rate as a function of traffic volume, one must expect a non-linear rather than linear relationship.

Previous studies have shown inconsistent findings about the relationship between number of crashes and the hourly volume. These findings include concave, convex, U-shape or other relationships. In fact, these discoveries rely on to what level the study disaggregated the data. Our study shows mixed functional

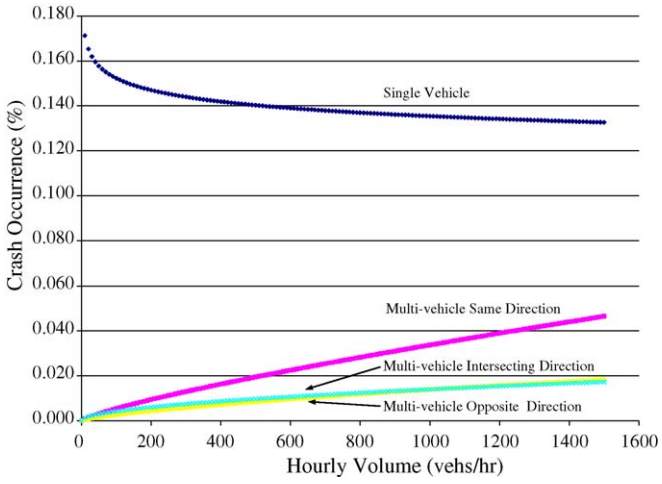

Fig. 2. Predicted number of crashes vs. flow rate (CT, 3 p.m.–11 p.m.).

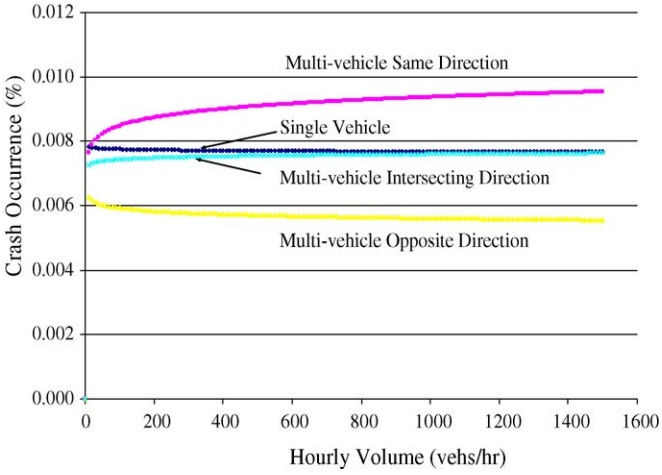

Fig. 3. Predicted number of crashes vs. flow rate (CT, 11 p.m.–7 a.m.).

tendencies: there is a negative relationship (convex downward) between single-vehicle crash occurrence and hourly volume at some times of day, but a concave upward relationship at other times. Multi-vehicle crash occurrence shows either a concave or convex upward relationship with the hourly volume. Consequently, the relationship between total vehicle crash occurrence and hourly volume may display a U-shape if the relationship for single-vehicle crashes is convex downward while that for multi-vehicle crashes is convex upward.

6. Conclusion

This paper describes an investigation into the relationship between crash occurrence and hourly traffic volume on rural two-lane highway segments. We used a hierarchical Bayesian framework with Markov Chain Monte Carlo (MCMC) algorithms to estimate the posterior distributions for crash probabilities as a function of hourly volume, time of day and so on. The findings show that even accounting for time of day, the relationship between crash occurrence and traffic volume is indeed non-linear for each of the four crash types: single-vehicle, and multi-vehicle same direction, opposite direction and intersecting direction. Consequently, the crash exposure proposed in this study is a function of the hourly volume and segment length with significant exponents different from one in each case.

In particular, the findings in this study demonstrate that crash risk prediction functions vary by crash type and time of day. In other words, the expected total crash count on two equal length segments with the same AADT and physical characteristics will vary according to the distribution of traffic volume through the day. Also, it provides a new modeling technique using hierarchical Bayesian binary response model which makes the models more flexible and extracts more information from the data.

Besides the findings mentioned above, we note some of the exposure factor such as hourly volume and risk factors such as roadway width, speed limit are inconsistent from Connecticut to Michigan, indicating the necessity for calibrating the model for transferability. Because only a few study segments in two states are employed in this study, the small sample size limited the prediction accuracy and significance of the covariates. Also, other highway characteristics such as population or driveway density may be more relevant than the covariates used in this study. The findings presented here indicate that time of day clearly affects the parameter estimates in crash prediction models. These findings should be validated and clarified through estimation with a larger data set that will result in more significant parameters estimates and more accurate crash predictions. Note that it is helpful to have good estimates of hourly volumes by time of day or better, actual traffic counts in more locations. Consequently, related research into estimating or extrapolating accurate hourly or time of day traffic volumes would be indispensable for increasing opportunities for building a larger data set. Finally, the study can be expanded to more facility types such as intersections and multi-lane highways.

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 authors thank Michigan Department of Transportation and Connecticut Department of Transportation for providing us with the data and assisting us in interpreting it.

References

Brown, R.J., 1981. A method for determining the accident potential of an intersection. Traffic Eng. Control 22 (12), 648–651.

Ceder, Avishai, Liveh, Moshe, 1982. Relationships between road accidents and hourly traffic flow-i analyses and interpretation. Accident Anal. Prev. 14 (1), 19–34.

Chang, Jaenam, Oh, Cheol, Chang, Myungsoon, 2000. Effects of Traffic Condition (v/c) on safety at freeway facility sections. Transportation Research, E-Circular, Fourth International Symposium on Highway Capacity Proceedings, pp. 200–208.

Chapman, R.A., 1973. The concept of exposure. Accident Anal Prev. 5 (2), 95–110.

Garbarino, S., Nobili, L., Beelke, M. Carli, F.D., Ferrillo, F. 2000. The Contributing role of sleepiness in highway vehicle accidents. Daytime Sleepiness, Accepted for publication, p. 203–206.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities. J. Am. Stat. Assoc. 85, 398–409.

Gelfand, A.E., Dey, D., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling based methods (with discussion). In: Bernardo, J., et al. (Eds.), Bayesian Statistics 4. Oxford University Press, pp. 147–167.

Gelman, Andrew, Carlin, John B., Stern, Hal S., Rubin, Donald B., 1995. Bayesian Data Analysis. Chapman & Hall/CRC.

Gwynn, D.W., 1967. Relationship of accident rates and accidents involvement with hourly volume. Traffic Quart. 21 (3), 407–418.

Hauer, E., Ng, J.C.N., Lovell, J., 1996. Estimation of safety at signalized intersection. Transport. Res. Rec. 1285, 42–51.

Ivan, J.N., Wang, C.-Y., Bernardo, N.R., 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accident Anal. Prev. 32, 787–795.

Jovanis, P.P., Delleur, James, 1983. Exposure-based analysis of motor vehicle accidents. Transport. Res. Rec. 919, 1–7.

Langlois, P., Smolensky, M., His, B., Weir, F., 1980-1983. Temporal patterns of reported single-vehicle car and truck accidents in Texas, USA, during 1980–1983. Chronobiol. Int. 2, 131–146.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, London.

Persaud, Bhagwant N., Mucsi, Kornel, 1995. Microscopic accident potential models for two-lane rural roads. Transport. Res. Rec. 1485, 134–139.

Raftery, A. E., 1995. Bayesian model selection in social research (with Discussion by Andrew Gelman, Rubin, Donald B. Hauser, Robert M., and a Rejoinder). Sociological Methodology.

Tanner, Martin##A., 1993. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Springer-Verlag.

TR News Research, 1995. Pays Off: Laser Videodisc Technology Meets Changing Operational Demands. vol.176. pp. 24–25.

Wang, C.-Y., Ivan, J.N. 2000. Representing traffic exposure in multi-vehicle crash prediction for two-lane highway segments. Transportation Research Board Annual Meeting, Washington DC.