ELSEVIER

Accident Analysis and Prevention 39 (2007) 125–134

ACCIDENT
ANALYSIS
&
PREVENTION

www.elsevier.com/locate/aap

Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models

Do-Gyeong Kim[a,*], Yuhwa Lee[b,1], Simon Washington[b,2], Keechoo Choi[c,3]

[a] *Department of Urban Transportation, Seoul Development Institute, Seoul, 137-071, Republic of Korea*
[b] *Department of Civil and Environmental Engineering, Arizona State University, Tempe, AZ 85287-5306, United States*
[c] *Department of Transportation Engineering, Ajou University, San 5 Woncheon-dong, Paldal-ku, Suwon 442-749, Republic of Korea*

Received 4 January 2006; received in revised form 16 May 2006; accepted 21 June 2006

Abstract

It is important to examine the nature of the relationships between roadway, environmental, and traffic factors and motor vehicle crashes, with the aim to improve the collective understanding of causal mechanisms involved in crashes and to better predict their occurrence. Statistical models of motor vehicle crashes are one path of inquiry often used to gain these initial insights. Recent efforts have focused on the estimation of negative binomial and Poisson regression models (and related deviants) due to their relatively good fit to crash data. Of course analysts constantly seek methods that offer greater consistency with the data generating mechanism (motor vehicle crashes in this case), provide better statistical fit, and provide insight into data structure that was previously unavailable.

One such opportunity exists with some types of crash data, in particular crash-level data that are collected across roadway segments, intersections, etc. It is argued in this paper that some crash data possess hierarchical structure that has not routinely been exploited. This paper describes the application of binomial multilevel models of crash types using 548 motor vehicle crashes collected from 91 two-lane rural intersections in the state of Georgia. Crash prediction models are estimated for angle, rear-end, and sideswipe (both same direction and opposite direction) crashes. The contributions of the paper are the realization of hierarchical data structure and the application of a theoretically appealing and suitable analysis approach for multilevel data, yielding insights into intersection-related crashes by crash type.

Keywords: Hierarchical data; Multilevel models; Motor vehicle crashes; Transportation safety; Rural intersections

1. Introduction

Motor vehicle crashes are thought, in general, to be caused by a combination of factors such as driver characteristics (attention, mood, eyesight, reaction times, driving skills, etc.), roadway characteristics (sight distance, pavement surface, roadway alignment, signing and striping, traffic control, roadside environment, etc.), and environmental factors (weather conditions, visibility, wind, etc.). Statistical models are helpful for identifying factors associated with motor vehicle crashes, despite their vast limitations. To this end, many studies (Harwood et al., 2000; Vogt, 1999; Vogt and Bared, 1998; Oh et al., 2003; Chin and Quddus, 2003) have developed crash models focused on predicting total crashes (totals, fatals, injuries, etc.) for assessing the safety effects of various factors. While these models have great utility, they are limited because they fail to relate crash types with roadway, traffic, and environmental factors, as discussed in Kim et al. (2006). For example, these models cannot provide insight into what factors are associated with rear-end crashes, which may be ‘over-represented’ at certain intersections. Nor could one compare the effect of the presence of left-turn lanes at signalized intersections on angle and rear-end crashes. These types of insights are valuable because different crash types require different safety countermeasures or interventions.

Recent research has shown that crash types are associated with geometric characteristics and traffic conditions in varying ways, and thus the safety effects of crash-related variables on different crash types cannot be revealed through crash totals (Kim et al., 2006). Stated differently, models of crash totals provide weighted average effects of various factors on crashes. Shankar

[*] Corresponding author. Tel.: +82 2 2149 1105; fax: +82 2 21491120.
E-mail addresses: dokkang@sdi.re.kr (D. Kim), yu.h.lee@asu.edu (Y. Lee), Simon.Washington@asu.edu (S. Washington), keechoo@ajou.ac.kr (K. Choi).
[1] Tel.: +1 480 727 9805; fax: +1 480 965 0557.
[2] Tel: +1 480 965 3589; fax: +1 480 965 0557.
[3] Tel.: +82 31 219 2538; fax: +82 31 215 7604.

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.06.011

et al. (1995) estimated overall accident models to evaluate the effects of roadway geometric variables and environmental factors on crash frequencies. In addition to modeling overall crash frequency, they separately modeled the frequency of specific types of crashes, focusing on the identification of safety effects of environmental factors on different crash types. They concluded that separate regression models for specific types of accidents have the potential for providing greater explanatory power relative to single overall crash frequency models. Because of the suspected heterogeneity in underlying causal mechanisms associated with different crash types, it is reasonable to suspect that the probabilities of crash occurrence by crash type are associated with roadway, traffic, and environmental factors in different ways.

This study builds on previous work by Kim et al. (2006) and aims to identify factors that affect the probability that certain types of crashes will occur by exploiting the hierarchical structure of intersection crashes. The data consist of 548 motor vehicle crashes that occurred on 91 two-lane rural intersections in the state of Georgia. Hierarchical structure is essentially a statistical description of a data structure that is characterized by correlated responses within hierarchical clusters. Hierarchical models, in fact, are justified by the presence of correlation within clusters; otherwise non-hierarchical modeling methods are appropriate. The hierarchy in these intersection crash data is postulated as follows. The crashes themselves represent the lowest level of the hierarchy, while the intersection at which the crash occurred represents the higher-level hierarchy, or cluster. It is reasonable to claim that correlation exists among crashes occurring at the same intersection, since these crashes may share unobserved and/or unrecorded characteristics of the intersection. From a methodological viewpoint, it is somewhat problematic to estimate traditional logistic regression models because an assumption of logistic regression (and other models such as negative binomial, Poisson, linear regression, etc.) is that the residuals from the model are independent across subjects (Jones and Jørgensen, 2003). For instance, suppose unobserved factors affect the probability of crash types at intersections. These unobserved factors might include poor pavement condition, low pavement friction, or poor reflectivity of road signs or lane striping. Unobserved characteristics might also include excessive distractions at the site, nearby drinking establishments, heavy animal populations, etc. Because of these, unobserved factors, crash frequencies and types observed at a particular location may be correlated. The correlation within clusters (higher-level units) violates the assumption of residual independence assumed in many statistical methods. If significant correlation within clusters is left unchecked – i.e. modeled without considering hierarchy – the consequence is generally attenuation of effects (parameter estimates tend toward zero), biased parameter estimates, biased standard errors, or heterogeneity of the regression (Bryk and Raudenbush, 1992). These problems are overcome (when cluster level correlation exists) by applying multilevel modeling techniques.

This paper describes the development of multilevel binomial logistic models for predicting the probability of certain types of crashes. It postulates that intersection crashes are hierarchical in nature, with crash-level and intersection-level hierarchies. The crash types examined include angle, head-on, rear-end, and sideswipe (both same and opposite direction) crashes.

2. Data description

The structure of data from 38 counties in the state of Georgia for 2 years (1996–1997) is postulated as hierarchical as depicted in Fig. 1. The figure reveals that the data consist of two different levels: level 1 consisting of crash-level characteristics and level 2 consisting of intersection-level characteristics. Crash-level characteristics contain detailed information associated with individual crashes such as day of week, weather condition, surface condition, and so on, whereas intersection-level characteristics include geometric information across intersections.

Five crash-level and four intersection-level variables are used in the analysis. Descriptions of the variables used in the study are provided in Table 1. Note that a horizontal curve indicator variable and a vertical curve (grade) indicator variable are included in crash-level characteristics (level 1 covariates), even though those variables are related to roadway geometry. The reason for this inclusion is that the intersection crash definition consisted of all crashes occurring within 76 m (250 ft) from the intersection center point along the major and minor roads; therefore, the degree of curve and grade of locations at which crashes occurred could vary within an intersection, making them independent of the intersection.

Table 2 provides summary statistics of the intersection crash data. A total of 548 reported crashes represent the level 1 units: 274 (50%) occurred at unsignalized intersections and 274 (50%) occurred at signalized intersections. As shown in Table 2, 60.8% of all reported crashes occurred during clear weather conditions, and 21.1% of crashes occurred during wet road-surface con-

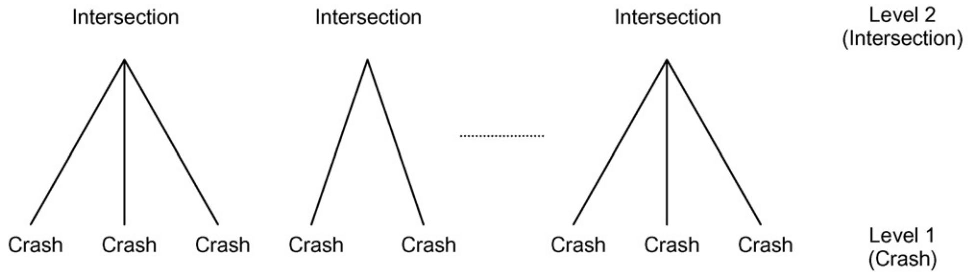

Fig. 1. The hierarchical structure of data.

Table 1
Descriptions of variables used in the analysis

Variable	Description
Dependent variables	
ANGLE	1 if angle crash, 0 otherwise
HEADON	1 if head-on crash, 0 otherwise
REAREND	1 if rear-end crash, 0 otherwise
SIDESAME	1 if sideswipe (same direction) crash, 0 otherwise
SIDEOPPO	1 if sideswipe (opposite direction) crash, 0 otherwise
Crash-level characteristics	
CLEAR	Clear weather indicator: 1 if crash occurred during clear weather condition, 0 otherwise
SURFACE	Surface condition indicator: 1 if crash occurred on a wet road-surface, 0 otherwise
DAYLIGHT	Daylight indicator: 1 if crash occurred during daylight, 0 otherwise
CURVE	Curve indicator: 1 if crash occurred on a horizontal curve, 0 otherwise
GRADE	Grade indicator: 1 if crash occurred on a vertical curve, 0 otherwise
Intersection-level characteristics	
SHOULDER	Shoulder indicator: 1 if shoulder exists on either major or minor roads, 0 otherwise
SIGNAL	Signal indicator: 1 if signalized intersection, 0 unsignalized intersection
DRIVEWAY	Driveway indicator: 1 if the number of driveways within 250 ft from the intersection center (both major and minor roads) is greater than 4 (median value), 0 otherwise
HAU	Intersection angle indicator: 1 if the degree of intersection angle is 90, 0 otherwise

Table 3
The number of intersections according to the number of crashes

Number of crashes	Number of intersections		
	Non-signal	Signal	Total
2	20 (21.98)	1 (1.10)	21 (23.08)
3	12 (13.19)	2 (2.19)	14 (15.38)
4	11 (12.09)	2 (2.20)	13 (14.29)
5	7 (7.69)	4 (4.40)	11 (12.09)
6	4 (4.40)	0 (0.00)	4 (4.40)
7	0 (0.00)	1 (1.10)	1 (1.10)
8	2 (2.20)	0 (0.00)	2 (2.20)
9	4 (4.40)	1 (1.10)	5 (5.50)
10	2 (2.20)	5 (5.49)	7 (7.69)
11	1 (1.10)	2 (2.20)	3 (3.30)
12	1 (1.10)	2 (2.20)	3 (3.30)
13	0 (0.00)	1 (1.10)	1 (1.10)
14	0 (0.00)	1 (1.10)	1 (1.10)
17	0 (0.00)	2 (2.20)	2 (2.20)
20	0 (0.00)	1 (1.10)	1 (1.10)
22	0 (0.00)	1 (1.10)	1 (1.10)
23	0 (0.00)	1 (1.10)	1 (1.10)
Total	64	27	91

Note. The percentages are in parentheses.

ditions. With respect to collision type, 43.6% of all reported crashes are angle crashes (including turning crashes), whereas 2.9% are head-on crashes representing the smallest crash type category.

A total of 91 rural intersections represent the level 2 units: 64 unsignalized intersections and 27 signalized intersections of two-lane roads. As indicated in Table 3, 23.08% of all intersections observed two reported crashes during the period, and 64.8% of intersections observed less than five reported crashes.

Table 2
Summary statistics of intersection crash data

	Unsignalized intersections		Signalized intersections		Total	
	No.	%	No.	%	No.	%
Number of crashes	274	50.0	274	50.0	548	100.0
Number of intersections	64	67.7	27	32.3	91	100.0
Dependent variables						
ANGLE	132	24.1	107	19.5	239	43.6
HEADON	8	1.5	8	1.5	16	2.9
REAREND	40	7.3	104	19.0	144	26.3
SIDESAME	14	2.5	19	3.5	33	6.0
SIDEOPPO	12	2.2	8	1.4	20	3.6
Crash-level characteristics						
CLEAR	156	28.5	177	32.3	333	60.8
SURFACE	67	12.2	49	8.9	116	21.1
DAYLIGHT	195	35.6	229	41.8	424	77.4
CURVE	241	44.0	263	48.0	504	92.0
GRADE	131	23.9	77	14.1	208	38.0
Intersection-level characteristics						
SHOULDER	45	49.4	22	24.2	67	73.6
SIGNAL	0	0.0	27	29.7	27	29.7
DRIVEWAY	12	13.2	14	15.4	26	28.6
HAU	2	2.2	10	11.0	12	13.2

Note. All of the variables are binary.

While all unsignalized intersections observed fewer than 13 crashes, the number of reported crashes for signalized intersections ranges from 2 to 23 crashes. The higher frequency of crashes at signalized intersections may be due to differences in exposure and is not unusual.

3. Model structure

Multilevel modeling techniques are appropriate when there is correlation among clusters of subjects, as described previously. For example, data obtained from surveys of individuals within households across different countries might constitute a three-level hierarchy—individuals (level 1), households (level 2), and countries (level 3). It is the presence of within-cluster correlation that justifies the use of a hierarchical (multilevel) model; without within-cluster correlation multilevel modeling does not provide benefit. Multilevel modeling techniques are commonly used in social contexts and individual behavior (Bryk and Raudenbush, 1992), but have rarely been applied to the area of road safety modeling. One reason for this might be because the possible existence of hierarchical structures in crash data is commonly ignored (Jones and Jørgensen, 2003). Alternatively, analysts may not be aware of the analysis technique or the motivation for adopting the technique. The authors identified one motor vehicle crash-related study by Jones and Jørgensen (2003), where multilevel models were fitted to a three-level hierarchy and used to identify factors affecting fatality risk for individual casualties.

In the literature, hierarchical models are given a variety of titles, including multilevel linear models (Goldstein, 1987; Mason et al., 1983), mixed-effects models and random-effects models (Elston and Grizzle, 1962; Laird and Ware, 1982), random-coefficient regression models (Rosenberg, 1973), covariance components models (Dempster et al., 1981; Longford, 1987), and hierarchical linear models (Bryk and Raudenbush, 1992). The term "multilevel models" is used throughout the remainder of this paper.

3.1. Standard multilevel models

Suppose we have collected data on N subjects (level 1) nested within J organizations or clusters (level 2). With two-level structure data, three different equations can be formulated: individual-level model (level 1 model), organization-level model (level 2 model), and combined model. Assuming normally distributed errors, for subject ij we have a level 1 model as

$$Y_{ij} \sim N(\hat{Y}_{ij}, \sigma_{ij}^2); \qquad r_{ij} \sim N(0, \sigma^2); \qquad \hat{Y}_{ij} = \hat{\beta}_{0j} + \hat{\beta}_{1j} X_{ij};$$

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \quad \text{(level 1 model)} \tag{1}$$

where β_{0j} is the intercept, β_{1j} the regression coefficient associated with the predictor X_{ij}, and r_{ij} is the residual accounting for level 1 random effect.

Although this formulation is similar to a linear regression model, there is an important difference in that both intercept and regression coefficients have subscript j, indicating that the intercept β_{0j} and the slope coefficient β_{1j} are permitted to vary across organizations (level 2). At the organization level, the units are organizations and the regression coefficients in the level 1 model for each organization are conceived as outcome variables depending on organization-level characteristics. Generally, there are five submodels in multilevel models depending on whether or not the intercept β_{0j} and the slope coefficient β_{1j} are assumed to vary across organizations. For more detailed information on the other submodels of multilevel models, see Bryk and Raudenbush (1992). In this application, the intercept β_{0j} is assumed to vary across organizations as a function of a grand mean, a single explanatory variable, and an error term, but the slope coefficient β_{1j} is assumed not to vary across organizations. Then, the intercept β_{0j} and the slope coefficient β_{1j} are formulated as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad \text{(level 2 model)} \tag{2}$$

and

$$\beta_{1j} = \gamma_{10} \quad \text{(level 2 model)} \tag{3}$$

In Eqs. (2) and (3), note that the gammas (regression coefficients) do not have subscript j because they are not assumed to vary across organizations. This model corresponds with a random-intercept model (Bryk and Raudenbush, 1992).

Substituting Eqs. (2) and (3) into Eq. (1) yields the combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij} + u_{0j} + r_{ij} \quad \text{(combined model)} \tag{4}$$

where Y_{ij} is the outcome variable for the ith subject at level 1 and the jth unit at level 2, γ_{00} the intercept, W_j the organization-level characteristic, X_{ij} the individual-level characteristic, γ_{01} and γ_{10} the regression coefficients associated with organization-level characteristic and individual-level characteristic, respectively, u_{0j} a random effect accounting for the random variation at level 2, where $u_{0j} \sim (0, \sigma_u^2)$ and r_{ij} is the individual-level random effect, where $r_{ij} \sim N(0, \sigma^2)$.

3.2. Multilevel binomial logit models

For a standard multilevel model described above, an outcome variable is continuously distributed. In case that the observed outcomes Y_{ij} are binary, however, a binomial logit model is appropriate. With hierarchical structure data, this translates to a multilevel binomial logistic model. A multilevel binomial logistic model is conceptually equivalent to a standard multilevel model except for the outcome variable. The multilevel binomial logistic model is well described for example in Guo and Zhao (2000).

Considering a binomial $Y_{ij} = (0, 1)$ outcome, and $p_{ij} = (\sum Y_{ij}/n)$, a combined model, Eq. (4), is rewritten as

$$Y_{ij} \sim \text{Bin}(\hat{\theta}_{ij}, n_{ij}),$$

$$\text{logit}(\theta) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij} + u_{0j} \tag{5}$$

where p_{ij} is the probability of the response equal to one, and γ_{00}, γ_{01}, γ_{10}, W_j, X_{ij}, and u_{0j} are as defined previously. As with the standard multilevel model, Eq. (5) is called a combined model and obtained with the assumption that the intercept β_{0j} varies across organizations as a function of a grand mean, a single explanatory variable, and an error term, but the slope coefficient β_{1j} does not vary across organizations. Therefore, this model is a random-intercept model.

Based on the model structure above, a multilevel binomial logistic model for outcome probabilities of intersection crash data used in this study is formulated as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{00} + \sum_{q=1}^{Q}\gamma_{0q}W_{qj} + \sum_{p=1}^{P}\gamma_{p0}X_{pij} + u_{0j} \qquad (6)$$

where p_{ij} is the probability that a type of crash will occur ($Y_{ij} = 1$), γ_{00} the intercept, W_{qj} a vector of intersection-level characteristics, X_{pij} a vector of crash-level characteristics, γ_{0q} and γ_{p0} the regression coefficients associated with the intersection-level characteristics and the crash-level characteristics, respectively, and u_{0j} is the random effect at level 2, where $u_{0j} \sim N(0, \sigma_u^2)$. Since Eq. (6) is a combined model, which corresponds to a random-intercept model, level 1 and level 2 models are given by

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \sum_{p=1}^{P}\beta_{pj}X_{pij} \quad \text{(level 1 model)} \qquad (7)$$

and

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \sum_{q=1}^{Q}\gamma_{0q}W_{qj} + u_{0j}\\ \beta_{1j} &= \gamma_{10} \qquad \text{(level 2 model)} \qquad (8)\\ &\vdots\\ \beta_{pj} &= \gamma_{p0}\end{aligned}$$

4. Model estimation issues and procedure

At first, fully unconditional models (models that are not conditional on model parameter vector, θ) are estimated for five types of crashes. Then, the proportion of the variance in the outcome (crash type) between the level 2 units (intersections) is examined by the intra-class correlation coefficient (ICC). In general, the variance of the outcome in standard multilevel models consists of two components: the variance of u_{0j} (τ_{00}) and the variance of r_{ij} (σ^2). The σ^2 parameter captures the within-group variability and τ_{00} captures the between-group variability. With these two variances, the intra-class correlation coefficient for standard multilevel models is calculated using the following equation to measure the proportion of the variance in the outcome that is between the level 2 units.

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \qquad (9)$$

If ρ is sufficiently close to zero, then there is effectively no variation in the subjects between the level 2 units, suggesting that standard subject level models are adequate for these data. For multilevel binomial logistic models, however, the variance of $r_{ij}(\sigma^2)$ is not available because the subject-level error term (r_{ij}) is not included in Eq. (6) unlike standard multilevel models. Instead, the variance of a standard logistic distribution is used, $\sigma_e^2 = \pi^2/3$. Thus, the variance at the lowest level is completely determined by the population proportion (Bryk and Raudenbush, 1992).

The next step is to estimate random-intercept models as described in Eq. (6). The estimation of multilevel binomial logistic models is performed using a GLIMMIX macro in SAS software. The GLIMMIX macro employs a pseudo-likelihood (PL) estimator proposed by Wolfinger and O’Connell (1993).

5. Estimation results

As has been postulated, the presence of within-organization (or cluster) correlation caused by unobserved characteristics at intersections, multilevel modeling techniques are used to analyze crash types in Georgia. Of prime interest is how various observed characteristics of intersections and environmental factors influence the probabilities of various crash types. Using five crash type variables, five separate multilevel models are estimated: angle, head-on, rear-end, same direction sideswipe, and opposite direction sideswipe crashes.

For obtaining estimates of between- and within-organization (or cluster) variance, unconditional models were estimated (see Table 4). The intra-class correlation coefficient (ICC) is 0.09 for angle crashes, indicating that 9% of the total variation in angle crashes exists between intersections, and therefore may be explained using intersection-level predictors. In contrast to angle crashes, the ICC for head-on crashes is 0.0, indicating that

Table 4
The estimation results of unconditional models

	AC[a]	HD[a]	RE[a]	SSSD[a]	SSOD[a]
Fixed effect					
Intercept	−0.2676 (0.109)	−3.5041 (0.254)	−1.2912 (0.139)	−3.2312 (0.237)	−4.0698 (0.287)
Random effect					
Intercept	0.3109 (0.165)	0.00 (0.00)	0.6918 (0.259)	2.2881 (0.616)	3.2934 (0.805)
ICC	0.09	0.00	0.17	0.41	0.50
$-2\log L$	2344.8	3508.8	2473.3	3038.9	3276.4

Note. For parameter estimates, standard errors appear in parentheses.

[a] AC: angle crash model; HD: head-on crash model; RE: rear-end crash model; SSSD: sideswipe same direction crash model; SSOD: sideswipe opposite direction crash model.

evidence is lacking that the variation in head-on crashes exists between intersections. As a result multilevel models for head-on crashes are not necessary and head-on crashes can be modeled using traditional binomial logistic regression methods.

The ICCs for rear-end and sideswipe crashes (both same and opposite direction) are 0.17, 0.41, and 0.50, respectively, indicating that between-intersection variation is present for these crash types. As a result intersection-level predictors are useful for estimating statistical models for these types of crashes. It should be noted that roughly 50% of the total variation in sideswipe crashes is attributable to between-intersection variability, suggesting that sideswipe crashes are significantly influenced by intersection characteristics, in contrast to head-on crashes which appear to be unaffected by them (and instead are influenced by crash-level characteristics which does include horizontal and vertical curvature effects).

Based on the unconditional model estimation results, one binomial logistic regression model and four multilevel binomial logistic regression models are separately estimated for all of the crash types.

5.1. *Angle crash, head-on crash, and rear-end crash models*

In the head-on crash model no statistically significant variables were ascertained. It is suspected that the small number of head-on crashes in the population of crashes (total of 16 head-on crashes) is too small to yield statistically meaningful results. Moreover, head-on crashes are thought to be rare and random events at intersections that are not strongly influenced by intersection features. This combination of factors is thought to have led to the inconclusive head-on crash results.

Table 5 presents the estimation results of angle and rear-end crash models, in which crash and intersection features are included as predictors. For logistic regression models, the odds ratio is used to interpret the actual effects of estimated coefficients. The odds ratios of estimated coefficients indicates how the odds of an event is affected by the presence of an intersection characteristics (indicator variable for characteristic assigned the value 1). For example, the estimated coefficient β_1 associated with an independent variable X represents the change in the log odds from $X = 0$ to 1. Therefore, the values of coefficients must be transformed back to their original scale (odds ratio) in order to interpret the actual effects of variables. The odds ratio is obtained by simply exponentiating the value of the coefficient associated with the variable. Then, subtracting 1 from the odds ratio and multiplying by 100 gives the percent change in the probability of the event's occurring with a unit change in the independent variable. Odds ratios are also provided in Table 5.

The results show that (all else being equal) angle crashes are more likely (odds = 2.35–1) to occur on horizontal curves compared to straight sections, are more likely to occur during clear weather conditions (odds = 1.46–1) relative to other weather conditions such as rainy and snowy days, and during the daytime (odds = 1.66–1) compared to the night-time. These effects are likely to be capturing exposure, whereas greater numbers of vehicles pass through the intersections during daytime and clear weather. However, when traffic volume was entered into the model it was not statistically significant (contrary to our expectations and repeated attempts to employ it in the model), therefore these variables appear to be superior predictors of exposure for these crash types. One must keep in mind that the predicted outcome is crash type and not crash frequency, and so traffic volume has a much less direct predictive effect.

It should be noted that comparison of odds ratios for these variables across crash types is more meaningful than assessing the odds ratio independently (since exposure is constant across crash types). Interestingly, angle crashes are more likely during clear weather conditions (compared to foul weather) whereas rear-end crashes are less likely during clear weather.

Table 5
Estimation results for angle and rear-end crashes

Variable	Angle crashes		Rear-end crashes	
	Estimate	Odds ratio	Estimate	Odds ratio
Fixed effects				
Intercept (γ_{00})	-1.021^{**}	0.36	-3.002^{***} (0.614)	0.05
Crash-level				
CLEAR (γ_{10})	0.381^{*} (0.229)	1.46	−0.365 (0.249)	0.69
SURFACE (γ_{20})	-0.649^{**} (0.289)	0.52	−0.188 (0.308)	0.83
DAYLIGHT (γ_{30})	0.506^{**} (0.231)	1.66	1.140^{***} (0.299)	3.13
CURVE (γ_{40})	0.855^{**} (0.370)	2.35	0.348 (0.420)	1.42
GRADE (γ_{50})	−0.131 (0.203)	0.88	−0.027 (0.232)	0.97
Intersection-level				
SHOULDER (γ_{01})	−0.236 (0.262)	0.79	0.143 (0.336)	1.15
SIGNAL (γ_{02})	-0.594^{**} (0.254)	0.55	1.311^{***} (2003)	3.71
DRIVEWAY (γ_{03})	−0.220 (0.262)	0.80	0.152 (0.319)	1.16
HAU (γ_{04})	0.228 (0.299)	1.26	−0.065 (0.335)	0.94
Random effects				
τ_{00} (u_{0j})	0.281^{**} (0.177)		0.548^{**} (0.261)	

Note. For parameter estimates, standard errors are within parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

This result is logical: angle crashes are often the result of aggressive driving of at least one driver (e.g. too fast for conditions, running red light, etc.) which seems more prevalent when conditions are favorable, whereas rear-end crashes occur when following distances are too short—a common occurrence in wet conditions.

Angle crashes at signalized intersections are also less likely (odds = 0.55–1) than at unsignalized intersections, whereas rear-end crashes are more likely at signalized intersections. These findings are consistent with warrants for installing signals and our understanding of the effects of signalization. Translating to a percentage, angle crashes are 45% ((0.55–1) × 100) less likely to occur at signalized intersections than at unsignalized locations. The pavement surface and daylight conditions are likely to reflect in large part the effect of traffic volumes (which were not statistically significant through repeated attempts to include them as continuous or as categories of VMT), which are greater during daylight hours and during dry pavement conditions (in Georgia anyway).

The horizontal curve effect must be taken in the context of angle crashes in close proximity of intersections. Recall that by definition crashes were coded as ‘intersection related’ when they occurred within 76 m (250 ft) of an intersection. Angle crashes under this definition, therefore, represent vehicles turning in the intersection (or before or after the intersection into driveways), or vehicles entering near the intersection from driveways where horizontal curves and thus restricted sight distance exists.

The probability of rear-end crashes is significantly greater at signalized intersections compared to unsignalized intersections (odds ratio = 3.71, or a 271% increase in the probability that rear-end crashes occur at signalized intersections). This finding is consistent with the results of a previous study (Greibe, 2003), and experience with signalized intersections around the US. Wang et al. (2003) suggested that this might be due to the combination of the leading vehicle’s unexpected deceleration by a signal and the ineffective response of the following vehicle’s driver to this deceleration. Another explanation would be differential risk between leading and following vehicle drivers—the following driver being more apt to enter the intersection during the yellow phase. Finally, sight restrictions caused by large leading trucks (or buses, RV’s, etc.) may result in rear-end collisions from sudden stopping of lead vehicles.

5.2. Sideswipe same direction crash and sideswipe opposite direction crash models

For the sideswipe same direction crash model, one crash-level characteristic and two intersection-level characteristics are statistically significant (see Table 6). A curve indicator variable is negatively associated with this type of crash, which indicates that crashes occurring on horizontal curves (near an intersection) are less likely to result in sideswipe same direction crashes. The likely reason is that same direction sideswipe crashes are associated with lane changing/crossing situations, which may arise from avoidance of conflicts in a busy intersection such as sudden right- and left-turning movements by nearby vehicle drivers. In contrast, while on horizontal curves drivers typically face a more predictable driving environment with respect to nearby vehicles. With respect to intersection-level characteristics, sideswipe same direction crashes are less likely to occur at intersections with shoulders than at intersections without shoulders. Intersections with shoulders provide more room for vehicle collision avoidance maneuvering. Finally, same direction sideswipe crashes are more likely to occur at right-angled intersections (i.e. the angle of intersection is 90°) than at skewed intersections. This is probably due to higher volume intersections having less probability of being skewed and thus greater exposure (particularly in the absence of a volume-related variable).

Table 6
Estimation results for sideswipe crashes (both same and opposite direction)

Variable	Sideswipe crashes (same direction)		Sideswipe crashes (opposite direction)	
	Estimate	Odds ratio	Estimate	Odds ratio
Fixed effects				
Intercept (γ_{00})	-2.088^{***} (0.701)	0.12	−0.648 (0.846)	0.53
Crash-level				
CLEAR (γ_{10})	0.219 (0.391)	1.24	-0.776^{**} (0.339)	0.46
SURFACE (γ_{20})	0.389 (0.430)	1.48	-1.130^{***} (0.460)	0.32
DAYLIGHT (γ_{30})	−0.508 (0.330)	0.60	-0.889^{***} (0.332)	0.41
CURVE (γ_{40})	-1.010^{**} (0.398)	0.36	-1.592^{***} (0.553)	0.20
GRADE (γ_{50})	0.151 (0.308)	1.16	-1.144^{***} (0.412)	0.32
Intersection-level				
SHOULDER (γ_{01})	-0.942^{*} (0.560)	0.39	−0.362 (0.726)	0.70
SIGNAL (γ_{02})	0.759 (0.553)	2.13	−0.326 (0.807)	0.72
DRIVEWAY (γ_{03})	−0.202 (0.608)	0.82	−0.276 (0.790)	0.76
HAU (γ_{04})	0.901^{*} (0.461)	2.46	−1.114 (1.288)	0.33
Random effects				
τ_{00} (u_{0j})	2.306^{***} (0.656)		4.219^{***} (1.066)	

Note. For parameter estimates, standard errors are within parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 7
Comparison of coefficients with and without controlling for hierarchical structure

Variable	Angle crashes		Rear-end crashes		Sideswipe crashes (same direction)		Sideswipe crashes (opposite direction)	
	MLM[a]	NMLM[a]	MLM[a]	NMLM[a]	MLM[a]	NMLM[a]	MLM[a]	NMLM[a]
Fixed effects								
Intercept (γ_{00})	−1.021** (0.505)	−0.998** (0.478)	−3.002*** (0.614)	−2.705*** (0.595)	−2.088*** (0.701)	−1.597** (0.788)	−0.648 (0.846)	−0.990 (0.852)
Crash-level								
CLEAR (γ_{10})	0.381* (0.229)	0.344 (0.226)	−0.365 (0.249)	−0.336 (0.257)	0.219 (0.391)	0.351 (0.523)	−0.776** (0.339)	−0.603 (0.533)
SURFACE (γ_{20})	−0.649** (0.289)	−0.706** (0.285)	−0.188 (0.308)	−0.171 (0.316)	0.389 (0.430)	0.323 (0.586)	−1.130*** (0.460)	−0.539 (0.665)
DAYLIGHT (γ_{30})	0.506** (0.231)	0.519** (0.227)	1.140*** (0.299)	1.059*** (0.309)	−0.508 (0.330)	−0.724* (0.427)	−0.889*** (0.332)	−0.481 (0.509)
CURVE (γ_{40})	0.855** (0.370)	0.860** (0.358)	0.348 (0.420)	0.198 (0.427)	−1.010** (0.398)	−1.153** (0.508)	−1.592*** (0.553)	−0.716 (0.678)
GRADE (γ_{50})	−0.131 (0.203)	−0.094 (0.190)	−0.027 (0.232)	−0.026 (0.222)	0.151 (0.308)	0.290 (0.384)	−1.144*** (0.412)	−1.021* (0.581)
Intersection-level								
SHOULDER (γ_{01})	−0.236 (0.262)	−0.217 (0.214)	0.143 (0.336)	0.232 (0.258)	−0.942* (0.560)	−0.621 (0.404)	−0.362 (0.726)	−0.564 (0.504)
SIGNAL (γ_{02})	−0.594** (0.254)	−0.557*** (0.210)	−1.311*** (0.303)	1.157*** (0.242)	0.759 (0.553)	0.269 (0.462)	−0.326 (0.807)	−0.117 (0.523)
DRIVEWAY (γ_{03})	−0.220 (0.262)	−0.218 (0.212)	0.152 (0.319)	0.067 (0.236)	−0.202 (0.608)	−0.123 (0.468)	−0.276 (0.790)	0.188 (0.523)
HAU (γ_{04})	0.228 (0.299)	0.103 (0.267)	−0.065 (0.335)	−0.037 (0.276)	0.901* (0.461)	1.138** (0.515)	−1.114 (1.288)	−1.530 (1.088)

Note. For parameter estimates, standard errors are within parentheses, $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

[a] MLM: coefficients with multilevel modeling techniques; NMLM: coefficients without multilevel modeling techniques.

For sideswipe opposite direction crashes, all of the crash-level characteristics are statistically significant variables, but no statistically significant intersection-level characteristics were observed. All of the crash-level characteristics reveal negative relationships with opposite direction sideswipe crashes. Sideswipe opposite direction crashes are less likely to occur during daytime and clear weather conditions (59 and 54% reduction in probability, respectively). It is hypothesized (post hoc) that drivers can more easily determine potential conflicts during the daytime, that is, an opposing vehicle's trajectory is more readily determined and can be constantly tracked (directly or peripherally). At night-time, in contrast, vehicle headlights prohibit the constant tracking of opposing vehicle trajectories, and so conflicts are relatively more likely (compared to daylight conditions and compared to same-direction sideswipe crashes). In addition, crashes occurring on wet roads, horizontal curves, and/or vertical curves are less apt to be involved in sideswipe opposite direction crashes. This might be due to the fact that drivers traveling on wet roads, horizontal curves, and/or vertical curves are forced to decelerate their speeds by posted speed limit signs or for safety—thus, driver security may be reduced considerably in these driving situations are additional caution is exercised.

5.3. Comparison of coefficients with and without controlling for hierarchical structure

In order to illustrate the impact of using multilevel modeling techniques, all of the crash models were estimated using non-multilevel modeling techniques (traditional binomial logit models). Table 7 shows a comparison of the estimated parameters with and without controlling for hierarchical structure.

The results show that traditional estimation techniques provide similar results for all of the models, except for the sideswipe opposite direction crash model. For the sideswipe opposite direction crash model, all of the crash-level characteristics were found to be statistically significant when multilevel modeling techniques were applied, but only a grade indicator variable was found to be negatively associated with sideswipe opposite direction crashes when traditional estimation techniques were applied. With respect to rear-end crashes, there is no difference between the safety effects of estimated parameters with and without controlling for correlation among rear-end crashes within clusters. For angle and sideswipe same direction crashes, the estimation results of traditional estimation techniques are very close to those of multilevel modeling approaches. Even though there is not much difference between the estimation results of traditional estimation techniques and multilevel modeling approaches for angle, rear-end, and sideswipe same direction crashes, it should be noted that the parameters estimated may have undesirable properties since neglecting significant correlation within clusters results in biased estimators and mis-estimated standard errors. Moreover, the motivation for applying hierarchical methods rests in the desire to explore and estimate within-cluster variation. The multilevel approach provides insight into the influence of hierarchical units on components of variability—useful insight that is not available from non-hierarchical structures.

6. Conclusions and discussions

This paper describes the estimation of statistical models of various crash types given information on geometric, environmental, and traffic information. Using data from rural signalized and unsignalized intersections in Georgia, models that predict the probability of crash occurrence by crash types were estimated for angle, rear-end, and same and opposite direction sideswipe direction crashes. Since the structure of data used in the study is hypothesized to be hierarchical, multilevel modeling techniques were employed to describe the data. We hypothesize that the upper level hierarchy includes intersection characteristics, while the lowest level hierarchy includes crash-level characteristics.

The findings of this study suggest that crashes at rural intersections in Georgia are indeed hierarchical in structure. The modeling results indicate that the effects of geometric characteristics and environmental factors can be modeled using multilevel modeling techniques. Crash data may consist of a hierarchical structure: driver's characteristics are nested within crashes, crash characteristics are nested within site characteristics, site characteristics are nested within regional characteristics, and so forth. As described previously, correlation among crashes within these 'clusters' or hierarchies violates a common regression modeling assumption, which may lead to undesirable model properties. This limitation, however, is overcome through multilevel modeling techniques.

Examination of the random effects suggests that there is a significant variation in the probability of specific types of crashes occurring across intersections (i.e. all of the coefficients of random effects are significantly different from zero). It should be noted that the intersection-level fixed effects for predicting both angle and rear-end crashes capture a significant portion of the variation across intersections, as reflected by random effect coefficients close to zero (0.281 and 0.548, respectively). In terms of both sideswipe same direction and opposite direction crash models, in contrast, the intersection-level random error terms (u_{0j}) suggest that relatively greater unobserved variation exists regarding factors that influence the probability of these crashes across intersections (after the fixed effects intersection-level characteristics have been accounted for). This indicates a need to introduce additional intersection-related explanatory variables for explaining sideswipe crash types in future research investigations. These variables might include signal phasing- and timing-related variables, and intersection complexity-related factors (signage, billboards, channelization, etc.).

In addition to variables employed in this study, it is believed that particular types of crash outcome probabilities may also associated with personal characteristics (e.g. driver attentiveness, reaction times, vision, and aggressiveness) and vehicle characteristics (braking characteristics, mass, steering characteristics, condition of tires, etc.), and including these variables (or reasonable surrogates) into the models may improve the accuracy of the prediction models. The personal characteristics, however, were unavailable for this study.

Many probability models of crashes have revealed that AADT is a significant and important predictor of crash frequencies; however, AADT variables were not found to have significant effects on the probability that certain types of crashes will occur, as opposed to expectation. In order to explore the safety effects of AADT on the probabilities of crash types, a variety of transformation on AADT variables (i.e. categorical variable, indicator variable, log of AADT, etc.) were conducted in different ways, but all expressions of AADT variables were found to be non-significant for predicting crash types. It is known that average daily traffic volumes are strongly associated with crash frequencies because greater volumes lead to greater opportunities for collisions—and so this result is generally troubling. Perhaps, the following explanation for this finding sheds some light. When the fundamental unit of analysis in a model is 'sites', we know that $P(y|x, \theta)$ is the model output and provides the probability of observing y crash counts $y = 1, \ldots, n$ given site and crash covariates x and model parameters θ. In this current analysis, however, the fundamental unit of analysis is a crash (not a site) and the probability models is $P(O|x, \theta)$, where O is crash outcome 0, 1. Thus, increasing exposure clearly should be correlated with larger n in a frequency-based model, whereas increased n does not relate as directly to crash outcome probabilities.

References

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models. Sage, Beverly Hills, CA.

Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accid. Anal. Prev. 35, 253–259.

Dempster, A.P., Rubin, D.B., Tsutakawa, R.K., 1981. Estimation in covariance component models. J. Am. Stat. Assoc. 76, 341–353.

Elston, R.C., Grizzle, J.E., 1962. Estimation of time–response curve and their confidence bands. Biometrics 18 (2), 148–159.

Goldstein, H., 1987. Multilevel Models in Educational and Social Research. Griffin, London.

Greibe, P., 2003. Accident prediction models for urban roads. Accid. Anal. Prev. 35, 273–285.

Guo, G., Zhao, H., 2000. Multilevel modeling for binary data. Annu. Rev. Sociol. 26, 441–462.

Harwood, D.W., Council, F.M., Hauer, E., Hughes, W.E., Vogt, A., 2000. Prediction of the expected safety performance of rural two-lane highways. FHWA-RD-99-207, Washington, DC.

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accid. Anal. Prev. 35, 59–69.

Kim, D., Washington, S., Oh, J., 2006. Modeling crash outcomes: new insights into the effects of covariates on crashes at rural intersections. J. Transport. Eng. 132 (4), 282–292.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38 (4), 963–974.

Longford, N.T., 1987. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika 74 (4), 817–827.

Mason, W.M., Wong, G.Y., Entwistle, B., 1983. Contextual analysis through the multilevel linear model. Sociol. Methodol. 14, 72–103.

Oh, J., Lyon, C., Washington, S., Persaud, B., Bared, J., 2003. Validation of the FHWA crash models for rural intersections: lessons learned. Transport. Res. Rec. 1840, 41–49.

Rosenberg, B., 1973. Random coefficient models: the analysis of a cross-section of time series by stochastically convergent parameter regression. Ann. Econ. Social Meas. 2, 399–428.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometric and environmental factors on rural freeway accident frequencies. Accid. Anal. Prev. 27 (3), 371–389.

Vogt, A., Bared, J., 1998. Accident prediction models for two-lane rural roads: segments and intersections. FHWA-RD-98-133, Washington, DC.

Vogt, A., 1999. Crash models for rural intersections: four-lane by two-lane stop-controlled and two-lane by two-lane signalized. FHWA-RD-99-128, Washington, DC.

Wang, Y., Idea, H., Mannering, F., 2003. Estimating rear-end accident probabilities at signalized intersections: occurrence-mechanism approach. J. Transport. Eng. 129 (4), 377–384.

Wolfinger, R., O'Connell, M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. J. Stat. Comput. Simulat. 48, 233–243.