Accident Analysis and Prevention 99 (2017) 51–65

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A generalized nonlinear model-based mixed multinomial logit approach for crash data analysis

Ziqiang Zeng[a,b,c], Wenbo Zhu[c], Ruimin Ke[c], John Ash[c], Yinhai Wang[c,*], Jiuping Xu[b], Xinxin Xu[a,b]

[a] School of Tourism and Economic Management, Chengdu University, Chengdu, 610106, PR China
[b] Uncertainty Decision-Making Laboratory, Sichuan University, Chengdu, 610064, PR China
[c] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 5 August 2016
Received in revised form 8 November 2016
Accepted 8 November 2016
Available online 18 November 2016

Keywords:
Generalized nonlinear model
Mixed multinomial logit
Crash frequency
Crash severity
Traffic safety

ABSTRACT

The mixed multinomial logit (MNL) approach, which can account for unobserved heterogeneity, is a promising unordered model that has been employed in analyzing the effect of factors contributing to crash severity. However, its basic assumption of using a linear function to explore the relationship between the probability of crash severity and its contributing factors can be violated in reality. This paper develops a generalized nonlinear model-based mixed MNL approach which is capable of capturing non-monotonic relationships by developing nonlinear predictors for the contributing factors in the context of unobserved heterogeneity. The crash data on seven Interstate freeways in Washington between January 2011 and December 2014 are collected to develop the nonlinear predictors in the model. Thirteen contributing factors in terms of traffic characteristics, roadway geometric characteristics, and weather conditions are identified to have significant mixed (fixed or random) effects on the crash density in three crash severity levels: fatal, injury, and property damage only. The proposed model is compared with the standard mixed MNL model. The comparison results suggest a slight superiority of the new approach in terms of model fit measured by the Akaike Information Criterion (12.06 percent decrease) and Bayesian Information Criterion (9.11 percent decrease). The predicted crash densities for all three levels of crash severities of the new approach are also closer (on average) to the observations than the ones predicted by the standard mixed MNL model. Finally, the significance and impacts of the contributing factors are analyzed.

1. Introduction

Traffic crashes cost billions of dollars in life and property damage annually worldwide (Lord and Mannering, 2010). Hence, traffic safety improvement for reducing the frequency of crashes and their severity has long been a top strategic goal of highway agencies (Savolainen et al., 2011). The development of effective countermeasures requires a thorough understanding of the factors that affect the probability of crashes resulting in a given crash severity level.

It is particularly important to take into account crash severities in safety analysis of road segments, because the cost of crashes could be hugely different at different severity levels (Wang et al., 2011). This means that, for instance, a road segment with higher frequency of fatal crashes would almost certainly be considered more hazardous than a road segment with fewer fatal crashes, but more minor injury crashes. Another issue for the traditional frequency based modeling methods is their estimation assumption that fatality rates are identical across locations with different volumes (Milton et al., 2008). Significant error may be introduced when this assumption is violated. Therefore, it is necessary to estimate crash frequency for each severity level.

Exploring the relationship between crash severity and its contributing factors has long been an important problem in traffic safety analysis which attracts many researchers' attention and efforts. Due to the complicated nature of traffic accidents, involving interactions among driver, vehicle, roadway, traffic and environmental components, unobserved heterogeneity cannot be ignored when addressing the relationship between the crash severity and its contributing factors. Many methodological techniques have been developed to account for unobserved heterogeneity by avoiding the restriction that the observable variables must be the same across all observations, which may result in biased parameter estimates and erroneous inferences and predictions. Recently, Mannering et al. (2016) presented a comprehensive review on the

* Corresponding author.
E-mail address: yinhai@u.washington.edu (Y. Wang).

http://dx.doi.org/10.1016/j.aap.2016.11.008
0001-4575/

strengths and weaknesses of the statistical approaches available to address this unobserved heterogeneity. Among these approaches, mixed (random parameter) logit models (Milton et al., 2008; Gkritza and Mannering, 2008; Behnood and Mannering, 2015; Cerwick et al., 2014), latent class models with random parameters within classes (Xiong and Mannering, 2013), and Markov switching models (Malyshkina and Mannering, 2009; Xiong et al., 2014) are widely used in the front-line research on traffic accident severity analysis.

In light of the aforementioned studies, this paper attempts to extend the linear predictors in the mixed (random parameter) multinomial logit model (MNL) to generalized nonlinear predictors which may provide new insights into the exploration of the relationship between the crash severities and their significant contributing factors. In fact, most of the aforementioned studies on mixed logit models can be classified as generalized linear model-based (GLM-based) approaches. Typically, a GLM-based approach utilizes a linear regression to aggregate a series of independent variables, such as roadway curvature, shoulder width, speed limit, etc. and establish a mapping relationship between these variables and a dependent variable (which is typically the expected value of crash rate) through a specific link function (Lao et al., 2014). However, such a GLM-based approach is constrained by its linear model specifications and may lead to biased model estimation and interpretation when the independent variable data demonstrates strong nonlinear features. For instance, increasing lane width may not always reduce crash frequency with a certain severity degree (Lee et al., 2015). Thus, there is a need to consider the nonlinear effects of significant contributing factors on the crash severity in the presence of unobserved heterogeneity.

To address the GLM inherent linear predictor constraint and reveal the nonlinear impacts of significant contributing factors on different crash severity categories in the context of unobserved heterogeneity, the objective of this paper is to develop a generalized nonlinear model-based (GNM-based) mixed MNL approach which is capable of more precisely exploring the influence of the contributing factors and providing new insights into the reduction of crash frequency and severity degree on Interstate freeways in Washington State.

2. Literature review

2.1. Crash severity analysis models

There have been many studies that have applied a wide variety of methodological techniques to explore the relationship between crash severity and its contributing factors over the years (Ye and Lord, 2014), such as binary outcome models (Helai et al., 2008; Lee and Abdel-Aty, 2008), ordered discrete outcome models (Eluru et al., 2010; Quddus et al., 2010), and unordered multinomial discrete outcome models (Washington et al., 2011). Among these methodological techniques, the multinomial logit (MNL) approach has been employed as an unordered approach for analyzing crash severities (Wu et al., 2016; Chen et al., 2015; Murray-Tuite et al., 2014; Tay et al., 2011; Shankar and Mannering, 1996; Kumfer et al., 2015; Bham et al., 2012). Celik and Oktay (2014) performed a MNL analysis to determine the risk factors affecting the severity of traffic injuries in the Erzurum and Kars Provinces of Turkey. In Geedipally et al. (2011) study, crash data from police-reported motorcycle crashes in Texas were used to estimate MNL models to identify differences in factors likely to affect the severity of crash injuries of motorcyclists. In fact, some researchers prefer choosing multinomial logistic models over ordinal models because such models are not afflicted with some of the restrictions imposed by traditional ordered probit and logit models (Ye and Lord, 2014), such as Bayesian ordered probit model, and generalized ordered logit model (Savolainen et al., 2011). As aforementioned, in order to address the unobserved heterogeneity that traditional modeling methods cannot account for, some newer methodological approaches have been employed for crash severity analysis. Milton et al. (2008) demonstrated that the mixed logit model can be used to better understand the injury-severity distributions of accidents on highway segments in the presence of unobserved heterogeneity. Malyshkina and Mannering (2009) proposed two-state Markov switching MNL models for statistical modeling of accident-injury severities which can account for time-varying heterogeneity. Xiong and Mannering (2013) showed that the latent class models with random parameters within classes have the advantage of both the semi-parametric latent classes and fully parametric random parameters when addressing unobserved heterogeneity.

While the applications of these new methodological approaches have undoubtedly provided useful insights, some limitations and disadvantages still exist in these methods which need further exploration (Mannering et al., 2016). In the aforementioned models, the basic assumption of using a linear link function to map the relationship between crash severity and its contributing factors can be sometimes violated in reality. Therefore, extension of the linear predictors in the models to generalized nonlinear predictors in the presence of unobserved heterogeneity has a potential to provide new insights in crash severity analysis.

2.2. Generalized nonlinear models

In order to reflect nonlinear effects of variables to extract more complex relationship, researchers have applied different methods. Many previous studies (Wong et al., 2007; Abdel-Aty and Haleem, 2011) used the logarithm of annual average daily traffic (AADT) instead of AADT to deal with the nonlinear relationship between the crash risk and AADT. Xie et al. (2007) showed that the relationship between lane width and crash frequency is described in a "concave-downward" polynomial function, that is, crash frequency increases as lane width increases from 9 ft to 10 ft and decreases as lane width increases from 10 ft to 13 ft. Moreover, Turner and Firth (2012) developed a package in the statistical computing software package R to estimate the parameters in GNMs. Recently, Lao et al. (2014) proposes a GNM for application in crash analysis. Unlike GLMs, GNMs account for nonlinear effects of independent variables on a dependent variable using a "nonlinearizing" link function. The study demonstrated that right shoulder width, AADT, grade percentage, and truck percentage have nonlinear effects on rear-end crashes. They also found that GNMs can better reflect the nonlinear relationships than GLMs based on residual deviance. However, while the aforementioned studies have provided valuable insights for traffic safety analyses, most of them only studied the nonlinear relationship between the contributing factors and the crash frequency (or rate) without considering the impacts associated with the crash severity and unobserved heterogeneity. In fact, the presence of unobserved heterogeneity can create spurious non-linearity in conventional models. The non-linear predictors may simply be picking up unobserved heterogeneity and not actual non-linearity (Mannering et al., 2016). Thus, exploring the non-linear effects of the contributing factors in the presence of unobserved heterogeneity is critical in crash severity analyses.

3. Data description

In order to develop the GNM-based mixed MNL approach, appropriate data on crash frequency by severity levels and contributing factors to crashes are essential. In this section, the data collection is introduced, and then a data quality control method based on segment length is developed to remove short road segments and improve the quality of the collected data. A multi-collinearity analysis is used for addressing the collinearity among correlated contributing factors.

3.1. Data collection

This study was performed based on crash data records collected in Washington State from January 2011 to December 2014 (i.e., a four-year period). The data were obtained from the Washington State Department of Transportation (WSDOT), Highway Safety Information System (HSIS), and the Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net) platform at the University of Washington (UW). Four major datasets are included in this study: crash data, roadway geometric characteristics, traffic characteristics, and weather conditions. These datasets detail all of the information regarding crash frequency, locations, severities, roadway segment length, average number of lanes (NOL), horizontal curve type (HCT), curvature of the segment (COS), average width of outer shoulder (WOS), average width of inner shoulder (WIS), average width of median (WM), dominant lane surface type (DLST), dominant outer shoulder type (DOST), dominant inner shoulder type (DIST), dominant median type (DMT), average speed limit (ASL), AADT, AADT per lane, road surface conditions (RSC, i.e., dry, wet, snow/ice/slush), and visibility (good, bad). It should be noted that the traffic characteristics and weather conditions are regarded as dynamic factors, while the roadway geometric characteristics are regarded as static factors. There are three types of road surface conditions and two types of visibility, thus a single road segment could be under any one of six combinations of road surface and visibility conditions. For instance, in a single segment in the dataset on which if a total of 50 crashes occurred during last 4 years, 30 of them could have happened during days in which there was a dry surface condition and good visibility, 10 of them could have occurred under dry surface and bad visibility conditions and 5 of them could have occurred under wet surface and good visibility conditions; all other possible combinations could have zero corresponding crashes.

The lane surface types include asphalt (A), bituminous (B), and Portland cement concrete (P); for outer and inner shoulder, the types include A, B, curb (C), gravel (G), P, soil (S), wall (W), and other (O); for medians, the types include A, B, G, P, S, and O. If a certain surface type covers more length than other types within a road segment, it will be considered as the dominant type. From our analysis we found the dominant type and its percentage length of the segments have more significant impacts on the crash data than the non-dominant types. Thus, these categorical variables, i.e., DLST, DOST, DIST, and DMT, are selected as the explanatory variables for roadway geometric characteristics.

Crash data were collected in terms of crash frequencies with different severity levels. Usually, statistical models are produced for all crash severity levels (often referred to as KABCO, i.e., fatal (K), incapacitating-injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO or O)) or for different crash severity levels, such as fatal and nonfatal injury crashes (e.g., KABC) or for PDO crashes only. Although the data on crash frequency by severity are multivariate in nature, they have often been analyzed by modeling each severity level separately, without taking into account correlations that exist among different severity levels. In this study, due to limited numbers of crash records, the collected crash frequency data were classified into three categories: fatal, injury, and PDO.

The candidate sites for data collection were selected as a total of 21,396 roadway segments on various Interstate highways in Washington including I-5, I-90, I-82, I-182, I-205, I-405 and I-705. The total number of crashes recorded during the data collection period was 48,154, including 134 fatal crashes, 13,936 injury crashes, and 34,084 PDO crashes. Fig. 1 illustrates the study area for I-5, I-90, and I-82. The orange, red, and green lines denote the selected road segments on I-5, I-90, and I-82, respectively. Of all segments, those on I-5 combine to have the highest number of fatal crashes across all Interstate freeways in the US. Within the 276.54 miles of road segments included in this study that are on I-5, there were a total of 29,601 crashes.

3.2. Data quality control based on segment length

Special care was taken to screen out the incomplete and outlier data to enhance the data quality. For the continuous variables, the outliers located beyond three standard deviations from the mean were removed from the datasets with respect to all the variables in the datasets. For the categorical variables, the categories with less than 1 percent observations with respect to a total of 21,396 roadway segments were removed from the datasets. Furthermore, in order to guarantee higher quality of the datasets, the road segments with short lengths should be properly addressed.

In this research, segmentation of the roadways based on curvature is employed, i.e., a new segment starts when a tangent section transitions into curve. Thus, the segment lengths are not fixed. Generally, crash locations are based on the police reports. Police reports indicate the location based on the spot where the vehicle was located upon the police response and thus do not always indicate the place that was where the crash was caused. Further, the number of crashes is proportional to segment length. Therefore, if there is a very short segment which is <100 feet, for example, the probability of an accident occurring in such a location is very small. Thus, road segments with short lengths should be removed from the dataset.

Segmentation, when based on multiple variables, may lead to very short homogeneous segments (Resende and Benekohal, 1997). For example, when using the segmentation approach proposed by the *Highway Safety Manual* (HSM), the presence of very short segments does not allow proper statistical inference for several reasons (American Association of State Highway and Transportation Officials (AASHTO), 2010). Perhaps the most important reason is that the locations of crashes, often taken from police reports, are not exact (Quin and Wellner, 2012). Also, since crashes are rare events, a great number of segments end up with zero crashes during the analysis period. Lengthening segments to avoid these issues can sacrifice homogeneity.

In the literature, there are a number of different approaches to segmentation. Miaou and Lum (1993) suggested that short sections, less than or equal to 80 m in length could create bias in the estimation of some linear models, but not when using Poisson models. Similarly, Ogle et al. (2011) demonstrated that short segment lengths, less than 160 m, can lead to uncertain results in crash analyses. Cafiso and Di Silvestro (2011) showed that to increase performance in identifying correct positives as black spots, segment length should be related to AADT with lower AADT values requiring longer segment lengths. Quin and Wellner (2012) studied the relationship between segmentation and safety screening analysis using different lengths of sliding windows to identify hazardous sites, and they concluded that short segments as well as those that are too long (based on threshold values) create a bias in the identification of sites with safety problems.

Some studies focused on the relationship between crashes and road geometry in addressing segmentation. For example, Cenek et al. (1997), who investigated this relationship for data collected on rural roads, used a fixed segment length of 200 m. A similar study was done by Cafiso et al. (2008) using homogeneous sections with different lengths on a sample of Italian two lane rural roads and aggregating variables related to curvature and roadside hazards. They concluded that models that contain geometry

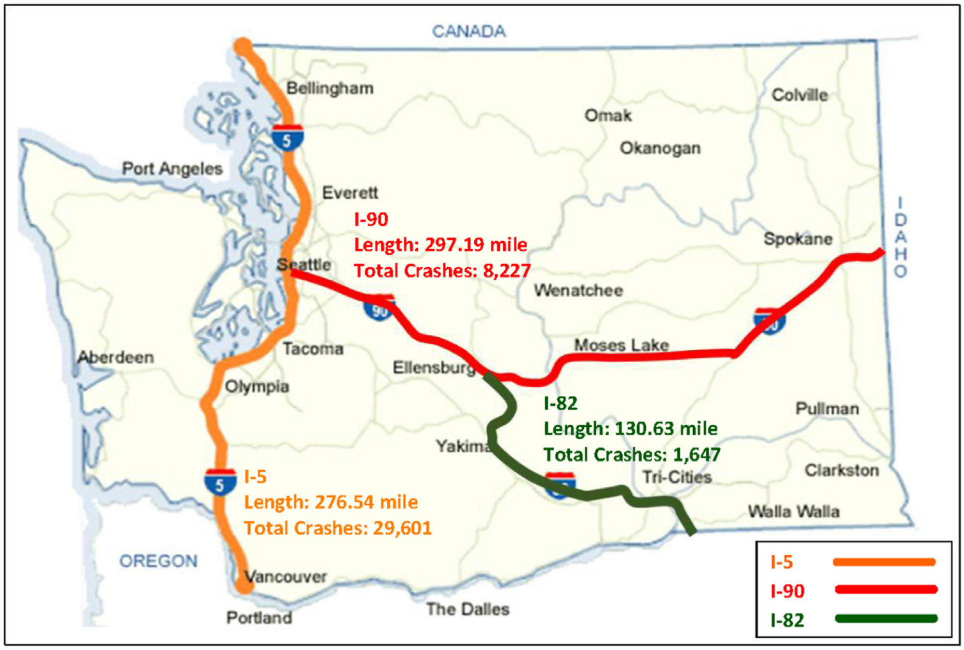

Fig. 1. Study area for I-5, I-90, and I-82 in Washington.

and design consistency variables are more reliable than those that do not. Other studies suggested different ways to aggregate segment data to avoid lengths that are too short. For example, Koorey (2009) proposed the aggregation of curves and tangents when the radius of curves exceeds a predetermined threshold value. The HSM (American Association of State Highway and Transportation Officials (AASHTO), 2010) recommends the use of homogeneous segments with respect to AADT, number of lanes, curvature, presence of ramps at interchanges, lane width, outside and inside shoulder widths, median width and clear zone width. However, there is no prescribed minimum segment length for application of the predictive models based on HSM guidance.

Based on the literature, there is no exact value for defining the length of a "short section." Ultimately, the best segment length depends on the particular dataset. In this research, a sensitivity analysis was made for determining the threshold value for short road segments. Table 1 shows the results of the sensitivity analysis for removing short segments that have a segment length less than the threshold values.

According to the results of the sensitivity analysis in Table 1, the trend of the loss of crashes (where a lost crash is one not included in the analysis as the short segment it occurred on was removed from the dataset) experiences a dramatic increase when the segment length threshold value exceeds 0.05 mile as shown in Fig. 2. In order to avoid losing too much information in the datasets, we select 0.05 mile as the threshold value, and remove all of the road segments having length less than 0.05 mile. The summary statistics of the continuous and categorical variables for the road segments after removing outliers and short segments are shown in Tables 2 and 3 respectively. The data of crash frequency with different severity levels for the road segments after removing outliers and short segments are shown in Table 4.

Additionally, a multicollinearity analysis was conducted on some highly-correlated contributing factors, which indicated that the AADT and AADT per lane have relatively high values of variance inflation factors (VIFs) (27.145 and 17.542, respectively). Ultimately, the presence of this multicollinearity leads to the insignificance of AADT as compared with AADT per lane; thus, in this research, AADT per lane was selected as one of the potential contributing factors for the following modeling analysis; the other contributing factors were not found to have significant multicollinearity with each other.

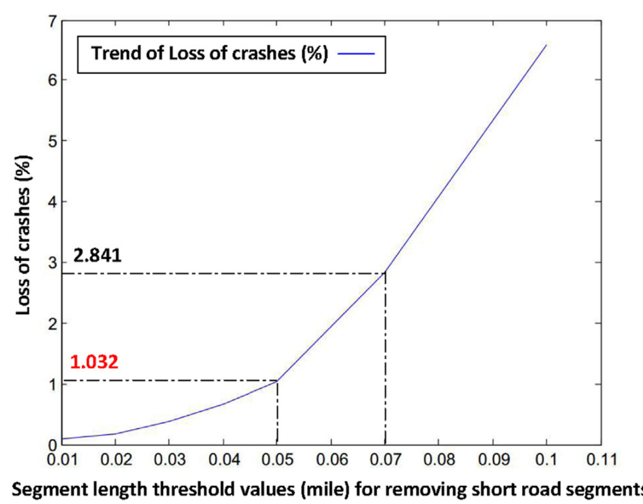

Fig. 2. Trend of the loss of crashes (%) in regarding to the segment length threshold values.

4. Methodology

In this section, the concept of a GNM is introduced and the extended GNM-based mixed MNL is developed for crash data analysis.

4.1. GNMs for crash prediction in context of unobserved heterogeneity

In classical linear regression models, the expectation of crash frequency (or rate) is formulated as an ordinary linear model. This

Table 1
Sensitivity analysis results for removing short segments that have a segment length less than the threshold value.

	Full dataset	Removing short segments that have a segment length less than						
		0.01 mile	0.02 mile	0.03 mile	0.04 mile	0.05 mile	0.07 mile	0.10 mile
Total observations	21396	21228	21084	20862	20598	20406	19482	17850
Total crashes	48154	48110	48071	47976	47831	47657	46786	44984
Loss of crashes (%)	–	0.091	0.172	0.370	0.671	1.032	2.841	6.583
Crash/year	12038.5	12027.5	12017.75	11994	11957.75	11914.25	11696.5	11246

Table 2
Summary statistics of continuous variables for Interstate freeway segments in the Washington State for years 2011–2014 after removing outliers and short segments.

Factor Type	Classification Type	Explanatory variables	Min.	Max.	Mean	Median	St.Dev.
Static	Roadway geometric characteristics	Segment length	264	71227.20	2342.11	1372.8	3615.54
		NOL	2	5	2.6	2	0.715
		COS	0	6.030	0.726	0	1.009
		WOS (feet)	0	18	6.596	8	3.741
		WIS (feet)	0	18	3.770	4	2.843
		WM (inch)	5.59	999	94.89	68	156.79
		ASL (mile/h)	46.67	70	65.58	70	4.87
Dynamic	Traffic characteristics	AADT	6700	229500	68305.42	44000	58910.24
		AADT per lane	1675	57375	12118.15	9500	8999.12

Note: NOL = number of lanes; COS = curvature of the segment; WOS = average width of outer shoulder; WIS = average width of inner shoulder; WM = average width of median; ASL = average speed limit; AADT = annual average daily traffic.

Table 3
Summary statistics of categorical variables for Interstate freeway segments in the Washington State for years 2011 – 2014 after removing outliers and short segments.

Factor Type	Classification Type	Explanatory variables	Number of categories	Category Types	Number of crashes (Percentage)
Static	Roadway geometric characteristics	HCT	3	S = Straight	32864 (68.96%)
				L = Left	7428 (15.59%)
				R = Right	7365 (15.45%)
		DLST	3	A = Asphalt	18266 (38.33%)
				B = Bituminous	1029 (2.16%)
				P = Portland Cement Concrete	28362 (59.51%)
		DOST	6	A = Asphalt	36362 (76.30%)
				B = Bituminous	1081 (2.27%)
				C = Curb	1896 (3.98%)
				P = Portland Cement Concrete	4739 (9.94%)
				W = Wall	2444 (5.13%)
				O = Other	1135 (2.38%)
		DIST	6	A = Asphalt	20251 (42.49%)
				B = Bituminous	1076 (2.26%)
				C = Curb	1765 (3.71%)
				P = Portland Cement Concrete	2428 (5.09%)
				W = Wall	3141 (6.59%)
				O = Other	18996 (39. 86%)
		DMT	4	A = Asphalt	16416 (34.45%)
				P = Portland Cement Concrete	3643 (7.64%)
				S = Soil	22463 (47.13%)
				O = Other	5135 (10.78%)
Dynamic	Weather conditions	RSC	3	Dry	30969 (64.98%)
				Wet	13104 (27.50%)
				Snow/Ice/Slush	3584 (7.52%)
		Visibility	2	Good	29630 (62.17%)
				Bad	18027 (37.83%)

Note: HCT = horizontal curve type; DLST = dominant lane surface type; DOST = dominant outer shoulder type; DIST = dominant inner shoulder type; DMT = dominant median type; RSC = road surface conditions.

Table 4
Data of crash frequency for Interstate freeway segments in the Washington State for years 2011–2014 after removing outliers and short segments.

Total crashes	Fatal crashes	Injury crashes	PDO crashes	Average crashes per year	Number of road segments
47657	134	13824	33699	11914.25	20406

model specification can be expressed as follows (McCullagh and Nelder, 1989):

$$E\left(y_i\right) = \mu_i = L_i \sum_{j=1}^{J} x_{ij}\beta_j + \beta_0 + \varepsilon_i, \quad (1)$$

where y_i denotes the crash frequency (or rate) along roadway segment i; $E(y_i)$ or μ_i is the expected crash frequency (or rate) along segment i during a certain time period; L_i is the segment length in miles; x_{ij} is the jth explanatory variable for segment i; β_j is the corresponding regression coefficient for the jth explanatory variable; β_0 is an intercept (constant) term; ε_i is an error term; and

J is the total number of explanatory variables considered in the model. Compared to the simplest linear regression, more complicated models, such as Poisson and negative binomial models for crash frequency and logit and probit models for crash severity have been used to interpret crash data. These models can be generalized by using a smooth and invertible linearizing link function to transform the expectation of the response variable, μ_i, to its linear predictor:

$$g\ (\mu_i)\ =\ L_i\sum_{j=1}^{J}x_{ij}\beta_j\ +\ \beta_0\ +\ \varepsilon_i, \tag{2}$$

where g(.) is the link function, which is monotonic, differentiable and used to connect the linear predictor of the explanatory variables with the expected crash frequency (or rate) in various formats, such as identity, log, logit, etc. In this research, the log link function is used for crash analysis.

As was discussed earlier, in many scenarios the relationship between the expected crash frequency (or rate) by severity level and its associated factors cannot be simply expressed by GLMs. GNMs are hence proposed as an extension of GLMs in order to satisfy such specific requirements by changing the linear predictor, to be nonlinear, in Eq. (2).

The GNM-based method uses a user-defined, customized function to extract the relationship between crash risks and contributing factors with more general assumptions. For the other explanatory variables, the diverse set of possible nonlinear predictors, $U(x)$, such as polynomial functions, exponential functions, logarithmic functions, etc. may be utilized to extract proper data features. In general, a model of the format of $U(x)$ can be determined based on statistical analysis of the crash rate and a specific explanatory variable. Notice that the defined nonlinear function $U(x)$ is an assumed relationship. This defined function can be revised based on further statistical analysis. Aggregating the nonlinear predictors for all the independent variables, Eq. (1) can be rearranged as:

$$E\ (y_i)\ =\ \mu_i\ =\ L_i\sum_{j=1}^{J}U_j\ (x_{ij})\ \omega_{(j)}\ +\ \beta_0\ +\ \varepsilon_i,\ i\ =\ 1, 2, \cdots, n, \tag{3}$$

where $U_j(x_{ij})$ is a nonlinear predictor for the jth explanatory variable; β_0 is an intercept term; ε_i is an error term; ω_j is the corresponding weight for $U_j(x_{ij})$; and n is the number of observations. Consequently, the GNM link functions becomes:

$$g\ (\mu_i)\ =\ \sum_{j=1}^{J}U_j\ (x_{ij})\ \omega_{(j)}\ +\ \beta_0\ +\ \varepsilon_i,\ i\ =\ 1, 2, \cdots, n, \tag{4}$$

If all of the $U_j(x_{ij})$ in the model are linear regressions of x_{ij}, a GNM will degrade to a GLM. Therefore, GLMs are simply special cases of GNMs. To make road sections comparable, in this research, $g(\mu_i)$ was considered as a logarithmic function and applied on the basis of crash density (i.e., crash frequency per mile) as shown below:

$$g\ \ (\mu_i) = \ln(d_i) = \ln\left(\frac{\mu_i}{L_i y_i}\right) = \sum_{j=1}^{J}U_j\left(x_{ij}\right)\ \omega_{\ (j)}$$
$$+\,\beta_0\ +\ \varepsilon_i\ =\ U_i\omega\ +\ \beta_0\ +\ \varepsilon_i,\ i\ =\ 1, 2, \cdots, n, \tag{5}$$

where $d_i = \mu_i/L_i y_i$, L_i, and y_i are the crash density, segment length, and time period length (years) of crash frequency of roadway segment i respectively; $\omega = [\omega_{(1)}, \omega_{(2)}, \cdots, \omega_{(J)}]^T$ is the coefficient vector for $U_i = [U_1(x_{i1}), U_2(x_{i2}), \cdots, U_J(x_{iJ})]$ when estimating the expected crash density; β_0 is an intercept term; and ε_i is an error term. In order to account for the accident-specific unobserved heterogeneity, a mixing distribution is introduced giving the ω vector a continuous density function $f(\omega|\phi)$, where ϕ is a vector of parameters characterizing the chosen density function (such as location and scale). If the contributing factors related to scale in the vector ω are determined to be significantly different from zero, there will be accident-specific variations in the effect of one or more contributing factors of the explanatory vector U_i on crash severity. If ω is fixed, this implies no accident-specific unobserved heterogeneity.

4.2. GNM-based mixed MNL approach

Logistic regression is generally used to handle categorical data (Bham et al., 2012). It can handle binary response variables, i.e., variables with two possible values, and can be extended to handle a multinomial response variable Y that takes a discrete set of values reflecting K categories (where K is greater than or equal to two). Since the response variable is nominal (unordered), a generalized mixed MNL model is suitable. This approach frames K-1 logits for the response variable to compare each categorical level with a reference category.

In this research, three categories are considered for the crash severity (i.e., PDO (k = 1), injury (k = 2), and fatal (k = 3)). Since crashes with a lower severity such as PDO collisions are more likely to go unreported, crashes of a relatively higher severity are often overrepresented and crashes with lower severity are often underrepresented in datasets. It has been widely accepted that fatal crashes have the highest reporting rate and PDO crashes have the lowest reporting rate, thus the observations of the fatal crashes are closer to the actual number than that of the PDO crashes. Ye and Lord (2011) demonstrated that fatal crashes should be set as the baseline severity for the mixed MNL model. To minimize the bias and reduce the variability of a model, in this paper, fatal crashes were used as the baseline severity category for comparison with the other categories. The crash severity type, denoted by Y, was the response variable, whereas contributing factors for roadway geometric characteristics, traffic characteristics, and weather conditions were the independent variables denoted by x_{ij} (j = 1, 2, . . ., J), where i denotes the observation and J denotes the number of independent variables. Y is then defined as follows:

$$Y = \begin{cases} 1,\ \textit{if crash is PDO}, \\ 2,\ \textit{if crash is injury}, \\ 3,\ \textit{if crash is fatal}, \end{cases} \tag{6}$$

Based on the GNM link functions in Eq. (5), when the response categories 1, . . ., K (K = 3) are unordered, Y is related to independent variables through a set of K-1 baseline category logits as shown in the following severity function S_{ki}:

$$S_{ki} = \ln\left(\frac{Pr\,(Y_i = \mathrm{k})}{Pr\,(Y_i = \mathrm{K})}\right) = \sum_{j=1}^{J}U_{kij}\left(x_{ij}\right)\omega_{kj} + \beta_{k0} + \varepsilon_{ki}$$
$$= U_{ki}\omega_k + \beta_{k0} + \varepsilon_{ki},\ i = 1, 2, \cdots, n;\ k = 1, 2, \cdots, K-1, \tag{7}$$

where $Pr(Y_i = k)$ is the probability of crash severity type k; $U_{ki} = \left[U_{ki1}\ (x_{i1}), U_{ki2}\ \left(x_{i2}\right), \cdots, U_{kiJ}\ (x_{iJ})\right]$ is the nonlinear predictor vector of observation i for contributing factors (i.e., roadway geometric characteristics, traffic characteristics, weather conditions); $\omega_k = [\omega_{k1}, \omega_{k2}, \cdots, \omega_{kJ}]^T$ is the coefficient vector for the kth category of the predictor vector; β_{k0} is an intercept term specific to crash severity type k; and ε_{ki} is an error term. McFadden (1981) has shown that if ε_{ki} is assumed to be generalized extreme value distributed, the standard MNL model is estimable by standard maximum likelihood techniques, resulting in,

$$Pr\left(Y_i = \mathrm{k}\right) = \frac{e^{U_{ki}\omega_k+\beta_{k0}}}{\sum_{k=1}^{K} e^{U_{ki}\omega_k+\beta_{k0}}},\ i = 1, 2, \cdots, n;\ k = 1, 2, \cdots, K, \tag{8}$$

By exponentiating both sides of Eq. (7), and solving for the probabilities, we get:

$$Pr\left(Y_i = \mathrm{k}\right) = Pr\left(Y_i = \mathrm{K}\right) e^{U_{ki}\omega_k+\beta_{k0}},\quad i = 1, 2, \cdots, n;\ k = 1, 2, \cdots, K-1, \tag{9}$$

Using the fact that all k of the probabilities must sum to one, we find:

$$Pr\ \left(Y_i = K\right) = \frac{1}{1+\sum_{k=1}^{K-1} e^{U_{ki}\omega_k+\beta_{k0}}},\ i = 1, 2, \cdots, n, \tag{10}$$

According to Eqs. (9) and (10), the other probabilities can be expressed as below:

$$Pr\ \ \left(Y_i = \mathrm{k}\right) = \frac{e^{U_{ki}\omega_k+\beta_{k0}}}{1+\sum_{k=1}^{K-1} e^{U_{ki}\omega_k+\beta_{k0}}},\quad i = 1, 2, \cdots, n;\ k = 1, 2, \cdots, K-1, \tag{11}$$

In order to account for the unobserved heterogeneity, let $\Omega = (\omega_1, \omega_2, \omega_3)$, as discussed previously, and note that the Ω vector has a continuous density function $f(\Omega|\Gamma)$, where Γ is a vector of parameters charactering the density function. According to Eqs. (10) and (11), the resulting mixed MNL crash severity probabilities are as follows (McFadden and Train, 2000; Train, 2003; Hensher and Greene, 2003):

$$Pr\left(Y_i = \mathrm{K}\right) = \int \frac{1}{1+\sum_{k=1}^{K-1} e^{U_{ki}\omega_k+\beta_{k0}}} f\ \left(\Omega|\Gamma\right) d\Omega,\quad i = 1, 2, \cdots, n, \tag{12}$$

$$Pr\left(Y_i = \mathrm{k}\right) = \int \frac{e^{U_{ki}\omega_k+\beta_{k0}}}{1+\sum_{k=1}^{K-1} e^{U_{ki}\omega_k+\beta_{k0}}} f\ \left(\Omega|\Gamma\right) d\Omega,\quad i = 1, 2, \cdots, n;\ k = 1, 2, \cdots, K-1, \tag{13}$$

Since U_{ki} is considered as a nonlinear predictor vector of observation i for contributing factors, Eqs. (12) and (13) are called the prediction functions of the GNM-based mixed MNL approach. In Eqs. (12) and (13), if all of the $U_j(x_{ij})$ in the nonlinear predictor vector are linear functions of x_{ij}, the GNM-based mixed MNL approach will degrade to a standard mixed MNL approach. The GNM-based mixed MNL crash severity probabilities are then a weighted average for different values of ω_k across roadway segments where some elements of the vector ω_k may be fixed and some may be randomly distributed. If all the elements of the vector ω_k are fixed, the GNM-based mixed MNL approach is reduced to a GNM-based standard MNL approach.

4.3. Modeling of expected crash densities in different severity levels

By considering the unobserved heterogeneity, the expected crash density in a roadway segment i can be estimated as below:

$$d_i = \int e^{U_i\omega+\beta_0} f\left(\omega|\varphi\right) d\omega,\ i = 1, 2, \cdots, n, \tag{14}$$

where $d_i = \mu_i/L_i y_i$ is the crash density of roadway segment i during a certain time period.

According to Eqs. (12)–(14), the expected crash density for different severity levels can be estimated as follows.

(1) Expected PDO crash density:

$$d_{i1} = d_i \cdot Pr\left(Y_i = 1\right) = \int e^{U_i\omega+\beta_0} f\left(\omega|\varphi\right) d\omega \cdot \int \frac{e^{U_{1i}\omega_1+\beta_{10}}}{1+\sum_{k=1}^{2} e^{U_{ki}\omega_k+\beta_{k0}}} f\ \left(\Omega|\Gamma\right) d\Omega,\quad i = 1, 2, \cdots, n, \tag{15}$$

where d_{i1} is the expected PDO crash density along segment i during a certain time period; $\omega = [\omega_{(1)}, \omega_{(2)}, \cdots, \omega_{(J)}]^T$ is the coefficient vector for $U_i = [U_1(x_{i1}), U_2(x_{i2}), \cdots, U_J(x_{ij})]$ when estimating the expected crash density.

(2) Expected injury crash density:

$$d_{i2} = d_i \cdot Pr\left(Y_i = 2\right) = \int e^{U_i\omega+\beta_0} f(\omega|\varphi) d\omega \cdot \int \frac{e^{U_{2i}\omega_2+\beta_{20}}}{1+\sum_{k=1}^{2} e^{U_{ki}\omega_k+\beta_{k0}}} f\ \left(\Omega|\Gamma\right) d\Omega,\quad i = 1, 2, \cdots, n, \tag{16}$$

where d_{i2} is the expected injury crash density along segment i during a certain time period;

(3) Expected fatal crash density:

$$d_{i3} = d_i \cdot Pr\left(Y_i = 3\right) = \int e^{U_i\omega+\beta_0} f(\omega|\varphi) d\omega \cdot \int \frac{1}{1+\sum_{k=1}^{2} e^{U_{ki}\omega_k+\beta_{k0}}} f\ \left(\Omega|\Gamma\right) d\Omega,\quad i = 1, 2, \cdots, n, \tag{17}$$

where d_{i3} is the expected fatal crash density along segment i during a certain time period. Based on Eqs. (15)–(17), the expected crash densities can be simulated and analyzed in the following section.

5. Modeling and discussion

In this section, the parameters of the predictors and coefficients in the GNM-based mixed MNL are estimated and the expected crash density functions in different severity levels are also analyzed, which are detailed as follows.

5.1. Estimation of nonlinear predictors

It is necessary to determine the appropriate predictors $U_j(x_{ij})$ and $U_{kj}(x_{ij})$ in Eqs. (5) and (7) before the corresponding coefficients $\omega_{(j)}$ and ω_{kj} can be calibrated. To better illustrate the nonlinear contribution function estimation process, an example is detailed to formulate the contribution function for the variable, AADT per lane, as follows. Assume the number of crashes for each severity level

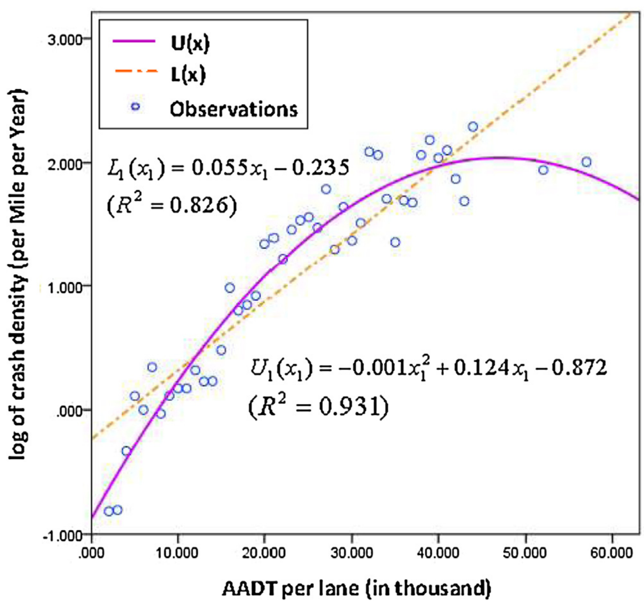

Fig. 3. Logarithm of the expectation of crash density (number of crashes per mile per year) from Interstate freeway segments in the Washington State for years 2011–2014, by AADT per lane.

follows the Poisson distribution, and all other dependent variables are approximately consistent across different AADT levels when the sample data is large enough. In this research, the confidence level of each contributing factor was set at $\alpha = 0.05$.

Let j = 1 denote the index of the contributing factor AADT per lane. To develop an appropriate form of the predictor $U_1(x_{i1})$, the visualized comparisons between the logarithm of the expectation of crash density (number of crashes per mile per year) and AADT per lane are illustrated in Fig. 3. The data points in the scatter plot show the logarithm of crash density from the Interstate freeway segments in Washington for the years 2011–2014, classified by AADT per lane. As one can see, the logarithm of the average crash density tends to increase when the AADT per lane increases at a variable rate. The rate of increase becomes smaller as AADT per lane increases, which indicates the inappropriateness of using a linear contribution function. To address this issue, a second degree polynomial was used as the nonlinear predictor to approximate the impacts of AADT on crash density:

$$U_1(x_1) = -0.001x_1^2 + 0.124x_1 - 0.872 \quad (18)$$

Compared to the linear predictor $L_1(x_1) = 0.055x_1 - 0.235$, the value of R^2 increases from 0.826 to 0.931 when the nonlinear predictor, $U_1(x_1)$, is utilized as shown in Fig. 3. Therefore, the nonlinear predictor appears to be more suitable to describe the relationship between crash density and AADT, and it is thus employed in this study. It should be noted that in the model estimation process, the second degree polynomial form of the nonlinear predictors in this study performs better than the logarithmic form which was employed in Lao et al. (2014). The nonlinearity of AADT and its impacts on crash frequencies have been found significant and these results are consistent with many previous studies (Lao et al., 2014; Wong et al., 2007; Abdel-Aty and Haleem, 2011).

Similar procedures can be applied to develop the functional forms of the predictors $U_{11}(x_{i1})$ and $U_{21}(x_{i1})$ in severity function S_{ki} as shown in the following:

$$U_{11}(x_1) = -0.004x_1^2 + 0.227x_1 + 0.055 \text{ and } U_{21}(x_1) = 0.132x_1^2 - 0.126x_1 + 0.108 \quad (19)$$

where the value of R^2 increases from 0.543 to 0.803 as compared to the linear predictor $L_{11}(x_1) = 0.066x_1 + 1.311$ by using the nonlinear predictor $U_{11}(x_{i1})$, and from 0.709 to 0.780 as compared to the linear predictor $L_{21}(x_1) = 0.133x_1$-0.422 by using the nonlinear predictor $U_{21}(x_{i1})$. Fig. 4 shows the regression results for estimating the two nonlinear predictors for AADT per lane in the severity function.

For the other continuous contributing factors, similar procedures can be carried out to develop the nonlinear predictor functions. Fig. 5. illustrates the regression results for estimating the nonlinear predictors for the continuous contributing factors NOL (j = 2), WOS (j = 3), WIS (j = 4), WM (j = 5), and ASL (j = 6). The regression results shown in Fig. 5 indicate significant nonlinearity of the predictors in the severity function when considering average number of lanes (NOL), average width of outer shoulder (WOS), average width of inner shoulder (WIS), average width of median (WM), and average speed limit (ASL). As a result, quadratic and cubic regression equations are employed for these predictors to extract proper data features. For the predictors of the logarithm of the expectation of crash density by NOL and ASL, no strong nonlinear associations are observed between the R^2 values of the nonlinear predictors and linear predictors are less than 0.015) and linear predictors thus appear to be sufficient to characterize their impacts on crash data which are consistent with the results of the previous study by Lao et al. (2014). Table 5 summarizes the estimated parameters for the predictor functions of the continuous contributing factors.

For the categorical contributing factors, including HCT (j = 7), DLST (j = 8), DOST (j = 9), DIST (j = 10), DMT (j = 11), RSC (j = 12), and Visibility (j = 13), an N-category class function is developed as follows:

$$U_j(x_{ij}) = \begin{cases} \pi_{1j}, \ \textit{category}\ 1, \\ \ldots \\ \pi_{Nj}, \ \textit{category}\ N, \end{cases} \quad (20)$$

where π_{nj} is an estimated sub-predictor function for category n of a given contributing factor. Consider HCT as an example. This contributing factor contains three categories, i.e., straight, left, and right. The curvature of the segment for each type of category is utilized as the independent variable to estimate the formation of the sub-predictor. Similar methods as were used for the continuous contributing factors were conducted to develop the sub-predictor functions of the categorical contributing factors. Summary statistical analysis on the datasets showed that the impact of the segment curvature is similar between the left and right categories. For the categorical contributing factors DLST, DOST, DIST, and DMT, the percentage length of dominant type in the segment is utilized as the independent variable to estimate the formation of the sub-predictor. Table 6 summarizes the estimated parameters for each categorical sub-predictor function.

It should be mentioned that the aforementioned empirical results of the nonlinearity or linearity of specific contributing factors are data-specific and cannot be generalized. This means the results of whether a nonlinear or linear predictor should be employed for a certain contributing factor could be different based on datasets used in the subject research.

5.2. Estimation of random coefficients in GNM-based mixed MNL model

The maximum simulated likelihood (MSL) estimation method (Train, 2003) is used to estimate the random coefficient vector ω in Eq. (5) and coefficient vector ω_k in Eq. (7). In this paper, for the functional form of the parameter density functions, the normal, log-normal, uniform, and triangular distributions are tested as potential mixing distributions. In fact, numerous distributional forms have been examined for model parameters in past research, includ-

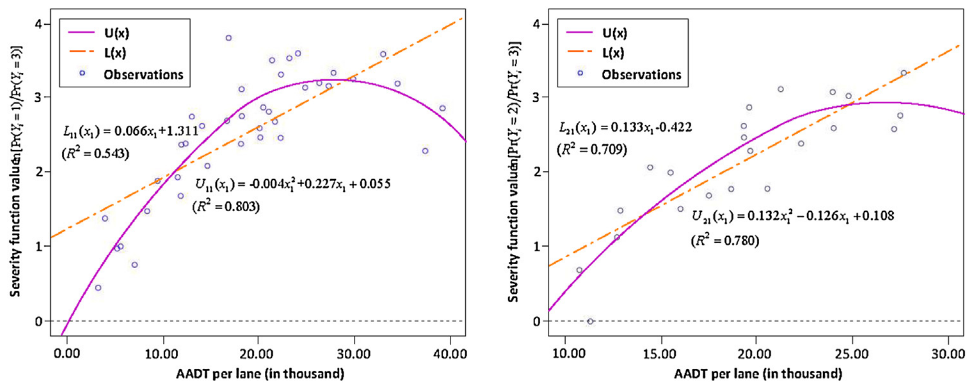

Fig. 4. Regression results for estimating the two nonlinear predictors for AADT per lane in the severity function.

Table 5
Summary of the estimated parameters for the predictor functions of the continuous contributing factors.

Index Number	Contributing Factor	Predictor	Coefficients				F-value	R^2
			b_0	b_1	b_2	b_3		
1	AADT	Crash Density	−0.872	0.124	−0.001	–	283.66	0.931
	per	PDO/Fatal	0.055	0.227	−0.004	–	72.315	0.803
	lane	Injury/Fatal	0.108	−0.126	0.132	–	45.352	0.780
2	NOL	Crash Density	−13.257	5.048	–	–	941.782	0.915
		PDO/Fatal	1.347	−1.868	1.225	−0.158	235.903	0.669
		Injury/Fatal	3.850	−1.883	0.419	–	38.036	0.817
3	WOS	Crash Density	5.560	−0.323	−0.07	–	126.389	0.751
		PDO/Fatal	3.301	−0.349	0.032	–	71.514	0.827
		Injury/Fatal	3.806	−0.616	0.048	–	86.147	0.910
4	WIS	Crash Density	5.333	−0.352	−0.288	–	151.923	0.781
		PDO/Fatal	3.237	−0.621	0.128	−0.006	256.623	0.795
		Injury/Fatal	3.222	−0.529	0.052	–	113.298	0.923
5	WM	Crash Density	7.010	−0.113	0.0004	–	186.891	0.815
		PDO/Fatal	3.704	−0.019	0.000055	–	49.614	0.750
		Injury/Fatal	2.717	−0.007	0.0000188	–	24.909	0.714
6	ASL	Crash Density	39.453	−0.602	–	–	1541.349	0.947
		PDO/Fatal	167.354	−4.887	0.036	–	90.453	0.866
		Injury/Fatal	171.181	−5.145	0.039	–	118.966	0.933

Note: b_0, b_1, b_2, b_3 are the estimated coefficients of the constant, linear, quadratic, and cubic terms of the predictor functions respectively.

Table 6
Summary of the estimated parameters for the sub-predictor functions of the categorical contributing factors.

Index Number	Contributing Factor	Category	Crash Density			PDO/Fatal			Injury/Fatal		
			b_0	b_1	b_2	b_0	b_1	b_2	b_0	b_1	b_2
7	HCT	Straight	0.41	–	–	1.71	–	–	1.17	–	–
		Left	0.41	0.22	–	1.66	0.98	−0.37	2.02	−1.11	0.25
		Right	0.38	0.27	–	−1.40	5.51	−1.87	−0.36	3.05	−1.06
8	DLST	Asphalt	2.76	−0.03	–	2.91	−0.01	–	1.73	−0.01	–
		Bituminous	−0.95	–	–	−0.34	–	–	−0.23	–	–
		Portland Cement Concrete	0.521	–	–	−9.75	0.12	–	−8.10	0.10	–
9	DOST	Asphalt	1.67	−0.15	–	2.471	−0.01	–	0.09	0.011	–
		Bituminous	−1.77	0.25	–	0.01	–	–	−0.03	–	–
		Curb	−0.49	0.02	–	3.21	–	–	2.45	–	–
		Portland Cement Concrete	1.26	−0.002	–	2.322	−0.006	–	−9.05	0.21	–
		Wall	1.24	0.002	–	4.95	−0.026	–	3.86	−0.02	–
		Other	−1.53	0.024	–	0.02	–	–	0.69	–	–
10	DIST	Asphalt	0.92	−0.006	–	2.87	−0.013	–	1.64	−0.01	–
		Bituminous	−2.46	0.032	–	0.02	–	–	0.01	–	–
		Curb	0.39	0.013	–	3.13	–	–	2.45	–	–
		Portland Cement Concrete	0.99	0.008	–	4.35	−0.016	–	3.86	−0.02	–
		Wall	0.88	0.008	–	0.88	0.017	–	0.81	0.015	–
		Other	1.56	−0.007	–	4.28	−0.025	–	1.02	0.003	–
11	DMT	Asphalt	1.04	−0.002	–	0.31	0.016	–	−2.14	0.035	–
		Portland Cement Concrete	0.22	0.008	–	8.19	−0.067	–	1.098	–	–
		Soil	1.33	−0.01	–	3.49	−0.02	–	1.88	−0.01	–
		Other	1.282	–	–	4.00	−0.011	–	1.79	–	–
12	RSC	Dry	0.665	–	–	1.724	–	–	1.213	–	–
		Wet	0.675	–	–	1.747	–	–	1.279	–	–
		Snow/Ice/Slush	0.345	–	–	1.539	–	–	0.652	–	–
13	Visibility	Good	0.648	–	–	2.089	–	–	1.314	–	–
		Bad	0.572	–	–	1.207	–	–	0.909	–	–

Note: b_0, b_1, b_2 are the estimated coefficients of the constant, linear, and quadratic terms of the sub-predictor functions respectively.

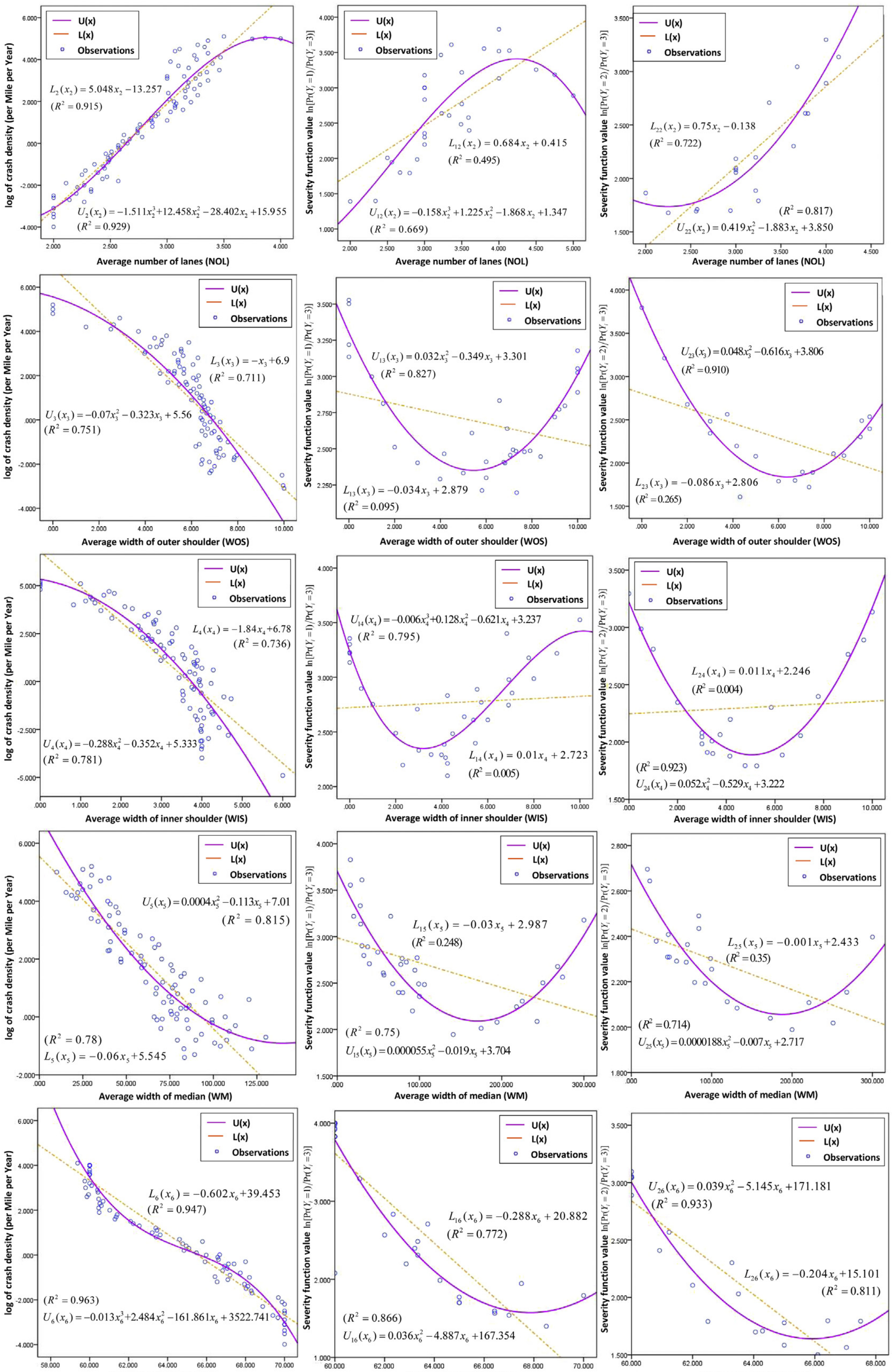

Fig 5. Regression results for estimating the nonlinear predictors for the continuous contributing factors NOL (j = 2), WOS (j = 3), WIS (j = 4), WM (j = 5), and ASL (j = 6).

ing normal, lognormal, uniform, triangular, dome, Erlang, Weibull, exponential, and nonstochastic distributions. The prevailing findings from the literature conclude that the normal distribution generally establishes the best fit for injury severity data (Milton et al., 2008; Gkritza and Mannering, 2008). Consistent with past research, the normal distribution was found to provide the best estimation results and thus was employed in this study.

The number of simulations required is strictly dependent upon the complexity of the model. More simulation points are required as the number of randomized variables increase. Past research has investigated the performance of various random draw routes, including Halton, random, and shuffled draws (Bhat, 2003; Hensher et al., 1999). Hensher et al. (1999) concluded that Halton draws of a specific number will produce results as accurate as ten times that number in random draws. Research by Bhat (2003), Anastasopoulos and Mannering (2009), and others have shown that 200 Halton draws is usually sufficient for accurate parameter estimation. In this paper, the random coefficients in the GNM-based mixed MNL model were estimated with NLOGIT 4.0 (Hensher and Greene, 2003) using 200 Halton draws and the typical simulation run took about one half to one hour to complete on a ThinkPad X1 Carbon Core i5 computer which is a time-consuming and computationally cumbersome process. Exploring more causes of heterogeneity, such as driver, vehicle, and environmental conditions, etc., would increase the computation time considerably. Table 7 shows the results of the GNM-based mixed MNL estimation which was based on the collected data from Interstate freeways in Washington from 2011 to 2014. Since the baseline or reference severity category in our study is fatal crash, the results are therefore reported in the following categories; one is for the likelihood of "PDO/Fatal" crash (i.e., logarithmic value of the probability a given crash results in PDO compared to a fatality), which refers to coefficients of the severity function (i.e., Eq. (7)) when $k = 1$; another is for the likelihood of "Injury/Fatal" crash (i.e., logarithmic value of the probability a given crash results in an injury compared to a fatality), which refers to the coefficients of the severity function (i.e., Eq. (7)) when $k = 2$; and the final one is for the likelihood of the logarithm of crash density, which refers to the coefficients of Eq. (5).

The estimated coefficients therefore show the mixed (random or fixed) effects and significance of a contributing factor on the likelihood of a "PDO/Fatal" crash, "Injury/Fatal" crash, and the logarithm of crash density. Overall, the model fit the data fairly well, with relatively small values of Akaike Information Criterion (AIC) (i.e., 287.492) and Bayesian Information Criterion (BIC) (i.e., 158.654) which were used as measures of the goodness-of-fit. Estimation findings indicate that the traffic characteristics, e.g, AADT per lane are best modeled as random-parameters for the likelihood of "PDO/Fatal" crash and the logarithm of crash density, and as fixed parameters for the likelihood of "Injury/Fatal" crash—while roadway geometric characteristics such as NOL, WOS, WIS, WM, ASL, and DOST are best modeled as fixed parameters for the likelihood of the logarithm of crash density, and as random parameters for the likelihood of "PDO/Fatal" and "Injury/Fatal" crash. The intercept terms, weather conditions such as RSC and visibility, and the rest of the roadway geometric characteristics such as HCT, DLST, DIST, and DMT were modeled as random parameters for all categories. It is important to note that since the preceding empirical results are data-specific, they cannot be generalized.

5.3. *Comparison of GNM-based mixed MNL and standard mixed MNL models*

In order to demonstrate the merit of using the GNM-based mixed MNL model, a comparison analysis was conducted between the GNM-based mixed MNL and the standard mixed MNL models. The nonlinear predictor vector in Eqs. (5) and (7) were regarded

Table 7
Estimation results for the GNM-based mixed MNL model (values in parentheses indicate the standard error of the random coefficients).

Contributing Factors	Crash density (ω_j)			PDO/Fatal (ω_{1j})			Injury/Fatal (ω_{2j})		
	Coeff.	Std. error	t-statistic	Coeff.	Std. error	t-statistic	Coeff.	Std. error	t-statistic
Traffic characteristics									
AADT per lane (j = 1)	0.493 (0.672)	0.031 (0.127)	12.074 (3.458)	0.513 (0.104)	0.127 (0.258)	2.753 (0.879)	0.006 (fixed)	0.002 (fixed)	1.936 (fixed)
Roadway geometric characteristics									
NOL (j = 2)	0.042 (fixed)	0.003 (fixed)	5.872 (fixed)	0.062 (0.088)	0.028 (0.103)	2.073 (1.214)	0.397 (0.487)	0.205 (0.322)	1.915 (0.878)
WOS (j = 3)	0.010 (fixed)	0.002 (fixed)	2.948 (fixed)	−0.236 (0.342)	0.302 (0.287)	−0.798 (1.043)	0.138 (0.175)	0.234 (0.197)	0.598 (0.787)
WIS (j = 4)	0.002 (fixed)	0.001 (fixed)	0.912 (fixed)	−0.018 (0.105)	0.047 (0.098)	−0.245 (0.158)	−0.039 (0.055)	0.259 (0.212)	−0.142 (1.331)
WM (j = 5)	0.004 (fixed)	0.002 (fixed)	6.437 (fixed)	−0.168 (0.274)	0.241 (0.344)	−0.601 (1.182)	−0.121 (0.214)	0.067 (0.081)	−1.483 (1.562)
ASL (j = 6)	0.069 (fixed)	0.006 (fixed)	8.116 (fixed)	−0.127 (0.312)	0.076 (0.128)	−1.588 (0.946)	−0.078 (0.103)	0.103 (0.122)	−0.842 (0.753)
HCT (j = 7)	−0.048 (0.076)	0.031 (0.185)	−1.502 (1.215)	−0.028 (0.088)	0.042 (0.086)	−0.712 (0.843)	0.268 (0.337)	0.272 (0.313)	0.976 (0.866)
DLST (j = 8)	−0.338 (0.129)	0.058 (0.097)	−4.723 (3.898)	0.025 (0.046)	0.018 (0.068)	1.732 (1.396)	−0.042 (0.056)	0.059 (0.091)	−0.584 (0.676)
DOST (j = 9)	0.004 (fixed)	0.003 (fixed)	2.108 (fixed)	0.503 (1.209)	0.236 (0.247)	4.813 (1.574)	−0.026 (0.048)	0.056 (0.058)	−0.876 (0.824)
DIST (j = 10)	−0.203 (0.187)	0.061 (0.083)	−3.408 (2.316)	0.011 (0.035)	0.078 (0.107)	0.117 (0.098)	0.004 (0.012)	0.138 (0.146)	0.015 (0.038)
DMT (j = 11)	−0.169 (0.233)	0.055 (0.042)	−2.948 (1.653)	0.037 (0.104)	0.072 (0.121)	0.603 (0.545)	0.102 (0.256)	0.172 (0.206)	0.552 (0.785)
Weather conditions									
RSC (j = 12)	1.572 (1.286)	0.087 (0.069)	18.013 (5.982)	−0.791 (1.287)	0.098 (0.086)	−2.215 (1.898)	0.063 (0.083)	0.615 (0.476)	0.097 (0.822)
Visibility (j = 13)	1.681 (1.342)	0.287 (0.355)	10.012 (7.385)	0.823 (0.684)	0.197 (0.267)	3.983 (5.212)	0.718 (0.693)	0.502 (0.342)	1.432 (1.677)
Intercept term									
β_0 or β_{k0}	−0.163 (0.122)	0.013 (0.078)	−6.892 (2.895)	0.042 (0.132)	0.038 (0.074)	2.136 (3.216)	−0.017 (0.108)	0.021 (0.065)	−1.064 (0.896)
AIC	287.492								
BIC	158.654								

as linear predictor vectors, and the same estimation method was employed to estimate the standard mixed MNL model based on the collected data from Interstate freeways in Washington from 2011 to 2014. Table 8 shows the results of the standard mixed MNL estimation. A few notable differences in Tables 7 and 8 were found between the two models. The roadway geometric characteristics such as HCT and DLST were modeled as fixed parameters for the likelihood of "PDO/Fatal" crash in the standard mixed MNL model, while they were modeled as random parameters for the likelihood of "PDO/Fatal" crash in the GNM-based mixed MNL model. The same situation applied to HCT and the intercept term for the likelihood of the logarithm of crash density. The comparison results also suggested a slight superiority of the GNM-based mixed MNL approach in terms of model fit measured by AIC (12.06 percent decrease) and BIC (9.11 percent decrease).

To further compare the two models, ten roadway segments were randomly selected from the collected data. The two models were employed to predict the crash densities for different severity levels and compared to the observed data for the selected ten roadway segments as shown in Table 9. It can be observed that the GNM-based mixed MNL approach predicted crash densities for all three levels of crash severities were closer (on average) to the observations than the ones predicted by the standard mixed MNL approach.

6. Results analysis

The results in Figs. 3–5 indicate that the nonlinear predictors outperform the linear predictors for most of the contributing factors in the GNM-based mixed MNL approach as they result in better goodness-of-fit based on the crash data collected in this study. Since the preceding empirical results are data-specific and cannot be generalized, the statistical superiority of the non-linear approach is not overwhelming. However, more sensible and applicable explanations for the relationship between crash densities at different severity levels and the contributing factors can be extracted by the GNM-based mixed MNL approach which paves the way for deeper research. The findings of this approach could be useful to identify significant influential factors and develop more robust and applicable countermeasures against crashes at different severity levels. Based on the results reported in Tables 5–7, the effects of different contributing factors on the different crash severity levels can be identified.

Among the roadway geometric characteristics, a total of ten contributing factors were identified as significant with respect to crash occurrence in the GNM-based mixed MNL approach, including NOL (j = 2), WOS (j = 3), WIS (j = 4), WM (j = 5), and ASL (j = 6), HCT (j = 7), DLST (j = 8), DOST (j = 9), DIST (j = 10), and DMT (j = 11). In this study, the estimation results in Table 7 indicate that only NOL (j = 2) has a positive fixed coefficient or is normally distributed with positive means for its predictors in all crash severity categories, the other contributing factors in terms of roadway geometric characteristics have a hybrid (positive or negative) mixed effect (fixed or random) on their predictors in different crash severity categories.

Statistical analysis was also applied on the DLST (j = 8), DOST (j = 9), DIST (j = 10), and DMT (j = 11). The results indicate that increasing the percentage length of the asphalt type on these contributing factors performs best in terms of reducing the probabilities of the total, PDO, and injury crashes, while increasing the percentage length of the soil type on DMT leads to a decrease in the likelihood of fatal crashes as compared with other types.

The weather conditions include two categorical contributing factors, i.e., RSC (j = 12) and visibility (j = 13), both of which were identified as significant in the GNM-based mixed MNL model. For the contributing factor RSC (j = 12), the average proportions of durations of dry, wet, and snow/ice/slush conditions in Washington

Table 8
Estimation results for the standard mixed MNL model (values in parentheses indicate the standard error of the random coefficients).

Contributing Factors	Crash density (ω_j)			PDO/Fatal (ω_{1j})			Injury/Fatal (ω_{2j})		
	Coeff.	Std. error	t-statistic	Coeff.	Std. error	t-statistic	Coeff.	Std. error	t-statistic
Traffic characteristics									
AADT per lane (j = 1)	0.472 (0.308)	0.054 (0.078)	6.326 (2.163)	0.489 (0.357)	0.134 (0.147)	3.605 (1.225)	0.008 (fixed)	0.003 (fixed)	2.064 (fixed)
Roadway geometric characteristics									
NOL (j = 2)	0.053 (fixed)	0.002 (fixed)	4.468 (fixed)	0.071 (0.105)	0.035 (0.084)	1.852 (1.452)	0.411 (0.592)	0.313 (0.503)	1.422 (0.564)
WOS (j = 3)	0.011 (fixed)	0.003 (fixed)	4.833 (fixed)	−0.218 (0.522)	0.285 (0.406)	−0.762 (1.335)	0.143 (0.201)	0.322 (0.243)	1.288 (0.674)
WIS (j = 4)	0.003 (fixed)	0.002 (fixed)	1.247 (fixed)	−0.016 (0.087)	0.065 (0.103)	−0.623 (0.548)	−0.041 (0.068)	0.341 (0.676)	−0.562 (0.815)
WM (j = 5)	0.005 (fixed)	0.001 (fixed)	0.881 (fixed)	−0.172 (0.356)	0.263 (0.238)	−0.833 (0.928)	−0.125 (0.115)	0.083 (0.143)	−2.182 (1.477)
ASL (j = 6)	0.073 (fixed)	0.004 (fixed)	1.002 (fixed)	−0.132 (0.564)	0.082 (0.116)	−1.322 (0.845)	−0.069 (0.198)	0.093 (0.105)	−0.524 (0.832)
HCT (j = 7)	−0.045 (fixed)	0.006 (fixed)	−0.911 (fixed)	−0.032 (fixed)	0.005 (fixed)	−0.636 (fixed)	0.273 (0.128)	0.385 (0.226)	0.816 (0.722)
DLST (j = 8)	−0.346 (0.113)	0.067 (0.104)	−2.355 (1.063)	0.028 (fixed)	0.006 (fixed)	0.858 (fixed)	−0.038 (0.098)	0.073 (0.102)	−0.645 (0.878)
DOST (j = 9)	0.004 (fixed)	0.003 (fixed)	1.575 (fixed)	0.722 (0.973)	0.321 (0.612)	1.808 (0.937)	−0.031 (0.077)	0.055 (0.101)	−0.731 (0.715)
DIST (j = 10)	−0.195 (0.203)	0.082 (0.098)	−2.112 (1.466)	0.013 (0.032)	0.082 (0.075)	0.122 (0.109)	0.005 (0.026)	0.121 (0.168)	0.087 (0.106)
DMT (j = 11)	−0.156 (0.126)	0.063 (0.057)	−1.266 (1.355)	0.049 (0.098)	0.083 (0.218)	0.722 (0.757)	0.097 (0.109)	0.089 (0.095)	0.411 (0.355)
Weather conditions									
RSC (j = 12)	1.513 (1.065)	0.094 (0.072)	11.336 (4.629)	−0.814 (1.054)	0.102 (0.095)	−1.017 (1.486)	0.067 (0.092)	0.524 (0.765)	0.168 (0.627)
Visibility (j = 13)	1.328 (1.284)	0.364 (0.466)	8.321 (5.446)	0.805 (0.747)	0.258 (0.322)	2.115 (3.007)	0.742 (0.815)	0.658 (0.523)	1.126 (1.078)
Intercept term									
β_0 or β_{k0}	−0.158 (fixed)	0.005 (fixed)	−0.901 (fixed)	0.045 (0.098)	0.056 (0.106)	0.967 (0.818)	−0.019 (0.083)	0.034 (0.077)	−0.946 (0.568)
AIC	326.918								
BIC	174.556								

Table 9
GNM-based mixed MNL and standard mixed MNL approach estimated crash densities for different crash severity levels compared to observations of ten roadway segments.

Roadway segments index	Crash density (per mile per year)								
	PDO			Injury			Fatal		
	GNM	Stand.	Observ.	GNM	Stand.	Observ.	GNM	Stand.	Observ.
Segment 1	1.623	1.742	1.562	0.332	0.376	0.312	0.085	0.082	0.104
Segment 2	13.826	14.243	13.043	2.336	2.574	2.173	1.026	0.964	1.086
Segment 3	16.932	18.461	17.307	5.548	5.326	5.769	1.885	1.874	1.923
Segment 4	86.705	88.387	84.091	46.895	46.238	47.727	2.218	2.136	2.272
Segment 5	76.565	81.428	78.125	30.879	30.088	31.250	3.104	3.067	3.125
Segment 6	21.332	19.897	20.588	7.245	7.659	7.352	1.453	1.412	1.471
Segment 7	4.269	4.102	4.457	0.948	0.912	0.969	0.184	0.172	0.193
Segment 8	15.338	15.174	15.625	7.678	7.326	7.812	1.535	1.498	1.562
Segment 9	18.105	17.846	18.750	6.146	6.468	6.250	1.025	0.956	1.041
Segment 10	37.889	37.232	38.462	16.452	17.238	16.761	1.354	1.303	1.389
Average deviation	0.783	1.376	–	0.228	0.510	–	0.041	0.070	–

Table 10
Impact intensity index for RSC and visibility.

Contributing Factors	Category Types	Total crash density			PDO crash density			Injury crash density			Fatal crash density		
		M_{q0}	P_q	I_{q0}	M_{q1}	P_q	I_{q1}	M_{q2}	P_q	I_{q2}	M_{q3}	P_q	I_{q3}
RSC	Dry	3.882	18.4	0.21	2.714	18.4	0.15	1.157	18.4	0.063	0.0115	18.4	0.00063
(j = 12)	Wet	1.685	6.2	0.27	1.204	6.2	0.19	0.480	6.2	0.077	0.0015	6.2	0.00024
	Snow/ice/slush	0.367	1	0.37	0.279	1	0.28	0.087	1	0.087	0.0010	1	0.001
Visibility	Good	2.551	2.32	1.09	1.777	2.32	0.76	0.767	2.32	0.33	0.0062	2.32	0.0027
(j = 13)	Bad	1.405	1	1.40	1.021	1	1.02	0.382	1	0.38	0.0028	1	0.0028

between January 2011 and December 2014 were employed to calculate the impact intensity index as below:

$$I_{qk} = \frac{M_{qk}}{P_q}, q = 1, 2, 3; k = 0, 1, 2, 3 \tag{21}$$

where I_{qk} is the impact intensity index; M_{qk} is the mean value of total (k = 0), PDO (k = 1), injury (k = 2), or fatal (k = 3) crash density; P_q is the proportion of duration of the dry (q = 1), wet (q = 2), or snow/ice/slush (q = 3) condition. The results in Table 10 showed that snow/ice/slush road surface conditions lead to higher impact intensity of total, PDO, injury, and fatal crash densities than dry and wet conditions which are consistent with previous studies (Brijs et al., 2008; Xu et al., 2013).

In regards to visibility (j = 13), similar calculations were applied and the results in Table 10 showed that bad conditions have higher impact intensity of total, PDO, injury and fatal crash densities than good condition. The differences of their impact intensity are more significant for lower crash severity levels.

7. Conclusions and future research

In order to help better understand the nonlinear effects of some specific contributing factors on crash frequency and severity, this paper extended the traditional mixed multinomial logit model by developing a GNM-based mixed MNL approach in which nonlinear predictors were utilized to replace linear predictors in the presence of unobserved heterogeneity. The goal of such modeling was to allow for better capturing of non-monotonic relationships between the independent and dependent variables. If the predictors estimated are found to be linear instead of nonlinear for all variables considered, the GNM-based mixed MNL model will degrade to a traditional mixed MNL model.

In previous studies that applied the traditional mixed MNL model, the basic assumption of using a linear function to explore the relationship between crash frequencies and their contributing factors can be violated in reality. The GNM-based mixed MNL approach is capable of having higher flexibility and providing more reasonable explanations and new insights into planning highway safety improvements. The crash data on seven Interstate freeways in Washington State were collected to develop the nonlinear predictors in the model. A total of 21,396 interstate freeway road segments in I-5, I-90, I-82, I-182, I-205, I-405 and I-705 in Washington State were selected as the candidate sites for crash data collection. 48,154 crashes reported by the police between January 2011 and December 2014 were collected from the aforementioned road segments and classified into three crash severity levels: fatal, injury, and property damage only. A data quality control method was employed to remove short road segments. Thirteen contributing factors in terms of traffic characteristics, roadway geometric characteristics, and weather conditions were identified to have significant mixed (fixed or random) effects on the crash densities at different crash severity levels.

A comparison of the GNM-based mixed MNL and the traditional mixed MNL models was made based on model fit and predicted crash densities for all three levels of crash severities that were considered. The results suggested a slight superiority of the GNM-based mixed MNL approach in terms of model fit based on the crash data collected for use in this study. The predicted crash densities for all three levels of crash severity of the GNM-based mixed MNL approach are also closer (on average) to the observations than the ones predicted by the traditional mixed MNL model. The expected crash density prediction functions for each severity level were developed for a further analysis.

The proposed GNM-based mixed MNL approach is useful to capture non-monotonic relationships by developing nonlinear predictors for the contributing factors while addressing the unobserved heterogeneity. The new approach has higher applicability than the traditional mixed MNL model in crash severity analyses and can be applied to hotspot identification to further the practice more accurate traffic and roadway network safety assessment.

Future research will focus on the following five directions: (1) Analyzing the effects of contributing factors on the crash densities at different severity levels by considering the collision types, such as rear-end, head-on, and run-off-road collisions; (2) Considering more types of contributing factors, including driver behavior-related factors, vehicular factors, and environmental variables; (3) Developing a GNM-based mixed MNL model with correlated random parameters to account for complex unobserved heterogeneity; (4) Combining the Markov-switching approach with the GNM-based mixed MNL model to account for time-varying heterogeneity; (5) Developing a new safety performance index based on the GNM-based mixed MNL approach for hotspot identification.

Acknowledgements

This research was supported in part by the Pacific Northwest Transportation Consortium, US Department of Transportation University Transportation Center for Federal Region 10, the Hi-Tech Research and Development Program of China (Grant No. 2014BAG03B04), the National Natural Science Foundation of China (Grant Nos. 71501137 and 51329801), the General Program of China Postdoctoral Science Foundation (Grant No. 2015M572480), the International Postdoctoral Exchange Fellowship Program of China Postdoctoral Council (Grant No. 20150028), the project of Research Center for System Sciences and Enterprise Development (Grant No. Xq16B05), and Sichuan University (Grant No. skqy201647). The authors would like to give our great appreciations to the editors and anonymous referees for their helpful and constructive comments and suggestions, which have helped to improve this paper.

References

Abdel-Aty, M., Haleem, K., 2011. Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accid. Anal. Prev. 43 (1), 461–470.

American Association of State Highway and Transportation Officials (AASHTO), 2010. Highway Safety Manual (HSM), first edition. American Association of State Highway and Transportation Officials, U.S.A.

Anastasopoulos, P.Ch., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.

Behnood, A., Mannering, F.L., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: some empirical evidence. Anal. Methods Accid. Res. 8, 7–32.

Bham, G.H., Jawadi, B.S., Manepalli, U.R.R., 2012. Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban U.S. highways in Arkansas. J. Transp. Eng.-ASCE 138 (6), 786–797.

Bhat, C., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transp. Res. Part B 37 (1), 837–855.

Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accid. Anal. Prev. 40 (3), 1180–1190.

Cafiso, S., Di Silvestro, G., 2011. Performance of safety indicators in identification of black spots on two-lane rural roads. Transp. Res. Rec. 2237, 78–87.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., 2008. Safety performance indicators for local rural roads: comprehensive procedure from low-cost data survey to accident prediction model. In: Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C.

Celik, A.K., Oktay, E., 2014. A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. Accid. Anal. Prev. 72, 66–77.

Cenek, P.D., Davies, R.B., McLarin, M.W., Griffith-Jones, G., Locke, N.J., 1997. Road environment and traffic crashes. In: Transfund New Zealand Research Report 79 (Wellington).

Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Anal. Methods Accid. Res. 3–4, 11–27.

Chen, C., Zhang, G.H., Tarefder, R., Ma, J.M., Wei, H., Guan, H.Z., 2015. A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Accid. Anal. Prev. 80, 76–88.

Eluru, N., Paleti, R., Pendyala, R., Bhat, C., 2010. Modeling multiple vehicle occupant injury severity: a copula-based multivariate approach. Transp. Res. Rec. 2165, 1–11.

Geedipally, S.R., Turner, P.A., Patil, S., 2011. Analysis of motorcycle crashes in Texas with multinomial logit model. Transp. Res. Rec. 2265, 62–69.

Gkritza, K., Mannering, F.L., 2008. Mixed logit analysis of safety-belt use in single- and multi-occupant vehicles. Accid. Anal. Prev. 40, 443–451.

Helai, H., Chor, C., Haque, M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accid. Anal. Prev. 40 (1), 45–54.

Hensher, D.A., Greene, W.H., 2003. The mixed logit model: the state of practice. Transportation 30 (2), 133–176.

Hensher, D., Louviere, J., Swait, J., 1999. Combining sources of preference data. J. Econ. 89, 197–221.

Koorey, G., 2009. Road data aggregation and sectioning considerations for crash analysis. Transp. Res. Rec. 2103, 61–68.

Kumfer, W., Wei, D.L., Liu, H.C., 2015. Effects of demographic and driver factors on single-vehicle and multivehicle fatal crashes investigation with multinomial logistic regression. Transp. Res. Rec. 2518, 37–45.

Lao, Y., Zhang, G., Wang, Y., Milton, J., 2014. Generalized nonlinear models for rear-end crash risk analysis. Accid. Anal. Prev. 62, 9–16.

Lee, C., Abdel-Aty, M., 2008. Presence of passengers: does it increase or reduce driver's crash potential? Accid. Anal. Prev. 40 (5), 1703–1712.

Lee, C., Abdel-Aty, M., Park, J., Wang, J.H., 2015. Development of crash modification factors for changing lane width on roadway segments using generalized nonlinear models. Accid. Anal. Prev. 76, 83–91.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A 44 (5), 291–305.

Malyshkina, N.V., Mannering, F.L., 2009. Markov switching multinomial logit model: an application to accident-injury severities. Accid. Anal. Prev. 41 (4), 829–838.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. J. Appl. Econ. 15, 447–470.

McFadden, D., 1981. Econometric models of probabilistic choice. In: Manski, C., McFadden, D. (Eds.), A Structural Analysis of Discrete Data with Econometric Applications. The MIT Press, Cambridge, MA, pp. 198–269.

Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prev. 25 (6), 689–709.

Milton, J., Shankarb, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40 (1), 260–266.

Murray-Tuite, P., Wernstedt, K., Yin, W.H., 2014. Behavioral shifts after a fatal rapid transit accident: a multinomial logit model. Transp. Res. Part F 24, 218–230.

Ogle, J., Alluri, P., Sarasua, W., 2011. MMUCC and MIRE: The role of segmentation in safety analysis. In: Transportation Research Board Annual Meeting Proceedings, Washington D.C.

Quddus, M., Wang, C., Ison, S., 2010. Road traffic congestion and crash severity: an econometric analysis using ordered response models. J. Transp. Eng. 136 (5), 424–435.

Quin, X., Wellner, A., 2012. Segment length impact on highway safety screening analysis. In: TRB 2012 Annual Meeting, Washington, D.C.

Resende, P., Benekohal, R., 1997. Effect of roadway section length on accident modeling traffic congestion and traffic safety. In: The 21 St Century Conference, ASCE, Chicago, IL.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. J. Safety Res. 27 (3), 183–194.

Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian-vehicle crash severity. Int. J. Sust. Transp. 5 (4), 233–249.

Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, UK.

Turner H., Firth D, 2012. Generalized nonlinear models in R: An overview of the gnm package. R package version 1.0-6. http://CRAN.Rproject.org/package=gnm.

Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accid. Anal. Prev. 43 (6), 1979–1990.

Washington, S., Karlaftis, M., Mannering, F., 2011. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.

Wong, S.C., Sze, N.N., Li, Y.C., 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. Accid. Anal. Prev. 39 (6), 1107–1113.

Wu, Q., Zhang, G.H., Ci, Y.S., Wu, L.N., Tarefder, R.A., Alcantara, A., 2016. Exploratory multinomial logit model-based driver injury severity analyses for teenage and adult drivers in intersection-related crashes. Traffic Inj. Prev. 17 (4), 413–422.

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

Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury seventies: a finite-mixture random-parameters approach. Transp. Res. Part B 49, 39–54.

Xiong, Y., Tobias, J., Mannering, F.L., 2014. The analysis of vehicle crash injury-severity data: a Markov switching approach with road-segment heterogeneity. Transp. Res. Part B 67, 109–128.
Xu, C.C., Wang, W., Liu, P., 2013. Identifying crash-prone traffic conditions under different weather on freeways. J. Safety Res. 46, 135–144.
Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models multinomial logit, ordered probit, and mixed logit. Transp. Res. Rec. 2241, 51–58.
Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Anal. Methods Accid. Res. 1, 72–85.