Accident Analysis and Prevention 129 (2019) 230–240

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Using latent class analysis and mixed logit model to explore risk factors on driver injury severity in single-vehicle crashes

Zhenning Li[a], Qiong Wu[a], Yusheng Ci[b], Cong Chen[c], Xiaofeng Chen[d], Guohui Zhang[a,*]

[a] Department of Civil and Environmental Engineering, University of Hawaii at Manoa, 2540 Dole Street, Honolulu, HI 96822, USA
[b] Department of Transportation Science and Engineering, Harbin Institute of Technology, 73 Huanghe Road, Harbin, Heilongjiang 150090, China
[c] Center for Urban Transportation Research, University of South Florida, 4202 East Fowler Avenue, CUT100, Tampa, FL 33620, USA
[d] School of Automation, Northwestern Polytechnical University, Xi'an, 710129, China

ARTICLE INFO

Keywords:
Unobserved heterogeneity
Latent class analysis
Mixed logit model
Driver injury severity

ABSTRACT

The single-vehicle crash has been recognized as a critical crash type due to its high fatality rate. In this study, a two-year crash dataset including all single-vehicle crashes in New Mexico is adopted to analyze the impact of contributing factors on driver injury severity. In order to capture the across-class heterogeneous effects, a latent class approach is designed to classify the whole dataset by maximizing the homogeneous effects within each cluster. The mixed logit model is subsequently developed on each cluster to account for the within-class unobserved heterogeneity and to further analyze the dataset. According to the estimation results, several variables including *overturn, fixed object*, and *snowing*, are found to be normally distributed in the observations in the overall sample, indicating there exist some heterogeneous effects in the dataset. Some fixed parameters, including *rural*, *wet*, *overtaking*, *seatbelt used*, *65 years old or older*, etc., are also found to significantly influence driver injury severity. This study provides an insightful understanding of the impacts of these variables on driver injury severity in single-vehicle crashes, and a beneficial reference for developing effective countermeasures and strategies for mitigating driver injury severity.

1. Introduction

The single-vehicle crash has been identified as a major type of crashes due to its high fatality rate. In the United States, 54% of motor vehicle crash deaths in 2017 occurred in single-vehicle crashes, although single-vehicle accounted for only about 30% of all traffic accidents (National Highway Traffic Safety Administration, 2018). In view of the massive casualties caused by single-vehicle crashes, considerable research has been conducted to investigate the frequency and injury severity patterns and their influencing factors in single-vehicle crashes (Behnood and Mannering, 2015, 2017b; Chen et al., 2016b; Feng et al., 2016; Jung et al., 2010; Li et al., 2018a; Wu et al., 2016a,b; Xie et al., 2012; Yau, 2004; Zeng et al., 2019). For instance, Li et al. (2018a) developed a random effects hierarchical Bayesian approach to examine the cross-level interactions between crash-vehicle and driver levels in a single-vehicle crash dataset, and provided detailed discussions on the impacts of significant variables. Chen et al. (2016a) adopted a support vector machine models to examine the contributing factors and their impacts on driver injury severity of single-vehicle overturn crashes.

Considering the nature of the conventionally used dataset for traffic crash analysis, the unobserved heterogeneity, resulting from unobservable contributing factors and data, has been recognized as a critical issue in traffic safety modeling (Amoh-Gyimah et al., 2017; Anastasopoulos et al., 2016; Li et al., 2019a; Mannering et al., 2016; Mannering and Bhat, 2014; Yasmin and Eluru, 2018). Many previous studies have already proved that ignoring potential unobserved heterogeneity may introduce biased estimation and erroneous prediction (Mannering and Bhat, 2014). In addition, driver injury severities in such dataset are often modeled as discrete severity outcomes (for instance, fatal injury, incapacitating injury, visible injury, complaint of injury or possible injury, and no apparent injury), once the crash is observed. Therefore, discrete choice models capable of interpreting unobserved heterogeneity are suitable for analyzing the normally collected crash datasets. Of all the approaches that can meet the aforementioned requirements, the mixed logit model has been widely adopted in previous studies (Barua et al., 2016; Behnood and Mannering, 2017a, 2016; Bhat et al., 2017; Chen et al., 2016a; Chen and Tarko, 2014; Coruh et al., 2015; Heydari et al., 2018; Kim et al., 2010; Milton et al., 2008; Russo et al., 2014; Seraneeprakarn et al., 2017; Ye and Lord, 2014; Zeng et al., 2017).This approach allows the

[*] Corresponding author.
E-mail address: guohui@hawaii.edu (G. Zhang).

https://doi.org/10.1016/j.aap.2019.04.001
Received 26 March 2018; Received in revised form 14 February 2019; Accepted 1 April 2019
Available online 05 June 2019

parameter to be randomly distributed across observations by assuming some pre-defined continuous distributions for the parameter (Li et al., 2018b). For example, Milton et al. (2008) applied a mixed logit model to investigate injury severity patterns in highway crashes. Gkritza and Mannering (2008) utilized mixed logit models to examine the heterogeneous impacts of roadway, vehicle and driver characteristics on seatbelt usage rate. Kim et al. (2010) evaluated pedestrian injury severity in pedestrian-vehicle crashes through a mixed logit model by simulating unobserved pedestrian heterogeneity regarding health, strength, and behavior. They also verified the heterogeneous effects of age and gender on driver injury severity in single-vehicle crashes with a mixed logit model analysis (Kim et al., 2013).

However, this approach also has its own drawbacks. For instance, the pre-defined parameter distributions in mixed logit models may not always hold for all observations in the dataset. To overcome this drawback, one may consider classifying the dataset by separating the entire dataset into different sub-datasets where the heterogeneity effects among these sub-datasets and the homogeneous of observations within each sub-dataset are both maximized (Depaire et al., 2008; Sasidharan et al., 2015). Cluster analysis has the advantage of partitioning sub-datasets without any beforehand partition criteria (Fernandes and Neves, 2013), and has been widely used in traffic crash data analysis as an assistive method in the first step with different frameworks, including K-means cluster (Ahmad and Dey, 2007; Feng et al., 2016; Yamashita, 2005), kernel-density function (Bíl et al., 2013; Prasannakumar et al., 2011), latent class analysis (de Oña et al., 2013; Depaire et al., 2008), network-based model (de Oña et al., 2013; Prato et al., 2012), geospatial and temporal statistical aggregation (Prasannakumar et al., 2011), etc. For instance, Li et al. (2013) identified six spatial-temporal clusters of drunk-driving based on location types and time slots for the drunk-driving pattern investigation. Chen et al. (2016c) utilized a K-means cluster analysis to identify sleeping patterns based on a naturalistic driving study. Wong et al. (2004) developed a hybrid approach based on cluster and autoregression analyses to evaluate the performance of implemented road safety strategies. Palamara et al. (2011) developed a hierarchical method based on a self-organizing map system and K-means cluster algorithm for occupation-based accident data.

Different with the standard cluster analysis techniques (e.g., K-means cluster analysis), the latent class analysis is a model-based clustering approach that derives clusters using a probabilistic model that both simple and complicated distribution forms can be used for the observed variables within clusters (Heydari et al., 2017; Li et al., 2018b; Mathew et al., 2014; Shaheed and Gkritza, 2014; Yasmin et al., 2014). Therefore, the latent class analysis approach allows the analysts to use a model to describe the distribution of the data, rather than selecting clusters with some arbitrarily chosen distance metrics by some standard cluster analysis techniques (Uebersax and Grove, 1990). Based on this model, it is easy to assess the probabilities that certain observations are members of certain latent classes. This feature also allows the latent class approach to have more ability to capture the between-class unobserved heterogeneity (Yu et al., 2017). Another advantage of the latent class approach is that it is not essential to make decisions about the scaling of the observed variables, which is always a critical issue for the standard cluster methods. In other words, the assumed distribution for the observations, in the latent class approach, will produce the same results irrespective of whether the variables are normalized (Behnood et al., 2014). In addition, there are also more formal criteria to make decisions about the number of clusters and other model performance features (Mannering and Bhat, 2014). The interested reader is referred to the article by Magidson and Vermunt (2002), and the references cited therein.

In order to identify driver injury severity patterns in single-vehicle crashes and comprehensively examine the contributing factors on driver injury severity outcomes, a two-step study using both latent class analysis and mixed logit models is developed in this research. The rest of the paper is organized as follows: the utilized dataset is introduced in the next section, Section 3 describes the detailed methodology design, research results and discussions are presented in Section 4, and the research effort is finally concluded in Section 5.

2. Data

The New Mexico single-vehicle crash dataset from 2010 to 2011, obtained from New Mexico Department of Transportation (NMDOT), is utilized in this study. The entire dataset consists of three sub-datasets, including the crash dataset, the vehicle dataset, and the driver dataset. The crash dataset documents crash-level information regarding collision types, crash time and location, road geometric and weather conditions. The vehicle dataset illustrates detailed characteristics of each vehicle, occupant injury outcomes, and the vehicle-specific traffic control information. The driver dataset demonstrates the demographic and behavior information of each driver involved in crashes. Five injury severity levels were originally defined in the dataset, including no injury, possible injury, visible injury, incapacitating injury and fatality. Due to the limited number of fatality records, the final two categories, incapacitating injury and fatality, are grouped into one category in this study. Finally, the severity levels were classified into four levels: *N* (original category: no injury), *P* (original category: possible injury), *I* (original category: visible injury), and *F* (original categories: incapacitating injury and fatality). In addition, the first injury severity level, N, is selected as the reference category. In order to facilitate the modeling process and better illustrate the influence of heterogeneous factors on driver injury severity outcomes, numeric variables are categorized accordingly based on previous traffic safety research and engineering experience (Chen et al., 2015b; Wu et al., 2016b, 2014). The detailed information of the studied dataset is shown in Table 1.

3. Methodology

3.1. Latent class model

In light of the article of Linzer and Lewis (2011), a latent class model is first conducted to group observations into several clusters. Assuming that driver injury severities have J levels (in this study, $j \in J$, indicating N, I, P, F severities, respectively), the conditional probability of the i th driver having the j th injury severity classified in r th ($r \in R$) latent class specified by an MNL model can be given by

$$\mathrm{Prob}(y_{ij|k}) = \frac{\exp(\boldsymbol{\beta}_{jr}^{T}\boldsymbol{X}_{ij} + \varepsilon_{ij|r})}{\sum_{j=1}^{J}\exp(\boldsymbol{\beta}_{jr}^{T}\boldsymbol{X}_{ij} + \varepsilon_{ij|r})} \tag{1}$$

where $\boldsymbol{X}_{ij}$ is the vector of explanatory variables, $\boldsymbol{\beta}_{jr}^{T}$ is the specific vector of parameters for r th ($r \in R$) class, and $\varepsilon_{ij|r}$ is error term indicating the unobserved heterogeneity. The class probability for the i th driver in the r th ($r \in R$) latent class can be given by

$$\boldsymbol{\pi}_{ir} = \frac{\exp(\theta_r^{T}\boldsymbol{z}_i)}{\sum_{r=1}^{R}\exp(\theta_r^{T}\boldsymbol{z}_i)} \tag{2}$$

where $\boldsymbol{z}_\mathrm{i}$ is a vector demonstrating the homogeneity among different individuals that resides in class r, and Θ_r is the specific vector for parameters accounting for the homogeneity within class r.

According to the Bayes' theorem, the unconditional probability of ith driver getting involved in the j th injury severity is given by

$$\mathrm{Prob}(y_{ij}) = \sum_{r=1}^{R}\boldsymbol{\pi}_{ir}\times\mathrm{Prob}(y_{ij|k}) \tag{3}$$

In this study, we begin by fitting the model with $R = 1$ and iteratively increase the value R by one until a suitable model fit is obtained. Bayesian Information Criterion (BIC) is usually used for model selection due to its conciseness (de Oña et al., 2013; Haughton et al., 2009; Li

Table 1
Descriptive Statistics of Studied Dataset.

Driver Injury Severity	N	P	I	F	All
	73.1%	11.7%	10.2%	5.0%	11429
Crash-Level Variables					
Weekday					
Sunday	69.4%	12.2%	12.0%	6.3%	1644
Monday	74.0%	11.9%	9.4%	4.7%	1624
Tuesday	74.4%	11.6%	9.3%	4.8%	1599
Wednesday	76.1%	10.5%	8.8%	4.6%	1473
Thursday	74.7%	11.3%	9.9%	4.1%	1630
Friday	72.6%	11.9%	10.4%	5.2%	1680
Saturday	71.1%	12.3%	11.3%	5.3%	1779
Intersection Related					
Intersection Related	75.0%	11.5%	9.4%	4.2%	96
Not Intersection Related	73.1%	11.7%	10.2%	5.0%	11333
Collision Type					
Overturn	47.1%	18.3%	22.1%	12.4%	2590
Other Non-collision	86.9%	6.6%	4.6%	2.0%	549
Railroad Train	60.0%	20.0%	0.0%	20.0%	5
Animal	92.8%	4.4%	2.3%	0.5%	1985
Fixed Object	75.6%	12.0%	8.7%	3.8%	5948
Other Object	90.6%	6.8%	1.7%	0.9%	352
Lighting Condition					
Daylight	71.2%	12.7%	11.1%	5.0%	5927
Dawn/Dusk	78.8%	9.4%	7.9%	4.0%	706
Dark	74.7%	10.8%	9.5%	5.1%	4796
Road Curvature					
Curved Road	68.7%	13.0%	12.4%	6.0%	2614
Straight Road	74.4%	11.3%	9.6%	4.7%	8815
Road Grade					
Level	73.0%	11.6%	10.3%	5.1%	8607
Hillcrest	74.1%	12.6%	9.1%	4.1%	657
On Grade	73.0%	11.9%	10.2%	5.0%	2050
Dip or Sag	75.7%	9.6%	10.4%	4.4%	115
Weather					
Clear	71.8%	11.8%	11.0%	5.4%	9463
Raining	77.3%	11.0%	7.3%	4.4%	520
Snowing	81.7%	11.2%	5.1%	2.0%	1176
Fog	72.0%	10.0%	14.0%	4.0%	50
Dust	64.3%	21.4%	7.1%	7.1%	14
Wind	76.2%	12.1%	7.8%	3.9%	206
Road System					
Rural	71.7%	11.6%	10.7%	6.1%	6304
Urban	74.9%	11.8%	9.6%	3.7%	5125
Crash Season					
Spring	70.6%	12.2%	11.6%	5.5%	2526
Summer	71.8%	11.3%	11.6%	5.3%	2724
Fall	72.3%	11.5%	10.7%	5.5%	2630
Winter	76.4%	11.8%	7.8%	4.0%	3549
Crash Hour					
Daytime (11 am – 4 pm)	69.6%	13.3%	11.4%	5.7%	2869
Night (9 pm–6 am)	73.6%	11.1%	10.3%	5.0%	4135
Peak (7 am–10 am/5 pm–8 pm)	75.0%	11.2%	9.3%	4.5%	4425
Vehicle-Level Variables					
Road Pavement					
Road Paved	73.3%	11.7%	10.1%	4.9%	10739
Road not Paved	69.7%	12.2%	11.6%	6.5%	690
Road Surface Condition					
Dry	71.4%	11.8%	11.2%	5.6%	8766
Wet	75.5%	12.0%	8.6%	4.0%	901
Snow	82.1%	10.6%	5.0%	2.3%	705
Ice	81.1%	11.3%	5.2%	2.4%	826
Loose	68.8%	11.5%	15.9%	3.8%	157
Water	73.7%	15.8%	5.3%	5.3%	19
Slush	82.1%	10.3%	5.1%	2.6%	39
Other Conditions	75.0%	12.5%	0.0%	12.5%	16
Traffic Control					
No Passing Zone	69.6%	12.4%	12.5%	5.6%	1458
Stop Sign	73.9%	11.7%	10.8%	3.6%	418
Traffic Signals	77.1%	12.5%	6.6%	3.9%	488
Railroad Gate	95.7%	0.0%	0.0%	4.4%	23
Yield Sign	84.9%	3.8%	9.4%	1.9%	53
No Controls	73.3%	11.6%	10.0%	5.1%	8989
Number of Lanes					
One Lane	72.5%	11.7%	10.8%	5.0%	4692
Two Lanes	74.1%	10.6%	10.1%	5.2%	5141
Three or More Lanes	71.9%	14.9%	8.8%	4.5%	1596
Vehicle Type					
Passenger Car	73.0%	12.7%	9.7%	4.6%	5494
Pickup	74.4%	10.2%	10.1%	5.3%	2954
Truck	77.2%	8.5%	9.8%	4.5%	530
Bus	91.2%	5.9%	0.0%	2.9%	34
Van	70.7%	11.9%	11.7%	5.8%	2417
Driver-Level Variables					
Driver Action					
Straight	72.4%	12.0%	10.4%	5.3%	10024
Overtaking	56.6%	13.2%	19.4%	10.9%	129
Right Turn	79.1%	9.4%	8.4%	3.0%	498
Left Turn	79.2%	10.2%	8.4%	2.2%	557
U-turn	94.3%	0.0%	5.7%	0.0%	35
Slowing	77.4%	11.9%	8.3%	2.4%	84
Backing	94.1%	0.0%	2.9%	2.9%	102
Driver Residency					
Non-New Mexico driver	74.3%	9.0%	11.5%	5.3%	2068
New Mexico driver	72.8%	12.3%	9.9%	4.9%	9361
Driver Seatbelt Use					
Seatbelt Used	74.5%	11.7%	9.9%	3.9%	11080
Seatbelt not Used	27.8%	10.6%	21.5%	40.1%	349
Driver Age					
16–20 Years Old	69.0%	14.0%	12.0%	5.0%	2104
21–34 Years Old	72.6%	11.5%	10.7%	5.2%	4076
35–44 Years Old	74.9%	12.3%	7.9%	5.0%	1811
45–54 Years Old	76.0%	10.5%	8.7%	4.8%	1582
55–64 Years Old	76.0%	10.7%	9.4%	4.0%	1099
65 Years Old or Older	72.4%	9.0%	12.4%	6.2%	757
Driver Under Influence					
Driver Under Influence	58.6%	11.8%	17.7%	11.9%	1440
Driver not Under Influence	75.2%	11.7%	9.1%	4.0%	9989
Driver Gender					
Male	75.6%	9.4%	10.0%	5.0%	7162
Female	68.9%	15.6%	10.5%	5.0%	4267

et al., 2019b), and is defined as:

$$BIC = -2\Lambda + \Phi \times \ln N \tag{4}$$

where Λ is the maximum log-likelihood of the model, Φ represents the total number of estimated parameters, and N is the number of observations in the studied dataset. In general, lower BIC value indicates a better model fit on the analyzed dataset.

3.2. Mixed logit model

For each sub-dataset generated by the latent class analysis, the mixed logit model is utilized to examine the contributing factors and assess their impacts on driver injury severity. In light of previous studies (Chen et al., 2015a; Wu et al., 2016c), the model is given as follows:

$$P_{ij} = \frac{\exp[\boldsymbol{\beta_j} \bullet \boldsymbol{X_{ij}}]}{\sum_{\forall I} \exp[\boldsymbol{\beta_n} \bullet \boldsymbol{X_{ij}}]} \tag{5}$$

where P_{ij} is the probability of the ith driver having jth severity level, $\boldsymbol{\beta_j}$ is a vector of parameters to be estimated for driver injury severity level n which may vary across observations, $\boldsymbol{X_{ij}}$ is a vector of explanatory variables. In order to allow the model to account for parameter variations across individual drivers, a mixed distribution is introduced giving driver injury severity probabilities (Train, 2009):

$$P_{ij}|\varphi = \int \frac{\exp[\boldsymbol{\beta_j} \bullet \boldsymbol{X_{ij}}]}{\sum_{\forall I} \exp[\boldsymbol{\beta_j} \bullet \boldsymbol{X_{ij}}]} f(\beta\,|\varphi)d\beta \tag{6}$$

where β is the driver-specific variations of the variables, $f(\beta|\varphi)$ represents the probability density function of β, and φ denotes a vector of parameters describing the probability density function (mean and variance). Considering the computational cost-efficiency, a simulation-based maximum likelihood estimation method is employed for model estimation. Simulation with 1000 Halton draws was conducted in this study. The normal distribution is selected for the parameter probability density function form due to its continuity and capability to describe the central tendency and variation of random variables (Wu et al., 2014).

3.3. Pseudo-elasticity analysis

Numerous studies have concluded that when assuming driver injury severities follow a multinomial distribution, the signs of the estimated parameter may not able accurately demonstrate the real impact of the variable on driver injury outcomes (Kim et al., 2013; Wu et al., 2014). Considering the binary form (with 0/1 outcome) of variables, direct pseudo-elasticity was proposed to assess the influences of statistically significant variables. The pseudo-elasticity, $E_{(p)}{}^{P_{kn}}_{X_{kni}}$, is defined as the percentage change in probability when an indicator variable is switched (i.e., from 0 to 1 or from 1 to 0), and can be given as (Li et al., 2019b)

$$E_{(p)}{}^{P_{ij}}_{X_{ijm}} = \frac{P_{ij}\,[given\ X_{ijm} = 1] - P_{ij}\,[given\ X_{ijm} = 0]}{P_{ij}\,[given\ X_{ijm} = 0]} \tag{7}$$

where X_{ijm} is the value of the m th variable for the i th crash in the propensity function with respect to the j th injury severity level, P_{ij} is the probability the driver of the i th crash having an injury severity level j for the given value of the variable X_{ijm} while holding other variables constant.

4. Results analysis and discussions

4.1. Latent class analysis results

The latent class analysis was conducted through the package “poLCA" in the R environment (Linzer and Lewis, 2011). In order to obtain the optimal number of clusters, different numbers of clusters from 1 to 10 were separately tested. As shown in Fig. 1, the BIC value reaches its minimum when the number of clusters is seven, indicating that classifying the original dataset into seven clusters could produce the best performance. Consequently, the seven-latent-class model is selected as the final model for further mixed logit analysis.

Similar to the work by Depaire et al. (2008), we focus on the differences between the various clusters and are particularly interested to see if the cluster models reveal new information. In order to conserve space, coefficients statistically insignificant at a 10% confidence level were omitted from the table. Variables which have significantly different percentages in the specific cluster with other clusters are set to be bold in Table 2. As illustrated in Table 3, each cluster is described and named by the representative variables.

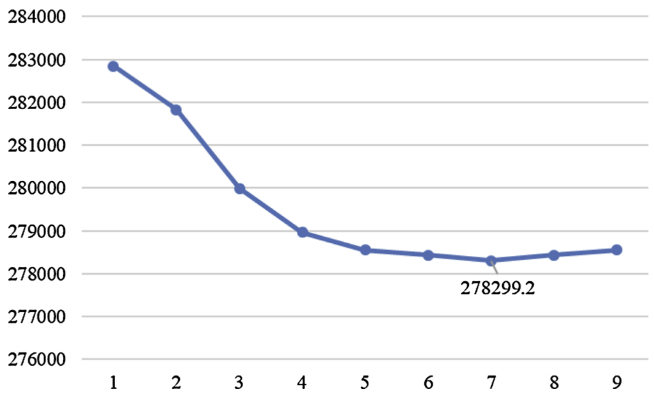

Fig. 1. BIC values for different number of clusters.

4.2. Mixed logit models results

All the eight mixed logit models were developed in this study using the NLOGIT software. The estimation results and pseudo-elasticity results of all statistically significant parameters ($P < 0.05$) are shown in Tables 4 and 5, respectively. The variables, *overturn*, *fixed object*, and *snowing*, are found to have random effects across the observations in the overall sample. Other models are degraded to MNL models because no random parameters are found in the model estimation processes. The detailed impacts of different variables on driver injury severity, especially on F injury severity, are discussed in the following sections.

As illustrated in Table 4, the parameter of the variable, *overturn*, for the overall sample model, specific to I injury severity, is found to follow a normal distribution. Although the mean of it is not significantly different from zero, we consider this parameter to be a random parameter because the significant standard deviation implies that the parameter is not fixed and has unobserved heterogeneity across observations. This finding is understandable, as many observed and unobserved factors, such as vehicle type and model year, speed, safety-feature indicators, driver height, driver weight, etc., have different impacts on the outcome of an overturn crash. These factors are highly aggregated in the overall sample, therefore resulting in more variations across different drivers. The variable is found to only have fixed effects in all other clusters, indicating the latent class model is able to explicitly eliminate the corresponding heterogeneity for this variable. In addition, the variable is also found to significantly increase the likelihood of F injury level for the drivers in the overall sample, Clusters 1, 3, 5, and 6, respectively, based on the pseudo-elasticity estimation results in Table 5.

The variable, *fixed object*, is found to be normally distributed in the observations in overall sample specific to F injury level, and in the observations in Cluster 6 specific to I injury level. The results indicate that this variable has some unobserved heterogeneity across the whole dataset and the observations in Cluster 6, and its effects on injury severity are not always constant. Similar to the overturn crash, many unobserved factors, such as human elements, vehicle characteristics, roadway characteristics, etc., can impact the overcome of this crash type. This finding once again demonstrates the necessaries of capturing for unobserved heterogeneity. The use of the latent class model enables the conventional mixed logit models to address heterogeneous effects in sub-datasets rather than the overall sample only. The pseudo-elasticity results show that this variable can decrease the possibilities of the driver in the overall sample, Clusters 1, 5, and 7, suffering F level injury by −13.2%, −27.5%, −18.1%, and −18.9%, respectively. In addition, the variable, *animal* is found to decrease the likelihood of drivers in the overall sample and Cluster 7 suffering F level injuries by −44.8% and −33.2%, respectively.

As illustrated in Table 5, with respect to *clear*, some of the adverse weather-related variables, including *raining*, *snowing*, and *wind*, are found to have significant influences on driver injury severities for the drivers in certain clusters or the overall sample. *Raining* is expected to decrease the possibilities of F level injuries for drivers in Cluster 7 by −11.9%. This result is understandable since the drivers always tend to drive slower than usual when driving at raining days, and it is in line with some previous studies (Jung et al., 2010). *Snowing* is also found to have favorable impacts on driver injury severities for the drivers in the overall sample, Clusters 1 and 6. In addition, this variable is found to be normally distributed across the observations in the overall sample specific to I level injury crashes. As figured out in the article of Mannering et al. (2016), due to different driver responses to snow, i.e., the extent to which they adjust driving speeds, the same amount of snow may have different effects in different geographical areas, which will affect the severity of individual injuries as well. The pseudo-elasticity results show that this variable has favorable impacts on severe injuries since it can reduce the likelihood of the drivers in the overall sample, clusters 1 and 5, being seriously injured by −22.3%, −6.1%, and −11.4%, respectively. It should be noted that the influence

Table 2
Summary of Variables Describing Cluster Characteristics.

Variable	OS	C1	C2	C3	C4	C5	C6	C7
Animal	17.4%	1.4%	**93.7%**	0.0%	0.0%	0.0%	5.4%	0.0%
Fixed Object	52.0%	38.3%	0.0%	66.6%	**91.9%**	58.9%	56.7%	55.7%
Daylight	51.9%	69.8%	20.9%	23.2%	61.9%	69.5%	59.3%	46.3%
Dark	42.0%	23.1%	**65.9%**	**69.1%**	30.7%	23.0%	31.8%	42.8%
Curved Road	22.9%	15.7%	7.2%	26.4%	16.0%	9.1%	**75.1%**	11.3%
Snow	6.2%	5.4%	0.4%	0.9%	4.9%	1.6%	9.0%	**34.2%**
Ice	7.2%	8.1%	0.1%	0.7%	5.8%	0.3%	8.1%	**46.8%**
Right Turn	4.4%	5.0%	0.1%	3.1%	**24.2%**	0.0%	4.5%	0.8%
Left Turn	4.9%	3.8%	0.0%	5.7%	**24.7%**	0.4%	5.8%	0.0%
No Passing Zone	12.8%	6.3%	13.8%	14.2%	0.6%	4.0%	**42.1%**	4.1%
No Controls	78.7%	84.6%	81.8%	72.9%	45.4%	92.0%	53.1%	91.5%
Pickup	25.9%	**53.5%**	27.0%	27.8%	11.2%	10.3%	22.1%	41.4%
Truck	4.6%	**29.8%**	3.4%	0.0%	5.0%	0.0%	0.0%	0.1%
Driver Under Influence	12.6%	1.9%	0.2%	**63.5%**	13.3%	2.3%	6.4%	0.0%
Male	62.7%	**92.4%**	57.6%	72.3%	59.0%	44.8%	54.5%	63.0%

OS = overall sample, C1 to C7 for Cluster 1 to Cluster 7.
The level of significance is p = 0.01 for the bold values.

mechanism of adverse weather conditions is still a controversial topic, but our results show that the influence is more stable within each cluster. Some researchers suggested that the reasons of adverse weather conditions having favorable impacts on driver injury severity because that the drivers tend to be more cautious and drive slowly and thus the subsequent crashes result in less severe injuries, which is known as risk compensation (Eluru et al., 2012).

The variable, *rural*, is expected to significantly impact injury severities for divers in certain clusters. This variable is found to increase the likelihood of F level injury for the driver in Clusters 1, 3 and 6, respectively (pseudo-elasticity result equals 10.2%, 12.5%, 11.7%, respectively). Previous studies also figured out that driver fatality rates in most rural counties are much higher than what they are in urban counties (Schwab, 2009; Wu et al., 2016b). The reasons are comprehensive, for instance, signages and road markings in rural areas are not common like that in urban areas. In addition, due to lower traffic density and less traffic control facilities, vehicles in rural areas are more likely to be speeding.

The action, *overtaking*, shows negative impacts on driver injury severities. For drivers in the overall sample and Cluster 7, this variable can increase the possibilities of them being seriously injured (I and F levels). Because the action requires more operations than going straight, the drivers in Cluster 7 are more difficult to control their vehicles on a snow/ice roadway without traffic control, resulting in more serious injuries. This finding is also in line with previous studies that this action is always associated with severe injuries (Li et al., 2018a; Pai, 2011).

This variable, *seatbelt used*, is expected to have favorable influences on driver injury severities. When the seatbelt is in used, the possibilities of drivers being severely injured (I or F level) are dramatically reduced, especially for the drivers in clusters with more severe injuries, i.e., Clusters 3, 5, and 6. Regarding the significant impacts, these results suggest that a stricter regulation or law is required for increasing the proportion of driving with the seatbelt on, since the overall ratio of using the seatbelt is only around 75%. This finding is also evidenced by other studies on driver injury severities (Lee and Li, 2014; Wang and Abdel-Aty, 2008).

As shown in Table 5, the variable, *driver age*, shows quite diverse impacts on driver injury severities for drivers in different clusters. For instance, in Cluster 1, drivers in 16–20 years are more likely to suffer serious injuries (14.4% and 9.7% for I and F level injuries, respectively), while the drivers in Cluster 5 with the same age tend to have less serious injuries (−4.4% and −4.7% for I and F level injuries, respectively). However, the old drivers (65 years old or older) are always associated with severe injuries in many clusters and the overall sample. As widely discussed in previous studies, the primary reasons for these results are due to the old drivers' chronic medical conditions and functional impairments (Li et al., 2018a). Their acute manifestations of chronic conditions and specific medical diagnoses are found often associated with impairment of skills necessary for successful motor vehicle operations and therefore lead to severe crashes. In addition, their inferior driving performance may also be impacted by their functional impairments, e.g., vision, cognition, mobility, and so on (Chen et al., 2015a).

This variable, *driver under influence*, is used to describe drivers being influenced by alcohol or drug while driving. It is not surprising that this variable can aggravate driver injury severities, and it is found to dramatically increase the possibilities of drivers in the overall sample, Cluster 4 and Cluster 6 being serious injured (I and F levels). Previous studies also concluded similar findings that the use of alcohol and drug will significantly influence the driver's state of consciousness (Behnood et al., 2014). This result provides convincible evidence for educating drivers to keep away from alcohol and drug usage before driving.

Compared to male drivers, female drivers seem to be more likely to suffer serious injuries. The variable, *female*, is found to increase the likelihood of I or F level injuries for drivers in the overall sample, Clusters 4, 5, 6, and 7. Note that there is not a very solid conclusion of the impacts of driver gender on driver injury severities, as some

Table 3
Cluster Definition and Description.

Cluster	Description	Abbreviation	Percentage
Cluster 1	Male driver heavy vehicle crashes	ManHeavy (MH)	11.1%
Cluster 2	Animal-vehicle crashes under dark conditions	AnimalDark (AD)	17.1%
Cluster 3	Impaired driving crashes under dark conditions	ImpairedDark (ID)	13.3%
Cluster 4	Hitting fixed objects while turning	FixedTurning (FT)	8.8%
Cluster 5	All others	Other (OT)	25.9%
Cluster 6	Crashes on curved roads with no passing zone sign	CurvedNopassing (CN)	13.8%
Cluster 7	Crashes on snow/ice surfaces	SnowIce (SI)	10.0%

previous studies also reported that male drivers are more likely to have serious injury outcomes (Li et al., 2018a). Some scholars argued that male drivers are more common to be speeding or involved in alcohol or drug-impaired driving, and therefore, male drivers are always associated with severe injury outcomes (Gray et al., 2008). While others inferred that male drivers have more driving experiences and are much calm down than female drivers when crashes occur, thus male drivers have less serious injuries. Regarding these inconsistent conclusions, more research is needed to figure out the impacts of the driver gender on injury severity (Kim et al., 2013).

4.3. Model comparison and evaluation results

Other than just identifying contributing factors for single-vehicle crashes, this study also aims to evaluate the performance of cluster-based data segmentation in injury severity analysis. Therefore, in this section, more efforts are focused on the discussions about the differences between the model developed for the overall sample and those for the seven clusters.

As shown in Table 4, 16 factors are found to significantly influence driver injuries only in the models for the seven clusters. These results provide a more comprehensive understanding of single-vehicle crashes

Table 4
Estimation Results for All Models.

Models		OS	C1 (MH)	C2 (AD)	C3 (ID)	C4 (FT)	C5 (OT)	C6 (CN)	C7 (SI)
Variable	SEV	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Constant	P	2.17*	3.22*	1.45*	2.20	1.97*	2.05	1.99	3.25*
	I	1.05*	2.20*	1.22	0.75	1.43*	0.52*	−0.46*	1.47*
	F	1.07*	1.24	0.66	0.87	0.73*	−1.52*	−0.68	−0.99
Collision Type (with respect to *Other Non-collision*)									
	I	0.83	1.92*		1.57*		3.30*	2.45*	
	SD	2.99*							
	F	2.52*	1.31*		1.96*		1.55*	1.24*	
Animal	P	−0.52*							
	F	−1.47*							−0.87*
Fixed Object	P	0.99*						0.69*	−0.55*
	I	1.36*	1.45*				1.70*	−1.22*	−1.33*
	SD							2.70*	
	F	−2.52*							
	SD	2.31*							
Lighting Condition (with respect to *Daylight*)									
Dusk	P			−1.42*					
Road Grade (with respect to *Level*)									
Hillcrest	P						0.77*		
Weather (with respect to *Clear*)									
Raining	I								1.30*
Snowing	P	−0.83*						−0.55*	
	I	−3.07*	−1.72*						
	SD	2.08*							
	F	−1.55*							
Wind	P		−1.23*						
Road System (with respect to *Urban*)									
Rural	P		1.44*						
	F				0.86*			0.92*	
Crash Hour (with respect to *Daytime*)									
Peak	I				−0.49*	−0.54*			
Road Surface Condition (with respect to *Dry*)									
Wet	F		1.23*						
Ice	I							−0.99*	
Traffic Control (with respect to *Traffic Signals*)									
No Passing Zone	P				0.54*				
	I	0.45*	0.80*						
Stop Sign	F				−1.26*				
Number of Lanes (with respect to *Two Lanes*)									
One Lane	P							−0.42*	
Three and More Lanes	P	0.33*							
	F						0.44*		
Vehicle Type (with respect to *Passenger Car*)									
Pickup	P			−0.42*					
	I			−1.10*					
Van	I					1.05*			
Driver Action (with respect to *Straight*)									
Overtaking	I	1.05*							1.29*
	F	1.44*							
Driver Seatbelt Use (with respect to *Seatbelt not Used*)									
Seatbelt Used	I						−0.99*		
	F				−2.25*			−1.79*	
Driver Age (with respect to *35 to 44 Years Old*)									
16 to 20 Years Old	I		0.85*						
	F		0.42*				−0.60*		
21 to 34 Years Old	P					−0.43*			
	F						−0.41*		
45 to 54 Years Old	P						−0.59*		

(*continued on next page*)

Table 4 (*continued*)

Models		OS	C1 (MH)	C2 (AD)	C3 (ID)	C4 (FT)	C5 (OT)	C6 (CN)	C7 (SI)
65 Years Old or Older	P			−1.29*					
	I	0.78*	0.56*			1.44*	0.65*		
	F	0.92*			1.58*				
Driver Under Influence (with respect to *Driver not Under Influence*)									
Driver Under Influence	P				−0.46*				
	I	1.78*				1.42*		1.33*	
	F	2.35*						1.09*	
Driver Gender (with respect to *Male*)									
Female	P	0.59*				0.61*			
	I	0.38*					0.34*		
	F	0.44*				0.87*	0.42*		
Model Performance Results									
Log Likelihood with Constant Only		−13471.70	−1497.68	−2298.96	−1793.45	−1187.77	−3486.73	−1854.72	−1348.03
Log Likelihood at Convergence		−6444.38	−731.21	−443.75	−1100.04	−493.55	−2144.16	−998.54	−532.74
ρ^2		0.52	0.51	0.81	0.39	0.58	0.39	0.46	0.60

OS = overall sample, C1 to C7 for Cluster 1 to Cluster 7.
SEV = Severity.
Coef. = Coefficient.
SD = Standard Deviation.
* Significant at 95% Confidence Level.

and justify the effectiveness of our proposed cluster-based models, because it confirms the assumption that performing traffic accident analysis on a large heterogeneous data set can obscure significant relations (Depaire et al., 2008). For example, *rural* is found to have significant impacts on injury severities for the drivers in Clusters 1, 3 and 6. Pseudo-elasticity results show that if a crash occurs in rural areas, the male heavy-vehicle driver (Cluster 1), the impaired driver driving at dark (Cluster 3), and the driver driving on a curved road with a no passing zone sign (Cluster 6), are more likely to be severely injured or killed. Therefore, specific countermeasures should be developed in rural areas for those kinds of drivers. However, this variable is not significant in the overall sample. If the countermeasures are only proposed based on the estimation results of the overall sample, the rural areas may not attract enough attention compared to those developed according to the estimation results of models for Clusters 1, 3 and 6.

Furthermore, although about 10 variables are found to both have significant impacts on injury outcomes for the drivers in the overall sample and in the sub-sets, their impacts on driver injury severity are not always consistent. For instance, *fixed object* is found to increase the possibilities of I level injuries for the drivers in the overall sample, Clusters 5 and 7, however, decrease the same level of injuries for the male drivers driving heavy vehicle in Cluster 1. Therefore, the latent cluster model reveals the variation of a variable's effect on the injury outcome probability between the overall sample and sub-datasets. In addition, the proposed models even reveal a different direction of the effect for some features in various clusters.

In addition, the model performance results also provide some shreds of evidence on the necessaries of classifying the overall sample into different clusters. As shown in Table 4, most clusters have better or similar goodness-of-fit compared to the overall sample, and only Clusters 3 and 5 have significantly worse performance than the overall sample.

5. Conclusions

In order to identify the impacts of the contributing factors on driver injury outcomes of single-vehicle crashes, a two-year dataset that contains all the single-vehicle crashes in New Mexico is adopted in this study. A latent class approach is conducted to address the across-class unobserved heterogeneity issue in the dataset, and classifies the whole dataset into seven sub-clusters by maximizing the homogeneous effects within each cluster. The mixed logit model is then separately developed for each cluster and the overall sample to capture the within-class unobserved heterogeneity and to further analyze the crash dataset.

According to the estimation results, several variables including *overturn, fixed object*, and *snowing*, are found to be normally distributed in the observations in the overall sample and certain clusters, indicating there exist some heterogeneous effects in the dataset. The superior of the proposed approach is also evidenced by the estimation results. First, classifying the whole dataset into different sub-clusters allows us to find more significant contributing factors which only exist in the sub-clusters. Second, some variables are examined to have different impacts on driver injury severity for the drivers in the overall sample and the sub-clusters, revealing that the influences of variables are not always consistent across all the observations. In addition, the model performance results also demonstrate that the classification could provide better goodness-of-fit and prediction accuracy.

It should be noted that some previous research efforts combined the latent class model and the mixed logit model together to account for the unobserved heterogeneity in the dataset (Li et al., 2018b; Xiong and Mannering, 2013). In their studies, the hybrid approach pre-defined the number of clusters and then allowed random parameters to be randomly distributed between classes and within classes, thus enabling the simulation of more sophisticated heterogeneous effects than conventional discrete choice models (Li et al., 2018b). Due to different model frameworks and different distribution assumptions for random parameters, the hybrid model may provide much fewer clusters than the proposed model. In the articles of Xiong and Mannering (2013), and Li et al. (2018b), the optimal number of latent classes are both two (seven in this study). While in the article of Morgan and Mannering (2011), similar to this study, user-specified classes were used to maximize heterogeneous effects, and a twelve-latent-class model is found to provide the best performance. It is difficult to come to the conclusion that which model is better than another model, because both two models have their own drawbacks. For the hybrid model, the complexity of the model estimation is quite cumbersome. For our proposed model, although increasing the number of clusters could increase the model performance, in the meantime, it may also introduce the over-fit issue. Future researchers may wish to investigate this apparent difference and compare the two approaches in detail.

Specific countermeasures could be implemented based on the estimation results. The use of lighting systems and adequately increasing illumination in the accident-prone areas can improve driving safety since these countermeasures can let the drivers have enough vision distances and available reaction time to recognize dangerous circumstances. Moreover, for the drivers in Cluster 3, who are always

Table 5
Pseudo-Elasticity Test Results for All Models.

Models	SEV	OS	C1 (MH)	C2 (AD)	C3 (ID)	C4 (FT)	C5 (OT)	C6 (CN)	C7 (SI)
Collision Type (with respect to *Other Non-collision*)									
Overturn	N	−20.2%	−25.2%		−30.2%		−20.1%	−19.2%	
	P	22.4%	27.6%		30.2%		10.3%	26.5%	
	I	19.3%	33.1%		29.7%		44.2%	18.3%	
	F	44.5%	40.3%		55.7%		24.6%	33.1%	
Animal	N	56.6%							−24.2%
	P	−13.2%							3.6%
	I	−27.6%							−22.1%
	F	−44.8%							−33.2%
Fixed Object	N	−1.7%	3.9%				2.0%		1.7%
	P	9.2%	12.1%				11.9%		10.6%
	I	6.6%	−7.3%				1.1%		3.7%
	F	−13.2%	−27.5%				−18.1%		−28.9%
Lighting Condition (with respect to *Daylight*)									
Dusk	N			−0.4%					
	P			−9.5%					
	I			10.9%					
	F			−3.4%					
Road Grade (with respect to *Level*)									
Hillcrest	N						−3.0%		
	P						−7.6%		
	I						11.7%		
	F						−1.6%		
Weather (with respect to *Clear*)									
Raining	N								7.5%
	P								3.0%
	I								−14.5%
	F								−11.9%
Snowing	N	32.3%	5.7%					14.5%	
	P	−30.5%	−12.1%					−18.2%	
	I	−49.9%	−8.8%					−22.3%	
	F	−22.3%	−6.1%					−11.4%	
Wind	N		13.1%						
	P		−17.2%						
	I		−20.3%						
	F		−10.6%						
Road System (with respect to *Urban*)									
Rural	N		−30.2%		−22.7%			−25.2%	
	P		22.3%		11.6%			15.1%	
	I		17.4%		10.4%			12.7%	
	F		10.2%		12.5%			11.7%	
Crash Hour (with respect to *Daytime*)									
Peak	N				7.7%	10.5%			
	P				4.3%	2.3%			
	I				−13.2%	−14.0%			
	F				−1.5%	−3.4%			
Road Surface Condition (with respect to *Dry*)									
Wet	N		−5.4%						
	P		2.1%						
	I		2.1%						
	F		3.5%						
Ice	N							11.1%	
	P							2.8%	
	I							−17.1%	
	F							−2.0%	
Traffic Control (with respect to *Traffic Signals*)									
No Passing Zone	N	−3.3%	−8.2%		−11.7%				
	P	3.3%	3.7%		6.2%				
	I	2.2%	3.5%		5.4%				
	F	3.7%	5.0%		7.0%				
Stop Sign	N				26.8%				
	P				−1.4%				
	I				−30.6%				
	F				−18.3%				
Number of Lanes (with respect to *Two Lanes*)									
One Lane	N							−1.7%	
	P							2.2%	
	I							−2.1%	
	F							2.3%	
Three or More Lanes	N	−15.5%					−17.4%		
	P	1.8%					8.7%		
	I	15.9%					18.4%		
	F	11.1%					11.9%		
Vehicle Type (with respect to *Passenger Car*)									

(*continued on next page*)

Table 5 (*continued*)

Models	SEV	OS	C1 (MH)	C2 (AD)	C3 (ID)	C4 (FT)	C5 (OT)	C6 (CN)	C7 (SI)
Pickup	N			33.2%					
	P			0.8%					
	I			−38.4%					
	F			−22.6%					
Van	N					−13.0%			
	P					9.2%			
	I					13.2%			
	F					8.8%			
Driver Action (with respect to *Straight*)									
Overtaking	N	−17.1%							−20.6%
	P	10.9%							12.3%
	I	23.5%							32.3%
	F	16.3%							22.7%
Driver Seatbelt Use (with respect to *Seatbelt not Used*)									
Seatbelt Used	N				9.2%		8.5%	10.9%	
	P				5.6%		3.2%	11.4%	
	I				−13.1%		−11.7%	−17.2%	
	F				−27.2%		−10.3%	−30.6%	
Driver Age (with respect to *35 to 44 Years Old*)									
16 to 20 Years Old	N		−13.5%				2.3%		
	P		9.4%				4.9%		
	I		14.4%				−4.4%		
	F		9.7%				−4.7%		
21 to 34 Years Old	N					7.2%	4.5%		
	P					8.9%	6.7%		
	I					−10.6%	−7.1%		
	F					−17.8%	−10.4%		
45 to 54 Years Old	N						10.4%		
	P						13.3%		
	I						−15.4%		
	F						−25.3%		
65 Years Old or Older	N	−7.9%	−9.2%	−8.4%	−13.7%	−17.4%	−12.5%		
	P	−9.2%	−4.9%	−7.8%	−2.8%	17.7%	15.9%		
	I	11.4%	12.1%	11.6%	17.1%	17.4%	11.8%		
	F	19.9%	17.6%	19.1%	21.8%	7.5%	2.5%		
Driver Under Influence (with respect to *Driver not Under Influence*)									
Driver Under Influence	N	−18.1%			−20.5%	−16.1%		−24.2%	
	P	−20.2%			−23.3%	−1.7%		−27.5%	
	I	25.7%			29.2%	19.6%		34.5%	
	F	45.1%			51.6%	24.6%		61.0%	
Driver Gender (with respect to *Male*)									
Female	N	−8.9%				−12.7%	−10.6%	−12.0%	−10.4%
	P	−4.3%				−7.7%	−5.8%	10.9%	2.7%
	I	12.8%				18.1%	15.1%	12.6%	12.2%
	F	12.3%				20.5%	16.0%	5.6%	11.5%

OS = overall sample, C1 to C7 for Cluster 1 to Cluster 7.
SEV = Severity.
Coe. = Coefficient.

associated with impaired driving behavior, it is necessary for law enforcement to perform regular driver with impairment (DWI) tests on the roadways for both urban and rural areas, especially at dark conditions. In addition, many countermeasures can be further developed to address the impaired driving issue, such as administrative license revocation (ALR), high visibility saturation patrols, passive alcohol sensors, etc. For the rural areas, considering the economic reason and traffic volume characteristics, some low-cost improvements may be implemented to protect the drivers, including using dynamic warning sign advising through traffic that a stopped vehicle is at the intersection and may enter the intersection, extension of the through edge line using short skip pattern to assist drivers to stop at the ideal point in the intersection, solar-powered LED flashing beacons on advance intersection warning and stop signs to provide traffic information ahead the roadways for the drivers, and so on (Hunter et al., 2012). For the old drivers, many countermeasures and strategies could help them to drive safer on the road. Since they always suffer functional impairments, some medical-related treatments (e.g., eyeglasses, vision-related surgery, and so on) and appropriate vehicle adaptations (e.g., extra rear view mirrors, parking sensors, reversing camera, extended gear shift levers, etc.) can definitely assist them in driving vehicles. In addition, formal courses or through communications and outreach provided to the older drivers, could assess their driving capabilities and limitations, and improve their driving skills when possible. However, sometimes, based on the results of regular physical examination for the drivers over a certain age, it is necessary to restrict or revoke driver licenses of the old drivers who are not able to drive safely in certain situations or at all (Goodwin et al., 2011). Besides, although the adverse weather conditions are examined not to aggravate driver injury severity, countermeasures such radiant advisory and regulatory variable speed limits, lighted variable message signs, and real-time information display devices, could significantly benefit driving safety under adverse weather conditions. More generally, considering the variables that could increase the possibilities of severe injuries and fatalities of both the overall model and cluster-based models, strategies like strengthening drivers and occupants seat belt use laws and ordinances, public information supporting enforcement, the use of variable speed limit zones, the use of automated red light cameras and speed cameras for recording vehicles that illegally enter intersections or exceeding a certain speed, etc., could benefit the overall traffic safety.

Although this paper has concluded some insightful findings, there still exist some limitations that may lead to inaccurate estimations and erroneous predictions. For instance, due to the limited occurrences, some variables may only contain a few records (e.g., railroad train, dust, bus, etc.). Estimations on these variables may be biased since there are not enough samples. In addition, though we carefully designed the latent class approach, and randomly selected the initial data partitioning and centroids for several times, the classification results are not constant and may still produce biased estimations. Besides, we only adopted the widely used normal distributions as assumptions for the randomly distributed variables in the mixed logit models. However, using other continuous distributions as assumptions, for example, uniform distribution, lognormal distribution, etc., could be more appropriate to reflect the nature of unobserved heterogeneous effects, and provides more reliable results and findings.

Acknowledgments

This research is jointly supported by the National Natural Science Foundation of China (No. 61803083) and the China Postdoctoral Science Foundation (No. 2018M630497).

References

Ahmad, A., Dey, L., 2007. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl. Eng. 63, 503–527. https://doi.org/10.1016/j.datak.2007.03.016.

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2017. The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. Anal. Methods Accid. Res. 13, 28–51.

Anastasopoulos, P.C., Sarwar, M.T., Shankar, V.N., 2016. Safety-oriented pavement performance thresholds: accounting for unobserved heterogeneity in a multi-objective optimization and goal programming approach. Anal. Methods Accid. Res. 12, 35–47. https://doi.org/10.1016/j.amar.2016.10.001.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Methods Accid. Res. 9, 1–15.

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.

Behnood, A., Mannering, F.L., 2016. An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. Anal. Methods Accid. Res. 12, 1–17. https://doi.org/10.1016/j.amar.2016.07.002.

Behnood, A., Mannering, F., 2017a. Determinants of bicyclist injury severities in bicycle-vehicle crashes: a random parameters approach with heterogeneity in means and variances. Anal. Methods Accid. Res. 16, 35–47.

Behnood, A., Mannering, F.L., 2017b. The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. Traffic Inj. Prev. 18, 456–462.

Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Anal. Methods Accid. Res. 3, 56–91.

Bhat, C.R., Astroza, S., Lavieri, P.S., 2017. A new spatial and flexible multivariate random-coefficients model for the analysis of pedestrian injury counts by severity level. Anal. Methods Accid. Res. 16, 1–22. https://doi.org/10.1016/j.amar.2017.05.001.

Bíl, M., Andrášik, R., Janoška, Z., 2013. Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accid. Anal. Prev. 55, 265–273. https://doi.org/10.1016/j.aap.2013.03.003.

Chen, E., Tarko, A.P., 2014. Modeling safety of highway work zones with random parameters and random effects models. Anal. Methods Accid. Res. 1, 86–95.

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

Chen, C., Zhang, G., Tian, Z., Bogus, S.M., Yang, Y., 2015b. Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations. Accid. Anal. Prev. 85, 186–198.

Chen, C., Zhang, G., Huang, H., Wang, J., Tarefder, R.A., 2016a. Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. Accid. Anal. Prev. 96, 79–87.

Chen, C., Zhang, G., Liu, X.C., Ci, Y., Huang, H., Ma, J., Chen, Y., Guan, H., 2016b. Driver injury severity outcome analysis in rural interstate highway crashes: a two-level Bayesian logistic regression interpretation. Accid. Anal. Prev. 97, 69–78. https://doi.org/10.1016/j.aap.2016.07.031.

Chen, G.X., Fang, Y., Guo, F., Hanowski, R.J., 2016c. The influence of daily sleep patterns of commercial truck drivers on driving performance. Accid. Anal. Prev. 91, 55–63. https://doi.org/10.1016/j.aap.2016.02.027.

Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with aggregated data: the random parameters negative binomial panel count data model. Anal. Methods Accid. Res. 7, 37–49.

de Oña, J., López, G., Mujalli, R., Calvo, F.J., 2013. Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks. Accid. Anal. Prev. 51, 1–10. https://doi.org/10.1016/j.aap.2012.10.016.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40, 1257–1266. https://doi.org/10.1016/j.aap.2008.01.007.

Eluru, N., Bagheri, M., Miranda-Moreno, L.F., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accid. Anal. Prev. 47, 119–127.

Feng, S., Li, Z., Ci, Y., Zhang, G., 2016. Risk factors affecting fatal bus accident severity: their impact on different types of bus drivers. Accid. Anal. Prev. 86, 29–39. https://doi.org/10.1016/j.aap.2015.09.025.

Fernandes, A., Neves, J., 2013. An approach to accidents modeling based on compounds road environments. Accid. Anal. Prev. 53, 39–45. https://doi.org/10.1016/j.aap.2012.12.041.

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. https://doi.org/10.1016/j.aap.2007.07.013.

Goodwin, A.H., Thomas, L.J., Hall, W.L., Tucker, M.E., 2011. Countermeasures That Work: A Highway Safety Countermeasure Guide for State Highway Safety Offices.

Gray, R.C., Quddus, M.A., Evans, A., 2008. Injury severity analysis of accidents involving young male drivers in Great Britain. J. Saf. Res. 39, 483–495.

Haughton, D., Legrand, P., Woolford, S., 2009. Review of three latent class cluster analysis packages: latent gold, poLCA, and MCLUST. Am. Stat. 63, 81–91. https://doi.org/10.1198/tast.2009.0016.

Heydari, S., Fu, L., Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: a joint analysis of pedestrian and cyclist injuries. Anal. Methods Accid. Res. 13, 16–27.

Heydari, S., Fu, L., Thakali, L., Joseph, L., 2018. Benchmarking regions using a heteroskedastic grouped random parameters model with heterogeneity in mean and variance: applications to grade crossing safety analysis. Anal. Methods Accid. Res. 19, 33–48. https://doi.org/10.1016/j.amar.2018.06.003.

Hunter, W., Srinivasan, R., Martell, C., 2012. Evaluation of rectangular rapid flash beacon at Pinellas Trail Crossing in Saint Petersburg, Florida. Transp. Res. Rec.: J. Transp. Res. Board 7–13.

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accid. Anal. Prev. 42, 213–224. https://doi.org/10.1016/j.aap.2009.07.020.

Kim, J.-K., Ulfarsson, G.F., Shankar, V.N., Mannering, F.L., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accid. Anal. Prev. 42, 1751–1758.

Kim, J.-K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in California: a mixed logit analysis of heterogeneity due to age and gender. Accid. Anal. Prev. 50, 1073–1081. https://doi.org/10.1016/j.aap.2012.08.011.

Lee, C., Li, X., 2014. Analysis of injury severity of drivers involved in single-and two-vehicle crashes on highways in Ontario. Accid. Anal. Prev. 71, 286–295.

Li, Y.C., Sze, N.N., Wong, S.C., 2013. Spatial–temporal analysis of drink-driving patterns in Hong Kong. Accid. Anal. Prev. 59, 415–424. https://doi.org/10.1016/j.aap.2013.06.033.

Li, Z., Chen, C., Ci, Y., Zhang, G., Wu, Q., Liu, C., Qian, Z., 2018a. Examining driver injury severity in intersection-related crashes using cluster analysis and hierarchical Bayesian models. Accid. Anal. Prev. 120, 139–151. https://doi.org/10.1016/J.AAP.2018.08.009.

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P.D., Ma, D.T., 2018b. Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. Anal. Methods Accid. Res. 20, 1–14. https://doi.org/10.1016/j.amar.2018.08.001.

Li, Z., Chen, X., Ci, Y., Chen, C., Zhang, G., 2019a. A hierarchical Bayesian spatiotemporal random parameters approach for alcohol/drug impaired-driving crash frequency analysis. Anal. Methods Accid. Res. 21, 44–61. https://doi.org/10.1016/j.amar.2019.01.002.

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian, Z., Prevedouros, P.D., Ma, D.T., 2019b. Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. Accid. Anal. Prev. 124, 219–229. https://doi.org/10.1016/j.aap.2018.12.020.

Linzer, D.A., Lewis, J.B., 2011. poLCA: an r package for polytomous variable latent class analysis. J. Stat. Softw. 42, 1–29.

Magidson, J., Vermunt, J., 2002. Latent class models for clustering: a comparison with K-means. Can. J. Mark. Res. 20, 36–43.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22. https://doi.org/10.1016/j.amar.2013.09.001.

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. https://doi.org/10.1016/j.amar.2016.04.001.

Mathew, D., Gkritza, K., Saad, M., Hans, Z., 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, 11–27. https://doi.org/10.1016/j.amar.2014.09.002.

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40, 260–266.

Morgan, A., Mannering, F.L., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. Accid. Anal. Prev. 43, 1852–1863.

National Highway Traffic Safety Administration (NHTSA), Traffic Safety Facts Annual Report Tables. 2018, https://cdan.nhtsa.gov/tsftables/tsfar.htm#.

Pai, C.W., 2011. Overtaking, rear-end, and door crashes involving bicycles: an empirical investigation. Accid. Anal. Prev. 43, 1228–1235.

Palamara, F., Piglione, F., Piccinini, N., 2011. Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. Saf. Sci. 49, 1215–1230. https://doi.org/10.1016/j.ssci.2011.04.003.

Prasannakumar, V., Vijith, H., Charutha, R., Geetha, N., 2011. Spatio-temporal clustering of road accidents: GIS based analysis and assessment. Procedia - Soc. Behav. Sci. 21, 317–325. https://doi.org/10.1016/j.sbspro.2011.07.020.

Prato, C.G., Gitelman, V., Bekhor, S., 2012. Mapping patterns of pedestrian fatal accidents in Israel. Accid. Anal. Prev. 44, 56–62. https://doi.org/10.1016/j.aap.2010.12.022.

Russo, B.J., Savolainen, P.T., Schneider, W.H., Anastasopoulos, P.C., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. Anal. Methods Accid. Res. 2, 21–29.

Sasidharan, L., Wu, K.-F., Menendez, M., 2015. Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. Accid. Anal. Prev. 85, 219–228.

Schwab, C.V., 2009. Agricultural Equipment on Public Roads.

Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: a random parameters approach with heterogeneity in means and variances. Anal. Methods Accid. Res. 15, 41–55.

Shaheed, M.S., Gkritza, K., 2014. A latent class analysis of single-vehicle motorcycle crash severity outcomes. Anal. Methods Accid. Res. 2, 30–38.

Train, K., 2009. Discrete Choice Methods with Simulation, second ed. Cambridge University Press, Cambridge.

Uebersax, J.S., Grove, W.M., 1990. Latent class analysis of diagnostic agreement. Stat. Med. 9, 559–572.

Wang, X., Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Accid. Anal. Prev. 40, 1674–1682.

Wong, S., Leung, B.S., Loo, B.P., Hung, W., Lo, H.K., 2004. A qualitative assessment methodology for road safety policy strategies. Accid. Anal. Prev. 36, 281–293. https://doi.org/10.1016/S0001-4575(03)00006-X.

Wu, Q., Chen, F., Zhang, G., Liu, X.C., Wang, H., Bogus, S.M., 2014. Mixed logit model-based driver injury severity investigations in single-and multi-vehicle crashes on rural two-lane highways. Accid. Anal. Prev. 72, 105–115.

Wu, Q., Zhang, G., Chen, C., Tarefder, R., Wang, H., Wei, H., 2016a. Heterogeneous impacts of gender-interpreted contributing factors on driver injury severities in single-vehicle rollover crashes. Accid. Anal. Prev. 94, 28–34.

Wu, Q., Zhang, G., Zhu, X., Liu, X.C., Tarefder, R., 2016b. Analysis of driver injury severity in single-vehicle crashes on rural and urban roadways. Accid. Anal. Prev. 94, 35–45.

Wu, Q., Zhang, G., Ci, Y., Wu, L., Tarefder, R.A., Alcántara, A.D., 2016c. Exploratory multinomial logit model-based driver injury severity analyses for teenage and adult drivers in intersection-related crashes. Traffic Inj. Prev. 17, 413–422. https://doi.org/10.1080/15389588.2015.1100722.

Xie, Y., Zhao, K., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accid. Anal. Prev. 47, 36–44. https://doi.org/10.1016/j.aap.2011.12.012.

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

Yamashita, E.Y., 2005. Using a K-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. J. Adv. Transp. 41, 69–89.

Yasmin, S., Eluru, N., 2018. A mixed grouped response ordered logit count model framework. Anal. Methods Accid. Res. 19, 49–61. https://doi.org/10.1016/j.amar.2018.06.002.

Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal. Methods Accid. Res. 1, 23–38.

Yau, K.K.W., 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accid. Anal. Prev. 36, 333–340. https://doi.org/10.1016/S0001-4575(03)00012-5.

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.

Yu, R., Wang, X., Abdel-Aty, M., 2017. A hybrid latent class analysis modeling approach to analyze urban expressway crash risk. Accid. Anal. Prev. 101, 37–43.

Zeng, Q., Wen, H., Huang, H., Abdel-Aty, M., 2017. A Bayesian spatial random parameters Tobit model for analyzing crash rates on roadway segments. Accid. Anal. Prev. 100, 37–43. https://doi.org/10.1016/j.aap.2016.12.023.

Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J., Hao, W., 2019. Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accid. Anal. Prev. 127, 87–95. https://doi.org/10.1016/j.aap.2019.02.029.