Accident Analysis and Prevention 131 (2019) 316–326

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Investigating injury severities of motorcycle riders: A two-step method integrating latent class cluster analysis and random parameters logit model

Fangrong Chang[a,c], Pengpeng Xu[b], Hanchu Zhou[a], Alan H.S. Chan[c], Helai Huang[a,*]

[a] *School of Traffic &Transportation Engineering, Central South University, Changsha, 410075, China*
[b] *Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, 999077, China*
[c] *Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, 99907, China*

ARTICLE INFO

Keywords:
Motorcycle crashes
Injury severity analysis
Latent class clustering
Random parameters logit model
Unobserved heterogeneity

ABSTRACT

Due to the wide existence of heterogeneous nature in traffic safety data, traditional methods used to investigate motorcyclist rider injury severity always lead to masking of some underlying relationships which may be critical for the formulation of efficient safety countermeasures. Instead of applying one single model to the whole dataset or focusing on pre-defined crash types as done in previous studies, the present study proposes a two-step method integrating latent class cluster analysis and random parameters logit model to explore contributing factors influencing the injury levels of motorcyclists. A latent class cluster approach is first used to segment the motorcycle crashes into relatively homogeneous clusters. A mixed logit model is then elaborately developed for each cluster to identify its unique influential factors. The analysis was based on the police-reported crash dataset (2015–2017) of Hunan province, China. The goodness-of-fit indicators and the Receiver Operating Characteristic curves show that the proposed method is more accurate when modeling the riders' injury severities. The heterogeneity found in each homogeneous subgroup supports the application of the random parameters logit model in the study. More importantly, the results demonstrate that segmenting motorcycle crashes into relatively homogeneous clusters as a preliminary step helps to uncover some important influencing factors hidden in the whole-data model. The proposed method is proved to have great potential for accounting for the source of heterogeneity. The injury risk factors identified in specific cases provide more reliable information for traffic engineers and policymakers to improve motorcycle traffic safety.

1. Introduction

Road traffic crashes are the eighth leading cause of death around the world, costing about 1.35 million people's lives each year (World Health Organization, 2018). Almost half of people killed on roads are vulnerable road users who travel with the least protection, i.e. motorcyclists, cyclists, and pedestrians. Given that motorcycle riders share the road with other motor vehicles, they account for nearly a quarter of all road traffic death. According to (National Highway Traffic Safety Administration, 2018), motorcyclist's fatality rate is reported to be 28 times more than that of passenger car occupants in motor vehicle crashes (25.85 and 0.98 deaths per 100 million vehicle miles traveled, respectively), highlighting an urgent need for further efforts to reduce motorcycle severe crashes.

Depending on traffic safety data, numerous researchers have tried to identify the alleviating or aggravating factors influencing motorcycle crash severity (e.g. Quddus et al., 2002; De Lapparent, 2006). However, traffic crashes might occur under distinct conditions, which result in heterogeneity of crash kinematics in nature, significant crash factors hidden, and different magnitude or even opposite direction of the effects that some specific factors have on injury outcomes (Pai and Saleh, 2007; Valent et al., 2002; Ulfarsson and Mannering, 2004). Some measures were therefore introduced to address the heterogeneity issue in motorcycle crash data: focusing on specific crash types (Pai and Saleh, 2008; Pai, 2009), crashes involving a specific age group (Lin et al., 2003), specific crash locations (Chang et al., 2016; Moore et al., 2011), or specific traffic control measures (Pai and Saleh, 2007). In most cases, the segmentation of traffic crash data is based on the objective of the research, methodologies used, and expert domain knowledge, which may lead to a workable segmentation of crash data but cannot guarantee a homogenous group in each segment (Depaire et al., 2008). Data mining techniques such as cluster analysis have then been applied to identify homogenous groups and reduce the heterogeneity in the data (Depaire et al., 2008; Yau, 2004; de Ona et al.,

[*] Corresponding author.
E-mail address: huanghelai@csu.edu.cn (H. Huang).

https://doi.org/10.1016/j.aap.2019.07.012
Received 31 January 2019; Received in revised form 15 July 2019; Accepted 15 July 2019
Available online 25 July 2019

2013). Depaire et al. (2008) used a clustering techniques-latent class as a preliminary analysis to identify homogenous traffic crash patterns and then applied multinomial logit approach to analyze the risk factors of each subgroup. The multinomial logit model combined with a latent class cluster approach was found to be more powerful than aggregately applying a single multinomial logit model to the whole data.

Although the identification of homogenous group can reduce the heterogeneity in the data to some extent, the within-group variation is expected to exist in each identified cluster (Mannering and Bhat, 2014). When analyzing motorcycle riders' injury risk factors, the impacts of explanatory variables are expected to vary across observations in each cluster because an individual has its own specific characteristics that may influence injury outcomes (Chang et al., 2016). However, few previous studies considered the within-cluster heterogeneity when analyzing the injury risk factors. For example, the multinomial logit model used by Depaire et al. (2008), binary logit model used by Sasidharan et al. (2015), and ordered probit model used by Mohamed et al. (2013) are to predict the factors influencing injury outcomes based on the identified clusters. In such cases, the well-known random parameters logit (RP-logit) model is capable as an alternate of addressing the unobserved within-cluster heterogeneity by considering potential variations in the effects of contributing factors.

This study aims at applying a two-step method integrating latent class cluster analysis (LCA) and RP-logit model to identify motorcycle crash patterns and revealing influential factors affecting rider injury severity in specific cases. Specifically, based on the 23,881 motorcycle crashes in Hunan province, China for the period between 2015 and 2017, an LCA is used to divide the whole data into homogenous groups, and then a binary RP-logit model is employed to investigate the significant factors for each subgroup.

2. Methodology

2.1. Latent class cluster analysis

LCA is a probability model based on the cluster analysis method, which assumes that the whole data is divided into exclusive latent classes by an unobserved or latent categorical variable (Collins and Lanza, 2010; Lanza and Rhoades, 2013). The class memberships of individual crashes can be inferred from the observed variables. In LCA, the probabilities of a crash to be in different clusters are estimated based on various models developed for different values of clusters specified. LCA assigns the probability of a crash in every cluster and labels the best index cluster for the cluster with the highest probability of including the crash (Sasidharan et al., 2015). LCA has two notable advantages over conventional cluster analysis: (a) LCA has the advantage of not specifying the number of clusters or labeling individual observation in advance. Different statistical criteria can be used to identify the most appropriate number of clusters (Sasidharan et al., 2015); (b) different types of variables, including counts, continuous, categorical, and nominal variables can be included in the LCA without standardization (Depaire et al., 2008; de Ona et al., 2013).

Cluster analysis aims to identify homogenous clusters in the heterogeneous data by maximizing the similarity within clusters and minimizing the similarity between clusters (Hair and Black, 1998). LCA posits that the data is from a mixture model of various probability distributions (Mohamed et al., 2013). The LCA Stata plugin developed by the Penn State methodology center to allow Stata users to perfume LCA is applied to identify the homogenous group in this study (Lanza et al., 2015).

Two sets of parameters (γ, ρ) provided in the LCA Stata Plugin output are estimated in this study. Gamma (γ) parameters are latent class membership probabilities and the basis for the interpretation of the latent classes; Rho (ρ) parameters represent the item-response probabilities conditional on latent class membership.

Suppose there are n_c latent classes and M crash characteristics associated with crash i; the vector $\boldsymbol{Y_i} = (Y_{i1},\ldots, Y_{iM})$ represent crash i's response to M attributes, and Y_{im} is one possible value of 1,…, r_m; $L_i = 1,2,\ldots, n_c$ is the latent class membership of crash i; I (y_m = k) which is the indicator function equals 1 if the attribute of m equals k, equals 0, otherwise; $\rho_{mk/l}^{I((y_m=k)}$ means the probability that a crash has the attribute k of characteristic m, conditional on membership in the latent class l. Then the response of crash i is (Lanza et al., 2015):

$$P(Y_i = y) = \sum_{l=1}^{n_c} \gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk/l}^{I(y_m=k)} \tag{1}$$

The maximum likelihood approach is used to estimate the parameters by means of the expectation maximization algorithm. Based on the Bayesian theorem, the posterior membership probability of a crash can be stated as:

$$P(L = l/Y = y) = \frac{(\prod_{m=1}^{M}\prod_{k=1}^{r_m}\rho_{mk/l}^{I(y_m=k)})\gamma_l}{\sum_{l=1}^{n_c}\gamma_l\prod_{m=1}^{M}\prod_{k=1}^{r_m}\rho_{mk/l}^{I(y_m=k)}} \tag{2}$$

The most appropriate number of clusters "n_c" is unknown in the LCA. By trying models with different number of clusters, we can find the most appropriate value of clusters. Based on the characteristics of a crash, we can calculate their maximum posterior probability and then assign them to a latent class. During this process, the choice of the number of clusters is to minimize assignment error (Collins and Lanza, 2010). Bayesian Information Criteria (BIC), Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (CAIC), and entropy-based measures are used to select the most appropriate number of clusters. The statistic figures of BIC, AIC, and CAIC model can measure the fitness of a model while considering the complexity. The number of clusters with a lower value of BIC, AIC, CAIC is superior. BIC is considered more reliable than AIC and CAIC when identifying the cluster number (Biernacki and Govaert, 1999). However, Bijmolt et al. (2004) suggested that the increased cluster number might not minimize the statistical figure when the sample is large enough. The percentage reduction in BIC between models with different clusters was then suggested.

2.2. Random parameters logit model

Although individuals in each cluster are considered relatively homogenous, there may still be heterogeneity among individuals in each subgroup. An RP-logit model is therefore used to identify the influential factors of motorcycle riders' injury in the whole data and individual clusters. Due to the serious underreporting problem of property damage only crashes (Elvik and Mysen, 1999), only injury and fatal crashes are included in the study. Compared to severe and slight injury crashes, the number of fatal crashes is small, especially in individual clusters. As such, fatal crashes are combined with severe injury crashes (KSI crashes) in the following analysis (Meng et al., 2017; Chang et al., 2019). Therefore, a binary RP-logit model is used to model riders' injury severities and is defined as:

$$S_{in} = \beta_i X_{in} + \varepsilon_{in} \tag{3}$$

Where S_{in} is a severity function deciding the injury outcome; $\boldsymbol{\beta}$ is a vector of estimable parameters; $\boldsymbol{X_{in}}$ is a vector of explanatory variables (e.g., rider, vehicle, road, and environment characteristics); ε_{in} is error term, which is assumed to follow a Generalized Extreme Value (GEV) distribution.

Maximum likelihood approach is used to estimate the random parameters and is shown as:

$$P_{ij} = \int \frac{EXP\,[\beta_i X_{in}]}{1 + EXP\,[\beta_i X_{in}]} f\,(\beta/\varphi) d\beta \tag{4}$$

Where $f\,(\boldsymbol{\beta}/\boldsymbol{\varphi})$ is the density function of $\boldsymbol{\beta}$ with $\boldsymbol{\phi}$ referring to a vector of parameters of the density function (mean and variance), and all other terms are as previously defined. $\boldsymbol{\beta}$ can account for the variations of the effects of **X** on the injury outcomes with the density function $\boldsymbol{f}\,(\boldsymbol{\beta} \mid \boldsymbol{\phi})$ used to determine $\boldsymbol{\beta}$. Parameters can be fixed or random. The logit probabilities are approximated by drawing values of $\boldsymbol{\beta}$ from $f\,(\beta/\varphi)$ for given values of $\boldsymbol{\varphi}$. The function form of the density function is assumed to be normal distribution. 200 Halton draws which have been shown to provide more efficient distribution of draws for numerical integration than purely random draws are used in this study when estimating parameters (Train, 2009).

3. Data

A three-year crash database was collected from the Traffic Management Sector-Specific Incident Case Data Report 2015–2017 maintained by the Traffic Administration Bureau of Hunan Department of Public Security. Located in southern China, Hunan is a typical province with a population of 67.4 million and an area of 211,800 km^2 and composed of 14 prefectures, 122 counties, and 2576 townships. The province ranked 9th among 31 provinces in terms of Gross Domestic Product value of 3.46 trillion CNY (i.e., 510 billion USD) in 2017 (National Bureau of Statistics, 2018). The number of registered motorcycles in Hunan was about 6 million in 2016, accounting for 55% of all motorized vehicles (Hunan Provincial Bureau of Statistics, 2017) According to the police records, a total of 30,033 motorcycle crashes were reported in Hunan Province during the period from 2015 to 2017, among which 5795 (19%) and 5467 (18%) crashes are property damage only (PDO) and fatal crashes, respectively. However, the proportion of PDO crashes which is almost equal to that of fatal crashes in the dataset is inconsistent with the fact that the number of crashes decreases with the increase of injury severities (Elvik and Mysen, 1999; Ahmed et al., 2017). Given the potential under-reporting of no injury crashes, PDO crashes were therefore excluded from our sample. In addition, another 357 crash records (1.47%) were also removed due to their absence of some important information. Thus, 23,881 motorcycle traffic crashes were retrieved for further analysis, among which 23% are fatal or serious injury level. It is noteworthy that the focus of our analysis is the injury severity sustained by motorcycle riders only.

In China, crashes and the injury severities are commonly collected and assessed by police officers at the traffic crash scenes. The injury severities are categorized as property damage only (i.e. no injury), slight injury (i.e., non-disability injury), serious injury (i.e., disability injury) and fatality (i.e., immediate or subsequent death from injuries within 7 days after a crash). As aforementioned that no injury crashes are often largely under-reported, these crashes were excluded from our sample and the whole crash data was divided into two categories: slight injury; killed and serious injury (KSI). By aggregating the crash and casualty injury profiles, the predictor variables reflecting demographic characteristics of motorcyclists (i.e., age and gender), riders' illegal behavior, crash characteristics (i.e. passenger on board or not, at-fault party, motorcycle type, and collision objects), geometric design features (segment or intersection, traffic control measures, and road type), together with the environment factors (i.e., crash time, weather, and lighting condition) were extracted and summarized in Table 1.

Regarding the variable classification, illegal behavior, road types and traffic control measures which account for less than 1% of the whole sample were combined as other violations, other types, and other measures, respectively. Based on the transport task, function and traffic volume, roads are divided into three types known as expressway, ordinary highway, and urban highway. Specifically, the ordinary highway is classified as first-class, second-class, third-class and fourth-class. Urban highway includes urban expressway and general urban street. Other road types include urban expressways and substandard roads which fail to meet the requirements of the National Highway Engineering Technical Standards. As for motor vehicle type, the heavy motor vehicle includes large and medium-sized coaches and trucks in our study.

Table 1
Descriptive statistics of motorcycle crashes.

Variables **Motorcycle crashes**	Number 23,881	Slight injury 18,407	Fatal/severe injury 5,474
Rider age			
Under 18	5.21%	5.56%	4.06%
18-24	6.80%	7.27%	5.21%
25-34	13.53%	14.02%	11.86%
35-44	17.58%	17.79%	16.88%
45-54	33.21%	33.21%	33.23%
55-64	18.37%	17.63%	20.84%
Over 65	5.29%	4.51%	7.93%
Rider gender			
Male	80.89%	80.10%	83.56%
Female	19.11%	19.90%	16.44%
Resident			
Urban resident	69.24%	70.37%	65.45%
Rural resident	30.76%	29.63%	34.55%
Illegal behavior			
No violations	17.76%	18.84%	14.14%
Drunk riding	2.14%	1.73%	3.54%
Violating traffic facilities	2.21%	2.22%	2.17%
Failing to yield	7.14%	7.46%	6.05%
Overtaking	1.71%	1.74%	1.59%
Riding without helmets	1.57%	1.10%	3.14%
Riding without licenses	42.64%	42.13%	44.37%
Riding in the wrong direction	2.17%	2.03%	2.65%
Approaching illegally	4.89%	4.93%	4.77%
Following too close	1.59%	1.55%	1.72%
Other violations	16.17%	16.26%	15.86%
Passenger	0.00%		
Without passenger	65.86%	63.42%	74.06%
With passenger	34.14%	36.58%	25.94%
Movement prior to crash			
Making U-turn	6.73%	6.79%	6.50%
Turing left	10.57%	10.16%	11.95%
Turning Right	3.18%	3.09%	3.47%
Going Straight	79.53%	79.96%	78.08%
Fault	0.00%		
At-fault	54.63%	51.86%	63.98%
Non-fault	45.37%	48.14%	36.02%
Motorcycle type			
Motorcycle	91.71%	92.11%	90.37%
Scooter	8.29%	7.89%	9.63%
Road type			
First-class highways	3.68%	3.13%	5.55%
Second-class or lower highways	63.14%	62.83%	64.18%
General urban streets	26.15%	26.91%	23.58%
Other types	7.03%	7.13%	6.69%
Traffic control measures			
Uncontrolled	50.36%	51.46%	46.64%
Signal-controlled	4.61%	4.60%	4.64%
Mark/sign-controlled	44.16%	43.05%	47.88%
Other measures	0.88%	0.89%	0.84%
Crash location	0.00%		
Segment	78.62%	78.31%	79.67%
Three-legged intersections	11.55%	12.00%	10.07%
Four-legged intersections	9.82%	9.69%	10.27%
Road surface			
Dry	82.35%	82.66%	81.29%
Wet	17.65%	17.34%	18.71%
Area (prefecture)			
Changsha	16.63%	17.69%	13.04%
Changde	12.71%	12.33%	13.99%
Chenzhou	11.04%	11.33%	10.05%
Hengyang	14.30%	15.33%	10.81%
Loudi	2.02%	1.83%	2.67%
Shaoyang	5.05%	5.07%	4.95%
Xiangtan	6.04%	5.76%	6.98%
Xiangxi	1.25%	1.09%	1.81%
Yueyang	4.34%	4.47%	3.89%
Huaihua	4.82%	4.11%	7.22%
Yongzhou	4.83%	4.48%	6.03%
Zhangjiajie	3.33%	3.07%	4.20%
Zhuzhou	11.52%	12.29%	8.91%
Yiyang	2.13%	1.14%	5.44%
Season			

(*continued on next page*)

Table 1 (*continued*)

Variables **Motorcycle crashes**	Number 23,881	Slight injury 18,407	Fatal/severe injury 5,474
Spring	25.16%	25.39%	24.41%
Summer	30.23%	30.69%	28.66%
Autumn	23.65%	23.31%	24.77%
Winter	20.96%	20.61%	22.16%
Weather			
Sunny	66.23%	66.27%	66.09%
Cloudy	19.86%	19.88%	19.78%
Rain/Snow	13.91%	13.85%	14.12%
Visibility			
Lower than 50m	13.17%	12.44%	15.62%
50m-100m	29.44%	29.68%	28.63%
100m-200m	22.45%	22.56%	22.09%
Over 200m	34.94%	35.31%	33.67%
Lighting condition			
Daylight	70.94%	72.56%	65.47%
Twilight	6.22%	6.12%	6.54%
Lighted dark	12.54%	12.25%	13.50%
Complete darkness	10.31%	9.06%	14.49%
Day of week			
Weekday	72.04%	72.26%	71.30%
Weekend	27.96%	27.74%	28.70%
Time of day			
00:00-06:59	8.52%	7.69%	11.33%
07:00-08:59 (morning peak)	10.84%	11.07%	10.07%
09:00-11:59	16.69%	17.27%	14.72%
12:00-16:59	31.06%	31.60%	29.23%
17:00-19:59 (evening peak)	21.68%	21.96%	20.73%
20:00-23:59	11.21%	10.40%	13.92%
Collision objects			
Single vehicle	48.32%	47.32%	51.66%
Pedestrians	2.68%	3.13%	1.13%
Non-motor vehicles	1.49%	1.73%	0.69%
Light motor vehicles	43.22%	44.47%	39.02%
Heavy motor vehicles	2.97%	2.19%	5.59%
Multi-vehicle crashes	1.32%	1.15%	1.90%

4. Results and discussion

This section shows the results of LCA and RP-logit for motorcycle crashes. To select the appropriate number of clusters, different criteria and the entropy values are discussed. A Receiver Operating Characteristic (ROC) curve analysis is also used to compare cluster analysis and the whole data analysis.

4.1. Cluster analysis

Models with the different number of clusters, from one to twelve, were preliminarily estimated for motorcycle crashes using all of the crash characteristic variables. AIC, BIC, and CAIC criteria were used to determine the most appropriate number of clusters. If these criteria values become stable during any stage from one to twelve clusters, then the appropriate number of clusters can be determined. Otherwise, this procedure will be continued from twelve clusters onwards. As shown in Fig. 1, three criteria decrease with the number of clusters increasing. However, the percentage decrease in BIC, AIC, and CAIC drops to less than 1% from six clusters onwards, which supports a good separation of the whole data. In addition, the entropy value of six clusters is 0.98, which is quite high and indicates a good fitness of model. Motorcycle crashes are thus classified into six clusters for further analysis (Table 2).

In terms of the entropy R^2 line in the figure, it is worth noting that the entropy had a precipitous decrease for five clusters and rose afterwards. As the entropy indicates how well one model can predict class memberships based on the observed variables, the lower value for entropy suggests that five clusters cannot separate the whole data very well, for the difference in the probability of an observation falling into the five clusters is not as significant as that in other clusters.

The six-cluster model provides univariate distributions for each variable based on each cluster, allowing us to identify each cluster as a specific crash pattern. Table A1 of Appendix showed the distributions of each variable in six clusters. The skewed feature distribution which differs between the clusters can help to identify the specific characteristics of each cluster.

For example, all crashes in cluster 4 happened in darkness with lighting while other clusters contain less than 8% of crashes under the same condition. As such, the variable of darkness with lighting can differentiate cluster 4 from other clusters. Cluster 4 can be defined as "motorcycle crashes in the lighted dark".

Similarly, about 98% of crashes in cluster 5 occurred during evening peak hours (17:00-19:59), while the other clusters have fewer distributions during this time period. Besides, most crashes in cluster 5 happened on second-class or lower highways. We refer to this cluster as "motorcycle crashes on second-class or lower highways during evening peak hours".

Complete darkness is specific to cluster 2 for accounting for 82% of crashes while less than 18% of crashes in the other clusters occurred under the same lighting condition. This cluster is thus characterized by "motorcycle crashes in complete darkness".

More than 76% of crashes in cluster 1 occurred in other types of road and 43% of crashes are in Changsha while the other clusters have a lower proportion for these two features. We refer to cluster 1 as "motorcycle crashes on other types of road in Changsha".

Cluster 3 overlaps with cluster 5 in terms of second-class or lower highways but distinguishes itself by an over-representation of the non-peak period in the afternoon (12:00-16:59). Cluster 3 can be therefore identified as "motorcycle crashes on second-class or lower highways during nonpeak hours in the afternoon".

Cluster 6 covers almost 100% of motorcycle crashes on urban streets and is hence described as "motorcycle crashes on general urban streets".

An overview of the identified clusters is shown in Table 3. Our analysis makes a distinction between motorcycle crashes in specific cases through road types, different time in a day, various lighting conditions, and areas.

4.2. Injury severity analysis

Random parameters logit models are developed and estimated for both the whole data and six clusters using maximization of the log-likelihood method. The ROC curves in Fig. 2 show that the Area Under ROC Curve (AUC) area for six-cluster models is 0.968, significantly greater than that for the whole sample, 0.823, indicating the predicted injury probability in the cluster-based model are more reliable than that predicted by the pooled model. In addition, in terms of goodness-of-fit measures for the models shown in Table 4, the separated models are definitely superior in terms of AIC statistics (approximately 110 points lower) and likelihood-ratio test (at 0% level of significance). Significant contributing factors identified in the models are shown in Table 5. The significance level used in this study is 5%.

4.2.1. General random parameters logistic regression analysis

Compared to the male, the female motorcycle riders were found less likely to be severely injured in crashes overall with a mean of -0.275 (0.038) and a standard deviation of 0.503 (0.048). According to the cumulative probability function of the normal distribution, the figures implied that 29% of female riders tended to be severely injured despite the decreased injury risks for the majority of females. This result is quite consistent with the findings from previous studies. For example, females were found to be less likely to be involved in severe injuries in some studies (Albalate and Fernandez, 2010; Shaheed et al., 2013) but more likely to sustain severe injuries in other studies (Savolainen and Mannering, 2007; Shaheed and Gkritza, 2014). The random effects across the female rider population are likely picking up a complex interaction among various human-factor related or physiological

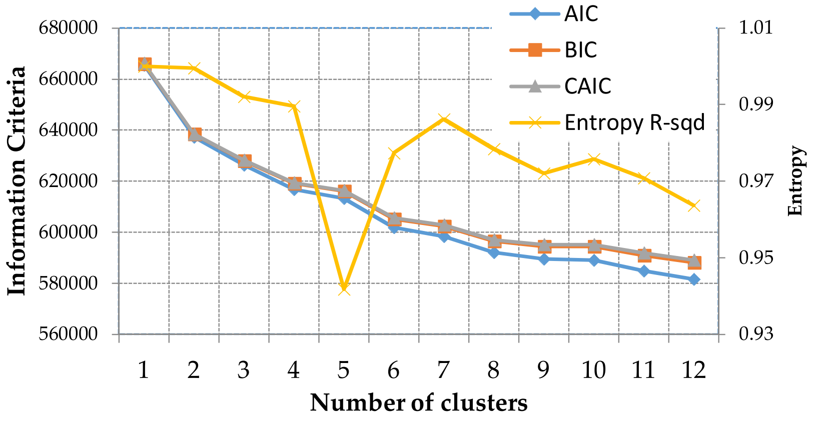

Fig. 1. Identification of the number of clusters for motorcycle crash analysis.

elements (such as perception/reaction times, experience, attention to the road, visual acuity, weight, height, and other) and the physics of the collision. Motorcycle riders in rural areas were found to have a higher probability of getting fatally or severely injured in crashes, which may be attributed to the lack of basic safety facilities on rural highways (The National People's Congress of the People's Republic of China, 2017). The variable for rural residents produced a normally distributed parameter with a mean of 0.074 (0.032) and a standard deviation of 1.255 (0.041), suggesting that 52% of rural motorcyclists had an increased probability of fatality or serious injury in crashes while the remaining were less likely to be severely injured. This result reflects the heterogeneity of motorcycle riders in rural areas.

With regard to riders' illegal behavior, alcohol-impaired motorcyclists are more likely to be involved in severe crashes. The indicator variable for drunk-riding leads to a random parameter with a mean of 0.303 (0.105) and a standard deviation of 2.551 (0.196). Given these distributions, 55% of the distribution was greater than zero while 45% was smaller than zero, revealing the heterogeneity in the drunk-riding motorcyclists. While drunk riding has a higher chance of causing motorcyclist serious injury due to alcohol-impaired riders' tendency of losing control of motorcycle (Kasantikul et al., 2005), some motorcycle drivers may have some risk-compensation operations for the influences under alcohol. Consistent with other studies (De Lapparent, 2006; Savolainen and Mannering, 2007; Xiao et al., 2018), motorcycle riders without helmets were also found to sustain a severe injury in crashes with the probability of 268% higher than helmeted riders. In fact, wearing a helmet has been widely reported to effectively protect riders from the higher risks of head injuries which have been demonstrated as the main cause in fatal motorcyclists (Rowland et al., 1996; Norvell and Cummings, 2002). The absence of license was found to significantly increase the injury severity of riders, which may be explained by unlicensed drivers' higher tendency towards risky driving behavior (Tsai et al., 2008).

Table 3
Clusters identified.

1	Motorcycle Crashes on other types of road in Changsha
2	Motorcycle crashes in complete darkness
3	Motorcycle crashes on second-class or lower highways during nonpeak hours in the afternoon
4	Motorcycle crashes in the lighted dark
5	Motorcycle crashes on second-class or lower highways during evening peak hours
6	Motorcycle crashes on general urban streets

Due to the unstable structure of motorcycle, the appearance of passengers may influence the rider's driving behavior from both mental and physical aspects. Generally speaking, riders tend to drive more cautiously when they have passengers on motorcycle. But it is also true that the passengers could increase the difficulty of controlling the vehicle, especially if the pillion passenger is a female who often uses an unbalanced sitting posture on motorcycle (Waseem et al., 2019). This may explain the random effects of passengers on rider injury as suggested by the estimated parameters with a mean of -0.779 (0.035) and a standard deviation of 1.304 (0.044). Towards the effect of motorcycle movement prior to the crash, left-turn movement imposes riders at an increased injury risk with the probability of 14% higher, compared to straight-going movement prior to the crash. This could be attributed to more conflicts that left-turn vehicles face than those going straight, including with left-turning traffic from the same approach or from the different approaches, and with through traffic from other approaches (Wang and Abdel-Aty, 2008). In terms of motorcycle type, the parameter for scooter is significant and random with a mean of 0.257 (0.050) and a standard deviation of 0.323 (0.067) which suggests the heterogeneity in scooter riders. More specifically, these distributions indicated that 79% of scooter riders have a higher likelihood of being fatally and severely injured while the remaining show an opposite

Table 2
The sample size of each cluster and distribution of interesting variables.

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
The number of observations	1,926	1,928	9,987	2,666	3,537	4,173
KSI	4,68	552	2,172	672	827	776
Slight injury	1,458	1,376	7,815	1,994	2,710	3,397
Time of day: 12:00-16:59	37.23%	0.00%	50.64%	0.00%	0.00%	39.37%
Time of day: 17:00-19:59 (evening peak)	22.69%	0.15%	0.00%	12.57%	98.16%	22.31%
Area: Changsha	43.46%	13.37%	10.81%	18.30%	11.51%	21.52%
Lighted darkness	2.91%	0.00%	0.12%	100.00%	7.35%	0.00%
Complete darkness	5.30%	81.56%	0.00%	0.00%	17.27%	3.74%
Second-class or lower highways	0.05%	61.27%	98.56%	23.63%	96.35%	0.00%
General urban streets	0.00%	9.94%	0.00%	70.48%	0.00%	99.95%
Other types of road	76.74%	5.69%	0.00%	3.38%	0.00%	0.00%

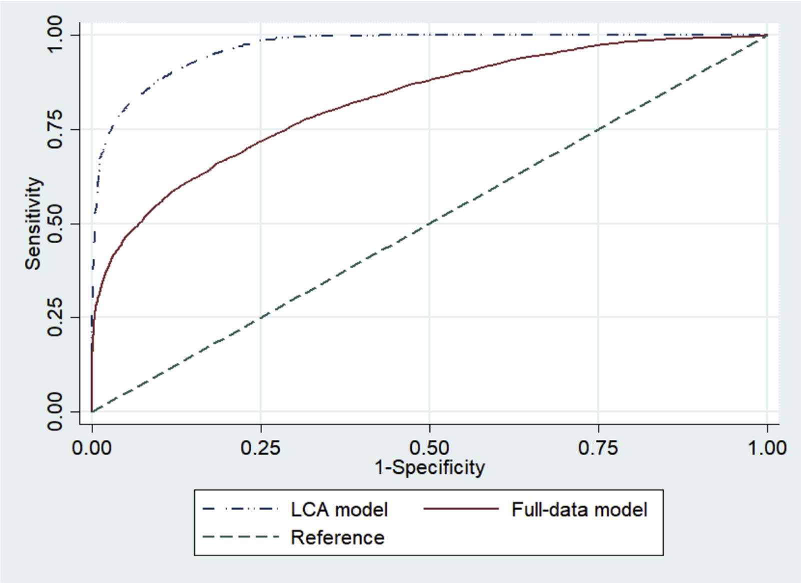

Fig. 2. ROC Curves for full-data model and LCA model.

trend.

With respect to road characteristics, first-class, second-class and lower highways are more related to severe crashes, compared to urban streets. The estimated parameters for these two variables are significant and random with a mean of 0.515 (0.086), 0.199 (0.038) and a standard deviation of 2.430 (0.147), 0.688 (0.026), respectively. The heterogeneity may result from motorcyclists' different driving habits. Given that higher-level roads are associated higher speed that has been broadly known to cause severe crashes (Shaheed and Gkritza, 2014), people with limited experience behind the wheel or who recently had an accident tend to have some risk-compensation behaviors to avoid high-risk situations on high-speed highways. Traffic signal- or signs/marks-controlled roads increase the probability of rider involving KSI crashes by almost 20% than uncontrolled roads, which is contradictory to the conclusions from Quddus et al. (2002) and Pai and Saleh (2008). One explanation could be found in the study by Haque et al. (2008) that the practice of earlier discharge during the initial green period is particularly risky for resulting in a serious crash with the crossing vehicles. As for the crash locations, motorcycle riders were found to sustain a higher-level injury in four-legged intersections than those in segments, which may be due to more complex traffic flows in cross intersections.

Concerning the environment variables, visibility lower than 50 m, twilight, complete darkness, and the time between 20:00-23:59 increase the probability of the occurrence of fatal/severe crashes. The reason behind the increases may be that drivers do not have good visibility or enough reaction time under these conditions. The result is consistent with previous studies (Quddus et la., 2002; Shaheed et al., 2013). Examining crash characteristics, collisions with heavy motor vehicles and multi-vehicle collisions were found to have significant effects on rider injury level. Both variables lead to normally distributed parameters which suggest that the majority of collisions with heavy motor vehicles and multi-vehicle are more likely to result in increased injury severity for motorcyclists. Similar findings were suggested by Chang et al. (2016).

4.2.2. *Cluster-based random parameters logistic regression analysis*

In terms of cluster-based random parameters logistic regression model results, some differences between the results of the pooled model and those of six reduced models were discovered. More importantly, the cluster models revealed new information which cannot be found in the model using the whole data and showed a great potential of exploring the source of heterogeneity.

First, as expected, the analysis of traffic crashes based on the heterogeneous data may obscure some significant and important contributing factors. For example, no violations committed by riders (as opposed to riders' other behaviors) is not significant in the full-data model. However, the estimated odds ratio for cluster 2 indicates that the probability of riders getting fatal or severe injuries increases by 61.93% even if motorcycle riders do not have any violations in darkness, which is critical information for improving motorcycle's visibility at night rather than curing riders' behavior blindly. In addition, travelling in the wrong direction is also an insignificant factor in the pooled model but lowers the probability of rider fatality or severe injuries by 31.55% in the lighted dark (cluster 4). Similarly, cluster 2 analysis suggests that motorcycle riders in single crashes are 34% more likely to be fatally/seriously injured in darkness but are not significantly different from the ones in other types of crashes in the pooled model. Both findings also highlight the importance of improving

Table 4
Goodness-of-fit measures for the whole-data model and cluster-based models.

	Whole data	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster6
Number of observations	23881	1926	1928	9987	2666	3537	4173
Number of parameters (K)	58	23	13	23	15	23	16
Log likelihood at zero	−12173.3	−976.6	−949.1	−4988.8	−1432.9	−1809.8	−1898.2
Log likelihood at convergence	−12148.6	−964.9	−948.0	−4983.0	−1428.0	−1804.7	−1897.3
AIC	24415.1	1975.8	1592	10017.9	2885.9	3655.5	3826.5
Likelihood-ratio test							
$X^2 = -2\,[\mathrm{LL}\,(\mathrm{fi}_{\text{whole data}}) - \mathrm{LL}(\mathrm{fi}_{\text{cluster-based}})$			224				
Degree of freedom			55				
P			0.000				

Table 5
Random parameters logit estimation results for injury severity analysis.

Variables	Whole data		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Female	−0.275	0.038	−0.772	0.162	−0.501	0.139	−0.423	0.063	−0.475	0.109				
s.d. female	*0.503*	*0.048*	*1.573*	*0.213*	*1.803*	*0.231*	*1.510*	*0.087*						
Rural resident	0.074	0.032	0.330	0.104	0.259	0.089	0.124	0.039	0.384	0.090			0.226	0.073
s.d. rural resident	*1.255*	*0.041*												
No violations					0.482	0.138								
Drunk riding	0.303	0.105			0.749	0.175							0.434	0.194
s.d. drunk riding	*2.551*	*0.196*												
Violating traffic facilities	−0.217	0.099	0.970	0.355					−0.505	0.178				
Failing to yield	−0.518	0.069					−0.192	0.076	−1.517	0.311	−1.804	0.278	−1.432	0.298
s.d. failing to yield	*0.961*	*0.085*							*2.395*	*0.436*	*3.778*	*0.403*	2.825	0.396
Overtaking	−0.275	0.111					−0.295	0.122						
Riding without helmets	1.303	0.099	2.027	0.543	1.483	0.387	0.995	0.144			2.805	0.323	0.810	0.191
s.d. riding without helmets			*2.337*	*0.950*			*1.144*	*0.228*						
Riding without licenses	0.074	0.032	0.284	0.098					−0.248	0.082			0.268	0.062
Riding in the wrong direction									−0.379	0.181				
Approaching improperly	−1.093	0.115									−0.667	0.174		
s.d. approaching improperly	*3.370*	*0.185*												
Following too close	−0.912	0.167					−0.667	0.160						
s.d. following too close	*2.980*	*0.278*												
With passenger	−0.779	0.035	−0.697	0.119	−0.641	0.107	−0.373	0.041	−0.354	0.081	−0.864	0.102	−0.429	0.069
s.d. *with passenger*	*1.304*	*0.044*			*1.247*	*0.150*								
Making U-turn	−0.130	0.056	−0.447	0.188							−0.419	0.201		
Turing left	0.131	0.048									0.609	0.152		
Turning Right											0.618	0.244		
Non-fault	−0.658	0.032	−0.466	0.102	−0.347	0.110	−0.675	0.044	−1.436	0.142	−1.819	0.142	−0.449	0.061
s.d. non-fault	*0.815*	*0.033*					*0.977*	*0.044*	*3.320*	*0.222*	*3.983*	*0.214*		
Scooter	0.257	0.050					0.226	0.069			0.615	0.152		
s.d. scooter	*0.323*	*0.067*												
First-class highways	0.515	0.086	−0.980	0.242	0.360	0.169	0.345	0.148						
s.d. first-class highways	*2.430*	*0.147*	*6.588*	*0.675*										
Second-class or lower highways	0.199	0.038									−0.762	0.214		
s.d. second-class or lower highways	*0.688*	*0.026*												
Other types	0.151	0.062												
Signal-controlled	0.189	0.073	−0.719	0.329			0.453	0.192						
Mark/sign-controlled	0.180	0.031			0.245	0.084	0.328	0.039						
s.d. mark/sign-controlled	*0.815*	*0.032*												
Other traffic control measures	−3.019	0.624												
s.d. other traffic control measures	*7.895*	*1.102*												
Three-legged intersections	−0.649	0.062	−1.644	0.631			−0.239	0.063			−0.623	0.163		
s.d. *three-legged intersections*	*2.239*	*0.091*	*8.000*	*1.546*										
Four-legged intersections	0.174	0.048											0.228	0.066
Spring	−0.122	0.034											−0.162	0.069
Summer	−0.327	0.036			−0.250	0.087	−0.126	0.044			−0.398	0.101		
s.d. summer	*1.342*	*0.044*					*0.564*	*0.053*						
Winter			0.412	0.118										
Cloudy							−0.136	0.049						
Visibility Lower than 50m	0.124	0.045	0.353	0.139									0.315	0.109
Visibility 50m-100m							−0.146	0.045						
Twilight	0.168	0.061												
Complete darkness	0.431	0.050											0.340	0.138
Weekend			−0.225	0.110										
00:00-06:59									0.258	0.083				
07:00-08:59	−0.126	0.053												
s.d. 07:00-08:59	*0.588*	*0.062*												
12:00-16:59	−0.095	0.039												
s.d. 12:00-16:59	0.526	0.037												
17:00-19:59	−0.326	0.044							−0.460	0.134	−1.393	0.244		
s.d. 17:00-19:59	1.285	0.051									*2.458*	*0.106*		
20:00-23:59	0.128	0.051												
Single vehicle			0.293	0.101					0.476	0.078	0.213	0.092		
Pedestrians	−1.594	0.132			−1.448	0.235	−1.516	0.318			−1.242	0.248	−1.375	0.417
s.d. pedestrians	*0.849*	*0.171*												
Non-motor vehicles	−1.180	0.149	−0.907	0.396			−0.755	0.226			−1.378	0.419	−0.875	0.327
Heavy motor vehicles	0.662	0.085	1.227	0.276	0.678	0.191	0.553	0.091	0.464	0.180	1.292	0.235	0.600	0.158
s.d. heavy motor vehicles	*2.521*	*0.162*			*0.759*	*0.277*								
Multi-vehicle crashes	0.580	0.120	1.251	0.471			0.451	0.142			4.309	0.784		
s.d. multi-vehicle crashes	*1.253*	*0.183*									*12.732*	*2.942*		

Note: S.E. is the abbreviation of standard error; s.d. denotes the abbreviation of standard deviation; the italicized represents estimates for the variables resulting in random parameters.

lighting conditions to alleviate motorcyclists' injury severities, which is ignored in the general model. Right-turn movement prior to the crash is another factor obscured in the general model. Turing right is not significantly different from that going straight in terms of the effects on rider injury severity in the whole-data driven model but has a higher probability of causing KSI crashes in cluster 5. Given that 98% of the crashes in cluster 5 occurred during evening peak hours, the higher likelihood of increasing rider injury severity could be explained by the aggressive driving and driving anger during rush hours (Deffenbacher et al., 1994; Zhang et al., 2015). After one day's work, drivers want to go home and have a rest, but their arrival goals are always blocked by a traffic jam or others' sudden cut-in during rush hours. In addition, right-turning vehicles do not need to obey traffic lights in China and thus have more conflicts with other road users in intersections. In this situation, driving anger could be triggered and drivers tend to behave aggressively, which endangers the road users especially motorcycle riders with fewer protection facilities. This result highlights the management of traffic condition during peak hours for safe traffic.

Second, the magnitude of the effects of factors is quite different between the general model and cluster-based models. Based on the results of the pooled model, riders who make U-turn are less likely to be fatally or severely injured with 12.19% lower compared to the ones going straight prior to the crash while the figures for making U-turn on other types of road (cluster 1) or on second-class or lower highways during peak hours (cluster 5) are much smaller with 36.05% and 34.23% lower, respectively. These lower probabilities may be explained by drivers' carefulness in these situations, for example, driving at a lower speed during rush hours when there are many vehicles and on other types of road with many pedestrians (e.g. the roads in the living areas). The effects of darkness (as opposed to daylight) are larger in the pooled model, compared to those for cluster 6. The difference of the estimated parameter in the pooled model and cluster models can also be observed in other factors including four-legged intersections, visibility distance lower than 50 m, and collisions with non-motor vehicles. In addition, the same factors might have a different degree of effects for different clusters. For example, the indicator variable for collision with non-motor vehicles shows a different level of influences on injury outcomes with the estimated odds ratio of 0.40, 0.49, 0.47, and 0.42 for clusters 1, 3, 5 and 6, respectively.

Third, the effects of predictors on riders' injury severities may be variable under some conditions but relatively fixed under other conditions. For instance, riding without helmets was found to be a fixed variable in the full-data model while it produced a random parameter with a mean of 2.027 and a standard deviation of 2.337 for cluster 1, and 0.995 and 1.144 for cluster 3 respectively. Given these distributional parameters, 81% of the distribution is greater than zero in two clusters, suggesting that 81% of non-helmeted riders tend to be severely injured on other types of road in Changsha and second-class or lower highways during nonpeak hours in the afternoon. This random effect has been well explained by Shankar and Mannering (1996). In addition, the parameter for the indicator variable of rural resident is normally distributed with a mean of 0.074 and a standard deviation of 1.255 in the pooled model, indicating that 52% of rural residents are more likely to be fatally or severely injured in crashes while the rest proportion (48%) have a lower probability of being fatally or severely injured. But the variable was observed to be significant and fixed for the clusters 1, 2, 3, 4 and 6, suggesting that rural riders have an increased probability of fatal or severe injuries in some specific cases.

Finally, cluster models even reveal the opposite effects of some variables on riders' injury outcomes. While riders who violate traffic facilities (compared to other behaviors influencing traffic safety) sustain a reduced probability of getting fatally or severely injured by 23.73% in the general model, the variable for violating traffic facilities was found to have different signs for cluster 1, resulting in negative effects on the injury outcomes of riders with the probability increasing by 163.79%. The negative effects of violating traffic facilities on rider injury may be caused by less attention to other vehicles paid by drivers on other types of road with a smaller number of vehicles than on urban roads. This result emphasizes education and enforcement actions for curbing riders' behavior of violating traffic-control information which is underestimated and even completely ignored based on the results of the general model. Similarly, the indicator variable for riding without licenses which leads to a significant and positive parameter in the general model was demonstrated to lower the probability of fatal or severe injury by 21.96% for cluster 4. Furthermore, compared to urban streets, riders' probability of being fatally or severely injured increases on highways (first-class, second-class and lower highway) due to the higher speed. On the contrary, highways show a positive effect on decreasing injury severities for both cluster 1 and 5.

The estimated parameter for the variable of first-class highways is normally distributed in the general model, whose statistics suggest that 58% of the distribution is greater than zero while 42% is smaller than zero. This heterogeneity could be explained to some extent by the model results of clusters 1, 2 and 3. The variable produced a random and general negative parameter for cluster 1, which account for its negative part in the general model, and a definitely positive parameter for both clusters 2 and 3, implying that the heterogeneity of the whole data has been decreased. These estimations also imply that the heterogeneity could be explained to some extent, for example, the heterogeneous effects of the first-class highways in the whole sample are from different situations that crash occurred (cluster 1, 2 and 3). Similarly, the estimated parameter for second-class or lower highways was found to be random with a mean of 0.199 (0.038) and a standard deviation of 0.688 (0.675), which suggest that 61% of second-class or lower highways tend to increase rider injury severity while 39% are likely to decrease rider injury. This variable produced a definitely negative parameter for cluster 5, which can account for the negative part of the estimation in the whole-data model, suggesting that during evening peak hours crashes occurring on the second-class or lower highways are less likely to pose severe injuries on riders. In addition, many other variables which showed random effects in the whole-data model and fixed effects in any cluster models, including female, rural resident, drunk riding, and failing to yield etc. could also indicate that the heterogeneity was reduced in the cluster-based approaches and the corresponding positive or negative part in the general model was accounted for in some degree. Therefore, clustering is indeed able to reduce the heterogeneity of the whole sample to some extent by making a separation between clusters and more importantly help achieve a better understanding of the source of heterogeneity.

Although clustering approach reduced the heterogeneity of the whole sample, the effects of some factors on rider injury are still random for some clusters, such as female, non-fault, and first-class highways, etc. This implies that the heterogeneity still exists within each subgroup, supporting the use of random parameters logit model.

5. Conclusion

As a convenient and affordable means of transportation, motorcycle is widely used for daily commuting, especially in less-developed cities and rural areas in China. Considering the high fatality rate of motorcycle riders, many researchers have attempted to reveal injury risk factors by applying conventional approaches. However, traditional methods used to analyze motorcyclist injury severities were found to be inadequate in revealing hidden relationships between some influential factors and injury severities, which may be vital for suggesting countermeasures and developing policy to improve traffic safety. To deal with the problem, this study applied a two-step method integrating latent class cluster analysis and random parameters logit model in analyzing motorcycle crash data from a typical province of China, i.e. Hunan.

The ROC curves confirm that applying LCA approach as a preliminary tool to segment the whole data into meaningful subsets before conducting rider injury severity analysis improves the model predictive accuracy. Comparing the general and cluster-based model results, several important findings are suggested: a) clustering can help reveal new information, including important contributing factors in subgroups which might be ignored in the pooled model, the ones with different influential magnitudes in cluster models, and factors showing opposite effects in the clustering sample. b) different contributing variables found in individual clusters and the whole data indicate that some factors are only influential under some specific conditions, such as no violation committed by riders, turning right prior to the crash, and single crashes, etc. c) clustering indeed has a great potential in reducing heterogeneity in the crash data and explaining the heterogeneity source. For example, the random effects of first-class highways are from different crash situations in cluster 1, 2 and 3. d) although observations in each cluster are relatively homogeneous in terms of some aspects (the characteristics extracted from the clustering sample), heterogeneity indicated by the estimated random parameters was still found in each subgroup, which supports the use of random parameters models.

From the practical perspective, the identified contributing factors that influence motorcycle crashes in specific cases can play a vital role in providing more reliable and detailed information for engineers, policy makers and planners to improve geometry, traffic control measures, traffic facilities, education, and enforcement actions. For example, single-vehicle crashes are not significant in the general model but significantly increase the probability of KSI on other types of road, on secondary- or lower-level highways, and in lighted darkness. This finding could provide guidance for policy makers that the management of roadside hazards such as barriers, posts, utility poles, etc. is necessary, for example, eliminating road hazards on other types of road in Changsha and secondary- and lower-level highways, and increasing the visibility of these roadside objects at night.

There are two limitations associated with the data used in this study. The data-processing procedure, for example, removing no-injury crashes and records with incomplete information may introduce bias in the analysis despite great efforts that have been made to avoid this issue. In addition, although many potential influential factors are included in the study, some predictors which may also have influences on rider injury levels are unavailable in police crash reports and hence not considered in this study, such as built environmental characteristics, collision speed, and traffic volume. In the future study, more complete police-reported crash data and the integration of police-reported data with other data sources (e.g. questionnaire surveys, field observations, driving simulations, and accident reconstruction simulation) are expected to achieve a more explicit understanding of the causal mechanism of motorcycle crashes.

Author Contribution Statement

The authors confirm contribution to the paper as follows: study conception and design: Pengpeng Xu, Fangrong Chang, and Helai Huang; data collection: Fangrong Chang and Helai Huang; data analysis and interpretation of results: Fangrong Chang, Hanchu Zhou; draft manuscript preparation: Fangrong Chang; manuscript revision: Pengpeng Xu, Helai Huang, and Alan H.S. Chan. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was jointly supported by: 1) the Joint Research Scheme of National Natural Science Foundation of China/Research Grants Council of Hong Kong (No. 71561167001 & N_HKU707/15); 2) National Key Research and Development Plan (No. 2018YFB1201601).

Appendix A

Table A1
The sample size of each cluster and percentage of all variables in 6 clusters.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
The number of observations	1926	1928	9987	2666	3537	4173
KSI	468	552	2172	672	827	776
Slight injury	1458	1376	7815	1994	2710	3397
Rider Age						
Under 18	4.72%	4.56%	5.23%	4.73%	5.20%	5.58%
18-24	6.39%	5.33%	6.89%	7.80%	6.22%	6.73%
25-34	13.76%	11.07%	13.43%	14.70%	12.81%	13.52%
35-44	17.29%	13.73%	17.63%	18.49%	17.61%	17.25%
45-54	33.90%	26.59%	33.04%	32.00%	34.41%	33.31%
55-64	17.91%	15.37%	18.81%	16.88%	17.98%	18.62%
Over 65	6.02%	4.92%	4.97%	5.40%	5.77%	4.98%
Rider gender						
Male	78.30%	85.50%	80.93%	83.01%	82.33%	78.94%
Female	21.70%	14.50%	19.07%	16.99%	17.67%	21.06%
Resident						
Urban resident	72.85%	73.62%	62.37%	79.93%	64.63%	81.72%
Rural resident	27.15%	26.38%	37.63%	20.07%	35.37%	18.28%
Illegal behavior						
No violations	18.22%	12.60%	17.70%	17.29%	14.65%	21.52%
Drunk riding	0.99%	4.41%	1.10%	5.48%	2.35%	1.63%
Violating traffic facilities	1.19%	0.82%	0.88%	5.21%	0.82%	5.58%
Failing to yield	5.76%	6.30%	7.01%	6.23%	8.31%	7.45%
Overtaking	1.30%	0.87%	2.46%	0.60%	1.78%	0.98%
Riding without helmets	1.19%	1.13%	1.54%	1.58%	1.72%	1.75%

(continued on next page)

Table A1 (*continued*)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Riding without licenses	38.01%	34.07%	44.72%	42.87%	44.70%	38.27%
Riding in the wrong direction	1.09%	1.74%	1.48%	3.75%	1.84%	3.59%
Approaching illegally	3.43%	2.66%	7.63%	0.71%	6.73%	0.74%
Following too close	1.25%	2.15%	1.63%	1.46%	1.95%	1.03%
Other violations	27.57%	14.81%	13.84%	14.82%	15.15%	17.45%
Passenger						
Without passenger	69.37%	76.13%	62.84%	64.03%	67.68%	69.21%
With passenger	30.63%	23.87%	37.16%	35.97%	32.32%	30.79%
Movement prior to crash						
Making U-turn	10.70%	5.38%	6.73%	6.38%	5.48%	6.21%
Turing left	9.35%	4.97%	13.41%	7.99%	9.58%	8.53%
Turning Right	2.28%	1.43%	4.68%	1.50%	2.86%	1.89%
Going Straight	77.67%	69.77%	75.19%	84.13%	82.08%	83.37%
Fault						
At-Fault	54.31%	72.64%	52.41%	58.51%	58.10%	50.18%
Non-Fault	45.69%	27.36%	47.59%	41.49%	41.90%	49.82%
Motorcycle type						
Motorcycle	95.74%	94.26%	92.38%	89.91%	92.48%	88.26%
Scooter	4.26%	5.74%	7.62%	10.09%	7.52%	11.74%
Road type						
First-class highways	23.21%	4.66%	1.44%	2.51%	3.65%	0.05%
Second-class or lower highways	0.05%	61.27%	98.56%	23.63%	96.35%	0.00%
General urban streets	0.00%	9.94%	0.00%	70.48%	0.00%	99.95%
Other types	76.74%	5.69%	0.00%	3.38%	0.00%	0.00%
Traffic Control measures						
Uncontrolled	57.17%	44.01%	63.44%	21.79%	58.30%	26.05%
Signal-controlled	3.79%	0.67%	0.78%	13.02%	0.90%	13.35%
Mark/sign-controlled	37.75%	36.01%	35.06%	64.59%	39.69%	59.62%
Other measures	1.30%	0.87%	0.72%	0.60%	1.10%	0.98%
Crash location						
Segment	89.82%	72.28%	83.40%	66.35%	86.54%	59.33%
Three-legged intersections	4.67%	5.79%	11.92%	13.32%	9.73%	15.98%
Four-legged intersections	5.50%	3.48%	4.69%	20.33%	3.73%	24.68%
Road surface						
Dry	85.41%	86.01%	83.04%	74.46%	83.18%	83.42%
Wet	14.59%	13.99%	16.96%	25.54%	16.82%	16.58%
Area						
Changsha	43.46%	13.37%	10.81%	18.30%	11.51%	21.52%
Changde	32.55%	12.65%	10.25%	9.94%	13.68%	9.30%
Chenzhou	0.36%	5.17%	12.15%	15.94%	12.27%	10.93%
Hengyang	0.36%	10.50%	19.80%	10.13%	16.03%	9.30%
Loudi	0.26%	2.00%	1.48%	3.34%	1.05%	3.93%
Shaoyang	0.16%	4.35%	6.81%	3.45%	5.80%	3.35%
Xiangtan	15.63%	4.87%	4.81%	5.48%	4.86%	5.97%
Xiangxi	0.00%	1.54%	1.39%	1.35%	1.55%	0.93%
Yueyang	0.57%	3.43%	5.39%	3.34%	4.72%	3.93%
Huaihua	0.26%	5.58%	5.03%	5.18%	6.33%	4.15%
Yongzhou	0.36%	4.76%	4.70%	5.14%	4.10%	7.26%
Zhangjiajie	5.40%	2.77%	3.01%	3.64%	3.19%	3.04%
Zhuzhou	0.16%	6.97%	12.36%	12.75%	12.02%	14.69%
Yiyang	0.47%	3.59%	2.02%	2.03%	2.88%	1.70%
Season						
Spring	25.70%	21.47%	24.44%	28.24%	22.87%	26.17%
Summer	30.63%	31.35%	30.27%	29.74%	27.62%	29.33%
Autumn	25.80%	16.03%	22.94%	23.78%	25.53%	24.18%
Winter	17.86%	12.70%	22.35%	18.23%	23.98%	20.32%
Weather						
Sunny	72.95%	52.25%	66.80%	58.66%	64.66%	68.75%
Cloudy or foggy	15.01%	18.49%	20.05%	21.79%	20.92%	18.45%
Rain or snow	12.05%	10.81%	13.16%	19.54%	14.42%	12.80%
Visibility						
Less than 50 m	9.45%	33.50%	7.11%	21.87%	21.15%	6.42%
50-100 m	21.96%	27.10%	25.76%	46.74%	34.01%	25.33%
100-200 m	23.73%	11.22%	23.10%	19.62%	20.53%	27.08%
Over 200 m	44.86%	9.73%	44.03%	11.78%	24.31%	41.17%
Lighting Condition						
Daylight	83.59%	0.00%	96.07%	0.00%	56.60%	89.48%
Twilight	8.20%	0.00%	3.80%	0.00%	18.77%	6.78%
Lighted dark	2.91%	0.00%	0.12%	100.00%	7.35%	0.00%
Complete darkness	5.30%	81.56%	0.00%	0.00%	17.27%	3.74%
Day of week						
Weekday	72.07%	76.38%	72.22%	71.83%	71.67%	72.39%
Weekend	27.93%	23.62%	27.78%	28.17%	28.33%	27.61%
Time of day						
00:00-06:59	7.94%	27.20%	3.57%	27.12%	1.55%	5.18%

(*continued on next page*)

Table A1 (*continued*)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
07:00-08:59 (morning peak)	12.67%	0.00%	17.49%	0.00%	0.00%	14.33%
09:00-11:59	19.42%	0.00%	28.30%	0.00%	0.00%	18.81%
12:00-16:59	37.23%	0.00%	50.64%	0.00%	0.00%	39.37%
17:00-19:59 (evening peak)	22.69%	0.15%	0.00%	12.57%	98.16%	22.31%
20:00-11:59	0.05%	54.20%	0.00%	60.32%	0.28%	0.00%
Collision objects						
Single vehicle	50.31%	42.21%	47.38%	48.99%	49.59%	46.82%
Pedestrians	1.82%	6.61%	1.12%	4.13%	5.20%	1.65%
Non-motor vehicles	2.70%	1.18%	1.12%	2.18%	1.47%	1.41%
Light motor vehicles	42.16%	26.13%	45.81%	40.36%	39.78%	46.54%
Heavy motor vehicles	2.18%	4.30%	3.10%	3.04%	2.80%	2.25%
Multi-vehicle crashes	0.83%	1.13%	1.46%	1.31%	1.16%	1.32%

References

Ahmed, A., Sadullah, A.F.M., Yahya, A.S., 2017. Errors in accident data, its types, causes and methods of rectification-analysis of the literature. Accid. Anal. Prev in press.

Albalate, D., Fernandez-Villadangos, L., 2010. Motorcycle injury severity in Barcelona: the role of vehicle type and congestion. Traffic Inj. Prev. 11 (6), 623–631.

Biernacki, C., Govaert, G., 1999. Choosing models in model-based clustering and discriminant analysis. J. Stat. Comput. Simul. 64 (1), 49–71.

Bijmolt, T.H., Paas, L.J., Vermunt, J.K., 2004. Country and consumer segmentation: multi-level latent class analysis of financial product ownership. Int. J. Res. Mark. 21 (4), 323–340.

Chang, F., Li, M., Xu, P., et al., 2016. Injury severity of motorcycle riders involved in traffic crashes in Hunan, China: a mixed ordered logit approach. Int. J. Environ. Res. Public Health 13 (7), 714.

Chang, F., Xu, P., Zhou, H., et al., 2019. Identifying motorcycle high-risk traffic scenarios through interactive analysis of driver behavior and traffic characteristics. Transp. Res. Part F Traffic Psychol. Behav. 62, 844–854.

Collins, L.M., Lanza, S.T., 2010. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. John Wiley & Sons.

De Lapparent, M., 2006. Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas. Accid. Anal. Prev. 38 (2), 260–268.

Deffenbacher, J.L., Getting, E.R., Lynch, R.S., 1994. Development of a driving anger scale. Psychol. Rep. 74, 83–91.

de Ona, J., López, G., Mujalli, R., et al., 2013. Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks. Accid. Anal. Prev. 51, 1–10.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40 (4), 1257–1266.

Elvik, R., Mysen, A., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transportation Research Record: Journal of the Transportation Research Board 1665, 133–140.

Hair, J.F., Black, W.C., 1998. Multivariate Data Analysis. Upper Saddle River.

Haque, M., Chin, H., Huang, H., 2008. Examining exposure of motorcycles at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board 2048, 60–65.

Hunan Provincial Bureau of Statistics, 2017. Hunan Statistical Year Book 2017. China Statistics Press.

Kasantikul, V., Ouellet, J.V., Smith, T., et al., 2005. The role of alcohol in Thailand motorcycle crashes. Accid. Anal. Prev. 37 (2), 357–366.

Lanza, S.T., Dziak, J.J., Huang, L., et al., 2015. LCA Stata Plugin Users' Guide (version 1.2). The Methodology Center, Penn State, University Park.

Lanza, S.T., Rhoades, B.L., 2013. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. Prev. Sci. 14 (2), 157–168.

Lin, M.R., Chang, S.H., Huang, W., et al., 2003. Factors associated with severity of motorcycle injuries among young adult riders. Ann. Emerg. Med. 41 (6), 783–791.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.

Meng, F., Xu, P., Wong, S.C., et al., 2017. Occupant-level injury severity analyses for taxi in Hong Kong: a Bayesian space-time logistic model. Accid. Anal. Prev. 108, 297–307.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., et al., 2013. A clustering regression approach: a comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Saf. Sci. 54, 27–37.

Moore, D.N., Schneider Iv, W.H., Savolainen, P.T., et al., 2011. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accid. Anal. Prev. 43 (3), 621–630.

National Bureau of Statistics, 2018. China Statistics Yearbook 2018. China Statistics Press.

National Highway Traffic Safety Administration, 2018. Traffic Safety Facts, 2016 Data. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812492.

Norvell, D.C., Cummings, P., 2002. Association of helmet use with death in motorcycle crashes: a matched-pair cohort study. Am. J. Epidemiol. 156 (5), 483–487.

Pai, C.W., 2009. Motorcyclist injury severity in angle crashes at T-junctions: identifying significant factors and analysing what made motorists fail to yield to motorcycles. Saf. Sci. 47 (8), 1097–1106.

Pai, C.W., Saleh, W., 2007. An analysis of motorcyclist injury severity under various traffic control measures at three-legged junctions in the UK. Saf. Sci. 45 (8), 832–847.

Pai, C.W., Saleh, W., 2008. Modelling motorcyclist injury severity by various crash types at T-junctions in the UK. Saf. Sci. 46 (8), 1234–1247.

Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. J. Safety Res. 33 (4), 445–462.

Rowland, J., Rivara, F., Salzberg, P., et al., 1996. Motorcycle helmet use and injury outcome and hospitalization costs from crashes in Washington state. Am. J. Public Health 86 (1), 41–45.

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.

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes. Accid. Anal. Prev. 39 (5), 955–963.

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

Shaheed, M.S.B., Gkritza, K., Zhang, W., et al., 2013. A mixed logit analysis of two-vehicle crash severities involving a motorcycle. Accid. Anal. Prev. 61, 119–128.

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

The National People's Congress of the People's Republic of China, 2017. Cooperative Governance of People, Vehicles and Roads to Ensure Road Traffic Safety. Available from: [cited 2018 Sep 26]. http://www.npc.gov.cn/npc/zgrdzz/2017-05/04/content_2021172.htm.

Train, K.E., 2009. Discrete Choice Methods With Simulation Cambridge University Press.

Tsai, V.W., Anderson, C.L., Vaca, F.E., 2008. Young female drivers in fatal crashes: recent trends, 1995–2004. Traffic Inj. Prev. 9 (1), 65–69.

Ulfarsson, G.F., Mannering, F.L., 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accid. Anal. Prev. 36 (2), 135–147.

Valent, F., Schiava, F., Savonitto, C., et al., 2002. Risk factors for fatal road traffic accidents in Udine. Italy. Accident Analysis & Prevention 34 (1), 71–84.

Wang, X.S., Abdel-Aty, M., 2008. Modeling left-turn crash occurrence at signalized intersections by conflicting patterns. Accid. Anal. Prev. 40 (1), 76–88.

Waseem, M., Ahmed, A., Saeed, T.U., 2019. Factors affecting motorcyclists' injury severities: an empirical assessment using random parameters logit model with heterogeneity in means and variances. Accid. Anal. Prev. 123, 12–19.

World Health Organization, 2018. Global Status Report on Road Safety. Available online: Accessed December 11, 2018. https://www.who.int/violence_injury_prevention/road_safety_status/2018/English-Summary-GSRRS2018.pdf?ua=1.

Xiao, Y., Huang, H., Peng, Y., Wang, Q.H., 2018. A study on motorcyclists head injuries in car- motorcycle accidents based on real-world data and accident reconstruction. J. Mech. Med. Biol. 18 (04), 1850036.

Yau, K.K.W., 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accid. Anal. Prev. 36 (3), 333–340.

Zhang, T., Chan, A.H., Zhang, W., 2015. Dimensions of driving anger and their relationships with aberrant driving. Accid. Anal. Prev. 81, 124–133.