Accident Analysis and Prevention 124 (2019) 50–57

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Pedestrian crash analysis with latent class clustering method

Ming Sun[a,*], Xiaoduan Sun[a], Donghui Shan[b]

[a] Department of Civil Engineering, University of Louisiana at Lafayette, Madison Hall 254Q, 131 Rex Street, Lafayette, LA 70504, United States
[b] CCCC First Highway Consultants Co., Ltd, 205 Keji 4th Rd, Yanta District, Xian, Shaanxi, 710065, China

ARTICLE INFO

Keywords:
Pedestrian safety
Latent class
Cluster analysis
Severity
Contributing factors

ABSTRACT

Pedestrians are the most vulnerable users of the highway transportation system. While encouraging "Green Transportation", a concerning fact emerges in the United States: pedestrian deaths are climbing faster than motorist fatalities, reaching nearly 6000 in 2016 - the highest in over two decades. In 2015, pedestrian fatalities reached 110, 14.6% of total traffic fatalities in Louisiana for that year. Consequently, the Louisiana pedestrian fatality rate (fatalities per 100,000 population) was 2.18, exceeding the U.S. average of 1.67. In an effort to effectively reduce the pedestrian crashes, this paper investigates this problem for Louisiana. However, with the heterogeneity of provided crash data, it is difficult to identify major causation that contribute to these crashes. This study will reveal the findings of the Latent Class Cluster (LCC) model, utilizing it as a preliminary tool for the segmentation of 14,236 pedestrian crashes in Louisiana, between the years of 2006–2015. Next, Multinomial Logit (MNL) models are used to identify the main factors in pedestrian crash severity, shown in the original dataset, by further analyzing the clusters previously obtained by the LCC model. The results shed lights on the crash characteristics that are not apparent without these combined data analysis methods. Certain variables that have not been identified as significant in whole data analysis are identified as significant for a specific cluster, such as pedestrian crossing and entering roads, crash hours between midnight to 6 pm, dark-unlighted conditions, dark-lighted conditions, and non-intersection locations. The study suggests that the LCC regression approach can reveal important, formerly hidden relationships in traffic safety analyses.

1. Introduction

Pedestrians are the most vulnerable roadway users of the highway transportation system. In 2016, there were 5987 pedestrian fatalities across the United States, nearly 16 pedestrian deaths every day, which represents the highest annual record of pedestrian traffic fatalities since 1990 (NHTSA, 2017; Retting, 2017). Though the total traffic deaths decreased by 9.2% from 2007 to 2016, the pedestrian fatalities increased by 27.4% at the same time as shown in Fig. 1 (Source: Traffic Safety Facts, 2016 Data, Pedestrians, NHTSA, 2007 data is used as baseline condition). The year of 2016 was the third consecutive year in which pedestrians accounted for 15% or higher of the total traffic fatalities as shown in Fig. 2 (Source: Traffic Safety Facts, 2016 Data, Pedestrians, NHTSA).

Pedestrian safety has been a long-standing problem in Louisiana. Although the total traffic deaths have declined significantly over a 10-year period (2006–2015), the progress in reducing pedestrian fatalities has been much less significant than that for total traffic fatalities. The pedestrian fatalities made up 14.6% of all traffic deaths in 2015 (Crash Data, Louisiana Department of Transportation and Development, 2018). 14,236 pedestrian crashes were reported to the police in the state during the 10-year period, and 1007, or 7.1%, were fatal crashes. Additionally, Louisiana experiences a significantly higher pedestrian fatalities rate (fatalities per 100,000 population) compared to the national average as illustrated in Fig. 3 (Source: Traffic Safety Facts, 2006–2015 Data, Pedestrians, NHTSA). The state has been identified as one of the most dangerous states for pedestrians (NHTSA, 2017; Smart Growth America, 2014).

To effectively reduce pedestrian crashes and improve pedestrian safety, it is important to identify why, where, and how pedestrian crashes occurred. Researchers usually rely on crash data to identify possible risk factors associated with crash frequency and severity. However, pedestrian crash data is too heterogeneous in nature, making it difficult to identify certain patterns. Previous researchers have tried to divide crash data into different subgroups, using crash location, time of day, roadway geometry, lighting conditions, and other inherent roadway characteristics. This data segmentation can help to develop and implement statistical model for pedestrian crash analysis, but the sample sizes varies greatly among the groups, which does not guarantee consistency of a homogenous group (Depaire et al., 2008). Data mining

* Corresponding author.
E-mail addresses: mxs1278@louisiana.edu (M. Sun), xsun@louisiana.edu (X. Sun), sdhcjj@126.com (D. Shan).

https://doi.org/10.1016/j.aap.2018.12.016
Received 5 June 2018; Received in revised form 19 November 2018; Accepted 20 December 2018
Available online 07 January 2019
0001-4575/

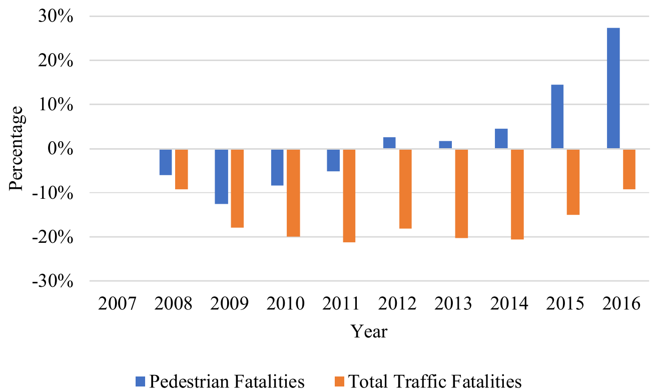

Fig. 1. Percentage change of pedestrian fatalities vs total traffic fatalities (2007–2016).

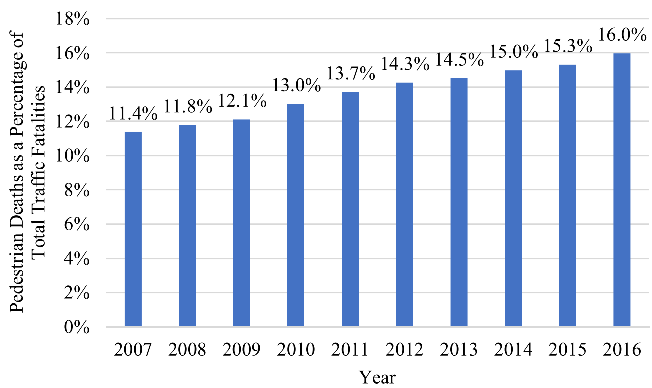

Fig. 2. Pedestrian fatalities as a percentage of total fatalities in the U.S. (2007–2016).

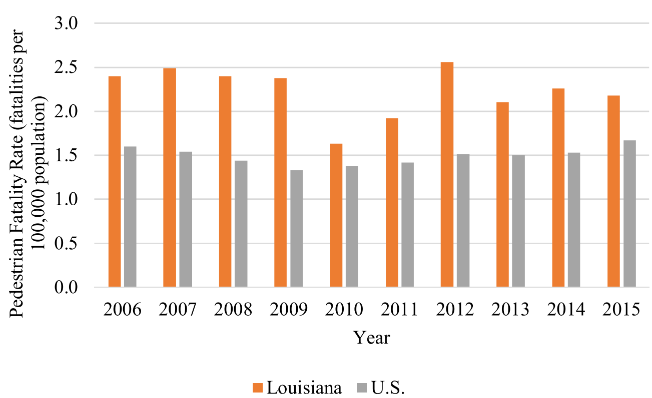

Fig. 3. Pedestrian fatality rate (per 100,000 population) Louisiana versus U.S. average (2006–2015).

techniques such as cluster analysis can be used as a preliminary tool in reducing the heterogeneity of pedestrian crash grouping.

2. Literature review

Pedestrian crash analysis has been conducted by many studies in the past. Investigating the characteristics of pedestrian crashes, identifying the risk factors related to the injury severity levels, and predicting pedestrian crashes through models were the issues focused on in these studies. Existing research has found that a wide range of factors may contribute to the occurrence and severity of pedestrian crashes, including behavioral, roadway, and environmental factors.

Haleem et al. (2015) found that crashes where pedestrians are at fault are more likely to result in severe injury or fatality. Lee and Abdel-Aty (2005) found that pedestrians between 15 and 24 years of age are less likely to suffer severe injuries in crashes, while older pedestrians are more likely to sustain severe injuries, as a result of weaker physical conditions. Ivan et al. (2000) used ordered probit models to evaluate the factors influencing pedestrian injury severity in rural Connecticut. They found that vehicle type, driver alcohol involvement, pedestrian alcohol involvement, and pedestrian age over 65 significantly increased pedestrian injury severity in these rural crashes. Roadway characteristics such as the functional classification, number of lanes, and the prevalent lighting conditions, were also observed to show an effect on pedestrian crash severity. Another study (Islam and Jones, 2014) found that pedestrian crash severity is influenced by dark lighting conditions and the number of lanes (i.e. two-lane roadways). Additionally, other research (Rothman et al., 2012) found that pedestrian crashes occurring at uncontrolled mid-block locations tend to be more severe than those occurring at controlled locations.

There are several statistical methods that can be used for analyzing the crash severity, such as binary logit models (Sze and Wong, 2007), ordered logit or probit models (Ivan et al., 2000; Lee and Abdel-Aty, 2005), mixed logit models (Islam and Jones, 2014; Haleem et al., 2015), and multinomial logit models (Rothman et al., 2012). Recently, data mining techniques such as clustering, classification and association rule mining have been used for crash data exploration and safety analysis. Kashani and Mohaymany, (2011) used a classification and regression tree (CART) to analyze the road accident data of Iran, finding that not using seat belts, improper overtaking, and excessive speeding affect the severity of accidents. Kweon and Kockelman, (2003) used Naive Bayes and a decision tree classification algorithm to analyze factor dependencies related to road safety. This research suggested the severity of accidents is directly concerned with the victim involved in the accident, and its analysis surrounds the injury, attempting to determine the factors responsible for its type and severity. Some researchers used k-means clustering algorithms to identify homogenous crash clusters (Kim and Yamashita, 2007; Mohamed et al., 2013; Hamzehei et al., 2014). Depaire et al. (2008) used LCC and MNL models to study the severity of traffic accidents. In their study, they identified seven clusters that represented different types of traffic accidents. Subsequently, they applied an MNL model to the full set of data, as well as each of the seven identified clusters. Their results showed that the clustered data provided information that would not have been obtained if only the full database had been used. The LCC model has also been used to segment databases and analyze vehicle crash data (Park and Lord, 2009; Park et al., 2010) and pedestrian crash data (Mohamed et al., 2013; Sasidharan et al., 2015).

The studies above suggested that pedestrian crash segmentation is associated with crash severities, and using the whole data involving all crashes without distinguishing the subgroups of crash data, makes the data heterogeneous. One of the implications is that safety countermeasures needed for preventions are different for each individual level of the injury severity. An accurate understanding of the factors and conditions, that play key roles in contributing to these crashes, is vital, because traffic engineers, policy makers, and planners rely on this information to determine the most effective safety countermeasures. Options include geometric improvement, traffic control measures, dedicated pedestrian facilities, modifying land use, educational actions, and enforcement actions. Therefore, an attempt is made to "see beyond" the systematic heterogeneity of the gathered data on pedestrian crashes in Louisiana by using an LCC model approach. Subsequently, an MNL model is used on each of the identified latent clusters to further identify different crash-contributing factors responsible for pedestrian injury severities.

3. Methodology

3.1. Latent class clustering analysis

Clustering is an unsupervised data mining technique where the main task is to group the data objects into different clusters, each displaying common traits throughout the data from which it is composed. This similarity-based technique includes two main approaches: the hierarchical approach (e.g. Ward's method, a single linkage method) and the partitioning approach (e.g. k-means). In this study, we focus on the partitioning approach, which divides the data into k-clusters with no hierarchical relationship. A general framework for categorical data analysis with discrete latent variables was proposed by Hagenaars (1990) and extended by Vermunt (1997). LCC analysis can be used as a probabilistic cluster analysis tool for crash data, an approach that offers many advantages over k-means clustering. These include:

- Being able to assign probability of a particular crash to belong in a specific cluster, by using subsequent membership probabilities estimated with a maximum likelihood method.
- Does not require the number of clusters to be predetermined.
- Allows for using different types of variables (numerical, nominal or a combination of them), without a prior standardization process.

In this study, the LCA package from Latent GOLD 4.5 is used to identify latent classes. Given a data sample of N crashes, let X represents the latent variable and Y_l one of the L observed variables, where $1 \leq l \leq L$. These variables form a latent class model with K classes. Suppose each observed value contains a specific number of levels D_l, a particular latent class is enumerated by the index x, $x = 1, 2, ..., K$ and a particular value of Y_l by y_l, $y_l = 1, 2, ..., D_l$. The basic LC cluster form is (Vermunt and Magidson, 2002):

$$\boldsymbol{P_{Y_l} = \sum_{k=1}^{K} P_{X_k} P_{Y_l|X_k}} \tag{1}$$

Here, P denotes probability of obtaining response pattern y; $P\ (X_k)$ represents the probability that a particular crash belongs to the latent k class, with $k = 1, 2, ..., K$; $P_{Y_l|X_k}$ is the conditional probability that a particular crash has a response pattern $Y_l = (y_1, ..., y_L)$, given its membership in the k class of latent variable X. The latent class model assumes that the L observed variables are mutually independent with each latent class, so the above equation can be re-written as:

$$P_{Y_l} = \sum_{k=1}^{K} P_{X_k} \prod_{l=1}^{L} P_{Y_l|X_k} \tag{2}$$

The parameters are estimated using a maximum likelihood algorithm. After model estimation, the probability of belonging to latent class x, which is often referred to as the posterior membership probability, can be obtained by the Bayes rule,

$$P_{X_k|Y_l} = \frac{P_{X_k} P_{Y_l|X_k}}{P_{Y_l}} \tag{3}$$

3.2. Number of clusters identification

The goodness-of-fit of an estimated LC model is usually tested by the Pearson or the likelihood-ratio chi-squared statistic, but in such cases, the frequency table may become very sparse and, as a result, asymptotic p values can no longer be trusted. An elegant, but somewhat time-consuming solution to this problem is to estimate the p values by parametric bootstrapping. One popular method is the use of information criteria such as Bayesian Information Criterion (BIC) (Raftery, 1986), Akaike Information Criterion (AIC) (Akaike, 1987) and Consistent Akaike Information Criterion (CAIC) (Fraley and Raftery, 1998). The appropriate number of clusters is determined by the one that minimizes the score of these criteria.

3.3. Severity model

When the crash severity outcomes are three or more, the usual logistic regression cannot be used. Therefore, Multinomial Logistic Regression (MNL) is a useful alternative for this scenario, in which the dependent variable is not limited to two categories. It is the most widely applied discrete-outcome modeling approach for crash severity analysis (Zhang et al., 2000; Bedard et al., 2002; Al-Ghamdi, 2002; Islam and Mannering, 2006; Ulfarsson and Mannering, 2004; Kim et al., 2007; Carson and Mannering, 2001; Shankar and Mannering, 1996; Valent et al., 2002; Yau, 2004.) Our injury risk analysis mainly follows the work of Ulfarsson and Mannering (2004); Islam and Mannering (2006), and Kim et al. (2007).

The general framework used to model the degree of injury severity sustained by a crash-involved individual begins by defining a linear function S that determines the injury outcome i for observation n as,

$$S_{in} = \beta_i X_{in} + \varepsilon_{in} \tag{4}$$

Where β_i is a vector of estimable coefficients, X_{in} is a vector of observable characteristics (pedestrian, roadway, and environmental factors) that impact the pedestrian injury severity sustained by observation n, and ε_{in} is a disturbance term that accounts for unobserved effects. If the disturbance terms are assumed to be independently and identically distributed as generalized extreme value distributed, the multinomial logit model results (see McFadden, 1981),

$$P_n(i) = \frac{exp(\beta_i X_{in})}{\sum_i exp(\beta_i X_{in})} \tag{5}$$

The MNL model does not account for ordinal nature of the independent variable, but allows independent variables' effects to vary among outcome levels. Specifically, it estimates a series of binary models, where one level of dependent variables is known as reference.

3.4. Variable selection

Crash data was obtained from the Louisiana Department of Transportation and Development (LADOTD) highway crash list database. The research team retrieved all pedestrian-related crashes for the period between 2006 and 2015. The crash database contains many variables, some of which are redundant in nature, such as pedestrian city, state, zip code, etc. These variables were omitted from the analysis. To focus on the main objective of this study, namely, identifying the contributing factors that affect the severity of pedestrian crashes, the research team identified a list of significant variables present from the past studies (Ivan et al., 2000;Lee and Abdel-Aty, 2005 Rothman et al., 2012; Mohamed et al., 2013; Islam and Jones., 2014; Haleem et al., 2015; Sasidharan et al., 2015), such as pedestrian-related factors (pedestrian age, gender, alcohol and drugs presence condition, pedestrian action, and injury severity), crash-related factors (crash hour, crash location: intersection or non-intersection and posted speed), and environmental factors (season, weather, lighting conditions, and setting: rural or urban).

4. Data

The preliminary dataset was prepared from ten years (2006–2015) of pedestrian crash data within statewide from the LADOTD highway crash list database. It was developed by merging three different datasets (crash data, pedestrian data, and vehicle data) from the Microsoft Access dataset. Each crash has a unique identification number that is common in different datasets to link records together.

A total number of 14,236 pedestrian crashes were collected. LADOTD's crash database uses an ABCDE scale to describe severity level of pedestrian crashes. 'A' indicates fatal injury, 'B', 'C' and 'D' indicate incapacitating or severe injury, non-incapacitating or moderate

Table 1
Characteristics of pedestrian crashes by severity.

Variables	Total crashes	Fatal/ severe crashes	Injury crashes	No injury crashes	Total
Pedestrian alcohol and drugs presence					
Neither alcohol nor drugs	8,475	7.26%	81.05%	11.69%	100%
Either alcohol or drugs	1,337	25.58%	67.24%	7.18%	100%
Other	4,424	16.23%	50.38%	33.39%	100%
Pedestrian action					
Crossing, entering road	6,604	11.57%	72.56%	15.87%	100%
Walking in road	2,079	14.53%	70.37%	15.10%	100%
Other inappropriate behavior	2,843	10.76%	72.70%	16.53%	100%
Other	2,710	11.18%	61.81%	27.01%	100%
Pedestrian age group					
< 15	4,017	5.75%	53.67%	40.58%	100%
15-25	2,621	11.48%	81.99%	6.52%	100%
25-40	2,858	15.57%	77.54%	6.89%	100%
40-65	3,719	15.51%	77.90%	6.59%	100%
> 65	620	15.97%	79.03%	5.00%	100%
Other	401	5.49%	22.19%	72.32%	100%
Pedestrian gender					
Male	8,219	14.47%	77.21%	8.32%	100%
Female	4,354	11.02%	82.84%	6.13%	100%
Unknown	1,663	0.36%	2.65%	96.99%	100%
Crash month					
Spring	4,015	10.09%	60.65%	29.27%	100%
Summer	3,530	11.27%	72.78%	15.95%	100%
Autumn	3,155	12.46%	75.25%	12.30%	100%
Winter	3,536	13.55%	74.07%	12.39%	100%
Crash hour					
Midnight-6am	2,439	14.10%	41.53%	44.36%	100%
6am-noon	2,413	8.12%	77.79%	14.09%	100%
Noon-6 pm	4,591	8.45%	77.11%	14.44%	100%
6 pm-midnight	4,793	15.59%	74.42%	9.99%	100%
Lighting condition					
Daylight	7,104	7.74%	78.03%	14.23%	100%
Dark-unlighted	1,635	27.71%	67.16%	5.14%	100%
Dark-lighted	4,000	15.05%	72.68%	12.28%	100%
Dusk/dawn	430	12.56%	77.67%	9.77%	100%
Other	1,067	1.50%	10.78%	87.72%	100%
Intersection					
Non-intersection	9,656	12.95%	70.72%	16.33%	100%
Intersection	4,580	9.28%	69.17%	21.55%	100%
Weather					
Clear	10,356	12.31%	75.24%	12.45%	100%
Cloudy	1,963	13.50%	75.09%	11.41%	100%
Rain	821	13.89%	71.74%	14.37%	100%
Snow	7	14.29%	85.71%	0.00%	100%
Other	85	15.29%	77.65%	7.06%	100%
Unknown	1,004	0.70%	6.97%	92.33%	100%
Posted speed					
< = 30	7,388	7.11%	68.71%	24.19%	100%
30-60	5,933	16.69%	72.36%	10.96%	100%
> = 60	250	38.00%	58.80%	3.20%	100%
Other	665	9.77%	72.33%	17.89%	100%
Setting					
Rural	3,995	18.87%	71.66%	9.46%	100%
Urban	10,241	8.99%	69.66%	21.35%	100%

injury, and possible or compliant injury, respectively. ‘E’ represents that there were no injuries occurred in the crashes. Based on this scale, there are 1007 (7.1%) fatal crashes, 1197 (8.4%) severe crashes, 5039 (35.4%) moderate crashes, 4958 (34.8%) possible crashes, and 2035 (14.3%) no injury crashes. In this study, the A, B levels of severity are combined as fatal and severe crashes, C, D levels of severity are combined as injury crashes. A comprehensive dataset containing information on crash, pedestrian and environment characteristics was created. Table 1 gives an overview of the descriptive statistics of pedestrian crashes and all variables used for this study.

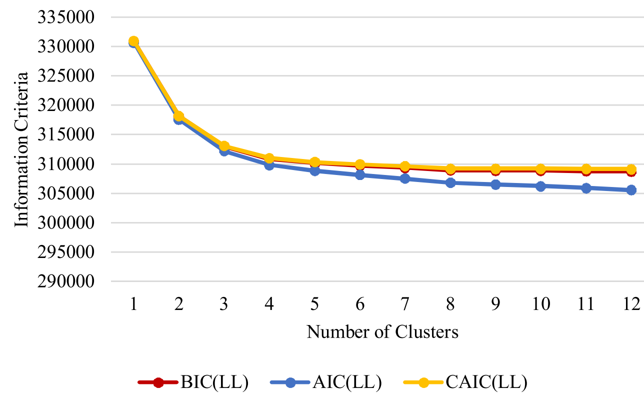

Fig. 4. Number of clusters identification for pedestrian crash analysis.

5. Results

5.1. Latent class analysis

Pedestrian crashes were clustered by variables in Table 1. To select the appropriate number of clusters, different numbers of clusters were tested, from one to twelve. The BIC, AIC, and CAIC criteria were used to select the final number of clusters. As shown in Fig. 4, all three criteria decrease as the number of clusters increases. BIC is more reliable than AIC, especially for large datasets (Biernacki and Govaert, 1999; Vermunt and Magidson, 2002). Since the BIC and other information criteria often do not reach a minimum value with an increasing number of clusters (Bijmolt et al., 2004), the percentage difference in BIC values was computed for different models. The results show that the percentage decrease in BIC drops to less than 1% from six clusters onwards. Furthermore, the quality of the clustering solution was assessed by calculating the entropy R squared criterion (McLachlan and Peel, 2000). The closer the criterion is to 1, the better the clustering. The entropy value estimated for six clusters was 0.90, which indicates a clear separation between the clusters identified. The AIC and CAIC values also support dividing the data into six clusters. Therefore, the pedestrian crash data was divided into six clusters for further analysis.

The final model was described by the proportion of each variable in each cluster. Much like the work of (Depaire et al., 2008), the clusters were analyzed and given names based on their variable distributions. For example, if one cluster has 95% of its crashes occurring in autumn, while the other clusters have balanced distribution over the season variable, this cluster would be the cluster of accidents happening in autumn.

The cluster profiles are shown in Table 2. For cluster 1, the variables are injury crash, crash hour between noon to 6 pm, and daylight conditions. In terms of severity, injury crashes represent approximately 81% of the crashes in this cluster. For the time of day, the crash occurs between noon to 6 pm in 58% of the cases. The lighting condition in this cluster is daylight for approximately 96% of the cases. Consequently, we referred to cluster 1 as “Crashes in daytime”. The other clusters were classified similarly. Cluster 2 resembles cluster 1 for injury crashes, but distinguishes itself by an over-representation of dark conditions and neither alcohol nor drugs involvement for pedestrians. Cluster 3 reveals a majority of the pedestrian alcohol or drugs involvement and crossing road behavior during nighttime. The special features of cluster 4 are the rural crashes at non-intersections. Two variables are specific to cluster 5: no injury (97.75%) and urban (90.41%). Finally, cluster 6 contains only about 3% of all data and covers the unknown or unreported values of different variables. This cluster shows the power of clustering as a pre-processing technique to cluster the missing data.

To summarize, the clustering is useful to segment the dataset into more homogeneous groups, allowing identification of the higher order variables that may have an influence on injury severity. Table 3 shows

Table 2
Summary of variables and their distribution in each cluster.

Variables	Whole dataset	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
Pedestrian alcohol and drugs presence							
Neither alcohol nor drugs	61.90%	73.58%	81.98%	10.20%	45.59%	60.06%	39.31%
Either alcohol or drugs	10.40%	3.71%	0.52%	36.08%	30.49%	2.52%	8.84%
Other	27.70%	22.70%	17.50%	53.72%	23.92%	37.42%	51.84%
Pedestrian action							
Crossing, entering road	46.50%	50.70%	47.01%	55.27%	20.56%	36.65%	39.94%
Walking in road	14.50%	10.42%	15.53%	13.36%	37.37%	6.92%	11.65%
Other inappropriate behavior	19.90%	21.56%	23.02%	9.44%	24.76%	13.07%	14.82%
Other	19.10%	17.32%	14.44%	21.92%	17.32%	43.37%	33.60%
Pedestrian age group							
< 15	23.30%	26.98%	20.80%	3.01%	2.99%	72.96%	65.64%
15-25	19.10%	19.34%	27.99%	13.61%	23.30%	0.01%	0.35%
25-40	21.40%	18.90%	22.63%	29.52%	35.52%	0.32%	0.55%
40-65	28.30%	27.62%	23.96%	48.44%	33.77%	0.68%	0.62%
> 65	4.80%	6.35%	4.14%	3.46%	3.72%	0.00%	0.00%
Other	3.00%	0.82%	0.47%	1.97%	0.71%	26.04%	32.84%
Severity							
No injury	14.40%	9.00%	6.98%	5.01%	5.00%	97.75%	96.47%
Injury	70.00%	81.07%	85.48%	53.65%	60.85%	1.61%	2.09%
Fatal/severe	15.60%	9.93%	7.53%	41.33%	34.15%	0.64%	1.44%
Pedestrian gender							
Male	61.80%	63.03%	60.18%	73.54%	76.43%	15.82%	19.60%
Female	32.10%	36.70%	39.59%	26.01%	23.47%	5.17%	7.22%
Unknown	6.10%	0.27%	0.23%	0.45%	0.10%	79.02%	73.18%
Crash month							
Spring	24.70%	23.85%	25.92%	25.84%	25.65%	23.13%	25.75%
Summer	25.20%	28.42%	21.78%	21.84%	21.20%	27.15%	21.07%
Autumn	23.50%	24.39%	21.01%	22.80%	25.44%	22.60%	23.62%
Winter	26.50%	23.34%	31.29%	29.52%	27.71%	27.12%	29.56%
Crash hour							
Midnight-6am	11.90%	0.42%	17.64%	32.03%	29.78%	0.95%	27.18%
6am-noon	18.00%	30.85%	2.93%	2.31%	4.18%	29.05%	2.05%
Noon-6 pm	33.80%	57.64%	7.34%	4.17%	3.56%	60.59%	4.62%
6 pm-midnight	36.40%	11.09%	72.09%	61.49%	62.48%	9.40%	66.15%
Lighting							
Daylight	52.40%	95.78%	0.06%	2.56%	2.42%	94.92%	1.57%
Dark-unlighted	12.90%	0.01%	12.70%	12.49%	79.41%	0.06%	18.25%
Dark-lighted	30.20%	0.01%	80.66%	81.46%	15.15%	1.46%	73.39%
Dusk/dawn	3.20%	3.26%	4.97%	1.34%	3.02%	1.58%	4.25%
Other	1.20%	0.94%	1.62%	2.14%	0.00%	1.99%	2.54%
Intersection							
Non-intersection	66.50%	64.35%	63.65%	64.41%	93.75%	49.29%	59.42%
Intersection	33.50%	35.65%	36.35%	35.59%	6.25%	50.71%	40.58%
Weather							
Clear	77.80%	78.77%	77.45%	74.05%	78.73%	80.15%	74.54%
Cloudy	14.70%	15.56%	13.31%	14.81%	13.93%	12.93%	14.86%
Rain	6.00%	4.76%	7.64%	8.77%	5.43%	5.08%	8.15%
Snow	0.10%	0.06%	0.05%	0.16%	0.00%	0.00%	0.00%
Other	0.60%	0.30%	0.80%	0.71%	1.90%	0.43%	1.00%
Unknown	0.70%	0.56%	0.74%	1.50%	0.00%	1.42%	1.45%
Posted speed							
< = 30	47.70%	57.17%	55.26%	31.60%	7.11%	54.50%	48.70%
30-60	45.40%	37.26%	39.53%	63.17%	74.57%	37.57%	44.75%
> = 60	2.00%	1.01%	0.00%	0.61%	13.82%	0.52%	0.00%
Other	4.90%	4.56%	5.21%	4.62%	4.50%	7.41%	6.55%
Setting							
Rural	29.30%	26.98%	24.57%	15.82%	79.19%	9.59%	14.61%
Urban	70.70%	73.02%	75.43%	84.18%	20.81%	90.41%	85.39%

Table 3
Cluster summary.

Cluster No.	Proportion of whole dataset	Number of crashes
Cluster 1	49.57%	7,057
Cluster 2	18.74%	2,668
Cluster 3	13.62%	1,939
Cluster 4	10.42%	1,483
Cluster 5	4.53%	645
Cluster 6	3.13%	446

an overview of the cluster descriptions and the size of each cluster.

5.2. Injury severity analysis using MNL

The goal of this study is to explore the variables influencing pedestrian crash severity. The MNL model was applied in which the severity outputs were considered as the dependent variable. We selected the injury consequences for the pedestrian crashes, which typically fell in to three categories: fatal and severe crash, injury crash (moderate or possible injury) and no injury crash (property damage only). In all models, no injury crash was selected as the base outcome. 11 explanatory variables were entered into the model, including: pedestrian

Table 4
MNL model estimation results for pedestrian crashes – whole data and clusters.

Reference group: no injury	Whole dataset		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
Variables	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.
Fatal and severe crash												
Neither alcohol nor drugs *(Ref. Either alcohol or drugs)*	−1.301	0.000	−2.506	0.000							−1.504	0.001
Crossing, entering road *(Ref. Other inappropriate behavior)*							1.029	0.058			0.960	0.071
Walking in road							0.897	0.010				
Age < 15 *(Ref. Age > 65)*			−1.649	0.001								
Age 15-25	−1.150	0.001	−1.342	0.000								
Age 25-40	−1.188	0.000	−1.622	0.000	−1.686	0.012						
Age 40-65	−0.820	0.013	−0.806	0.028	−1.167	0.070						
Male *(Ref. Female)*	−0.522	0.000	−0.461	0.017					−0.339	0.007	−0.907	0.089
Spring *(Ref. Winter)*					−0.854	0.013						
Midnight-6am *(Ref.6 pm-midnight)*							0.918	0.055				
6am-noon												
Noon-6 pm									−0.606	0.005		
Daylight *(Ref. Dusk/dawn)*			−1.181	0.019								
Dark-unlighted					2.477	0.028						
Dark-lighted					2.207	0.042						
Non-intersection *(Ref. Intersection)*			0.377	0.058								
Clear *(Ref. Other)*	−15.692	0.000									−16.152	0.000
Cloudy	−15.632	0.000			−0.926	0.079					−16.172	0.000
Speed < = 30 *(Ref. Speed > = 60)*							−1.482	0.000	−2.133	0.001		
Rural *(Ref. Urban)*					−0.720	0.024						

Only significant variables are shown in the table.

alcohol and drugs presence condition, pedestrian action, age, gender, season, hour, lighting condition, crash location (intersection or non-intersection), weather, posted speed, and setting (rural or urban). For all sets of nominal variables, variables coded as “Other” and “Unknown” were removed from the model, because of the uncertainty issues associated with these crashes. In total, six models were built using the maximization of log-likelihood method, one for the whole data set and one for each cluster except cluster 6, for which too many values were missing. Pedestrian crashes were assigned to a cluster if the posterior probability for that cluster was at least 90%. We selected a probability of 90%, which is high enough to have a correct assignment, but does not result in sample sizes that are too small.

Consequently, the estimated coefficients show the effects of a contributing factor on the probability of a fatal/severe or injury crash, compared to a no injury crash. The estimation results of the six models can be found in Table 4. In correspondence with Kim et al. (2007); Depaire et al. (2008); Mohamed et al. (2013), and Sasidharan et al. (2015), the significance level used in this study is 10%. To conserve space, coefficients statistically insignificant at a 10% confidence level were omitted from the table. All six models were statistically significant at a 10% confidence level.

Compared to no-injury crashes, the predictors with positive coefficients indicate an increase in the probability of occurrence of fatal/severe injury or injury crashes. Focusing on the whole dataset, variables that significantly increase the probability of fatal and severe crashes are as follows: pedestrian alcohol or drugs involvement condition, pedestrian age greater than 65 years old, female, and adverse weather conditions. Additional significant variables that increase the probability of injury are as follows: pedestrian crossing or entering road, winter and nighttime (6 pm to midnight).

Performing traffic accident analysis on a large heterogeneous data set can obscure significant relations. However, according to Table 4, we can see that the significance of the effects of variables is very different in whole data analysis and individual clusters. For example, the odds ratio of pedestrians age between 25–40 years old is 0.305 ($e^{-1.188}$) for whole data analysis, while it is 0.185 ($e^{-1.686}$) for cluster 2. The whole data analysis suggests that the odds of a pedestrian being involved in a crash of fatal/severe injuries when he is 25–40 years old is 69.5% lower than the base line condition (pedestrians older than 65 years old). However, the odds ratio estimated for cluster 2 indicates that pedestrians between 25–40 years old, under dark-lighted conditions, are 81.5% less likely to receive fatal/severe injuries compared to baseline.

Some variables are significant only within specific clusters, which provides additional information. For example, pedestrian crossing and entering roads are not significant in the whole data analysis, while the odds ratio is 2.798 ($e^{1.029}$) for cluster 3, which indicates that compared to the baseline condition (other inappropriate behavior such as standing in the roadway, pushing, working on vehicle in road, etc.), pedestrian crossing and entering roads with alcohol or drugs involvement during nighttime are 180% more likely to receive fatal/severe injuries. This is important information for educational and enforcement purposes, but it was hidden in the whole data analysis. Similarly, speeds less than 30 mph (compared to $\geq$ 60 mph) are not significant in the whole data analysis. This factor is 0.118 ($e^{-2.133}$, $p = 0.001$) for cluster 4, which indicates that compared to high posted speeds, pedestrians in rural, non-intersection areas on low speed roads are 88.2% less likely to receive fatal/severe injuries. In other words, high speeds in rural areas are more likely to cause fatally/seriously injured pedestrian crashes. The indicators for midnight to 6 pm, dark-unlighted, dark-lighted, and non-intersection are also not significant in the whole data analysis, but are highly significant in some clusters, indicating the importance of these variables to certain specific pedestrian crash groups. All of these findings indicate that the full-data model did not only hide these differences, it hid these variables completely. In the above examples, the cluster-based models reveal a more complete interpretation.

Furthermore, the cluster models reveal variations of a variables' effects on the injury outcome probability. For example, the probability of having a fatal/severe crash is the highest when the posted speed is equal to or greater than 60 mph, for crashes occurred in rural, non-intersections (cluster 4). Similarly, for daytime crashes (cluster 1), the probability of having a fatal/severe crash is the highest when the pedestrian is older than 65. In terms of crashes without alcohol or drugs involvement during nighttime (cluster 2), the probability of having a fatal/severe crash is the highest when the lighting condition is dark-lighted or dark-unlighted. For crashes with pedestrians with alcohol or drugs involvement, during nighttime (cluster 3), and urban crashes (cluster 5), the probability of having a fatal/severe crash is the highest when the pedestrian is crossing or entering the roadway.

6. Discussion

This paper presents an analysis of 14,236 pedestrian crashes in Louisiana from 2006 to 2015 with the LCC and MNL models used to investigate the potential relationship between pedestrian injury severity outcomes and a set of the following factors: pedestrian behavior, demographics, crash characteristics, and the built environment.

This study reveals that several variables, as well as their influence, are common in the pedestrian fatal/severe and injury crashes, and this information is further confirmed by the results of the MNL model. Fatal and severe crashes are closely linked to variables, such as pedestrian alcohol or drugs involvement, older than 65 years old, and adverse weather conditions. The probability of having an injury is also high for pedestrians crossing or entering roads between 6 pm and midnight during the winter. The findings are in accordance with previous studies in pedestrian safety analysis using other statistical methods, such as ordered logit or probit models (Ivan et al., 2000; Lee and Abdel-Aty, 2005), and mixed logit models (Islam and Jones, 2014).

Results also show that the combined LCC and MNL models can effectively discover the underlying patterns behind crash data and the MNL analyses are conducted to quantify the impacts of significant contributing factors on pedestrian injury severity outcomes. Certain variables, otherwise insignificant or unidentified in the whole data analysis, are identified, and found to be very significant for a specific cluster. For example, pedestrians crossing and entering roads with alcohol or drugs involvement during nighttime increases the likelihood of fatal crashes, which suggests a critical need of targeted education and ensured enforcement. Pedestrians and motorists need to know about the risk factors associated with sharing the road, and they need to understand, as well as obey, the of right-of-way rules they are legally obligated to follow at crosswalks and other locations. In terms of roadway and environment characteristics, non-intersections and bad visibility increase the likelihood of fatal accidents, which have been also identify in previous studies (Mohamed et al., 2013; Sasidharan et al., 2015). It is important that traffic engineers provide proper lighting conditions for the roadways, and law enforcement ensures traffic regulations are respected. This could be implemented by targeting locations with a high number of infractions.

With the safety strategy plan emphasizing the Destination Zero Deaths objective, Louisiana must pay close attention to the pedestrian safety. To reduce pedestrian crashes, particularly the fatal and injury crashes, it is critical to select countermeasures that target each specific problem at each specific location. Considering the study scope, we make the following recommendations based on the results, focusing on the planning level without involving project level. 1) Enhancing education to reduce inappropriate pedestrian behavior, such as “drunk walking”, may, and most likely will, improve pedestrian safety. 2) Providing pedestrian travel infrastructures, such as cross walks with activated warning lights, wide shoulders on two-lane highways (to serve as sidewalks), sufficient roadway lighting, and pedestrian signs for motorists, improve pedestrian safety in rural areas. It needs to be emphasized that the outcomes should be dealt with as part of new polices on state level to counter the causation of the pedestrian cashes/fatalities. Pedestrian crossing facility design, safety education and enforcement strategies must work together to ensure a safe pedestrian travel environment.

Finally, this work confirms that segmenting the pedestrian crash dataset into homogeneous subsets helps identify important contributing factors that are not evident when using the dataset as a whole. As demonstrated in this study, clustering techniques can be used, not only for descriptive analysis, but also as a preliminary segmentation tool for a more detailed, standard statistical analysis. Segmenting the crash data into homogeneous subgroup would better isolate crash factors under certain conditions. Although much of the past research has suggested that a crash is a confluence of human factors, roadway design, and built environment, this study suggests that more could be done to further decrease the number of crashes. The same crash types can be grouped under unique, homogenous clusters, which will help in the identification of more effective countermeasures, targeting the specific reasons revealed in those identified clusters of crashes. The application of latent class clusters is not simply limited to pedestrian safety, but can be extended to all types of crashes for use in identifying better solutions for improving safety.

7. Conclusion

In this paper, we combined LCC and MNL models to investigate the statistical relationship between pedestrian injury severity outcomes and contributing factors, such as pedestrian behavior, demographics, crash characteristics, and the built environment. The LCC model is utilized to segment the crash data, and then the MNL models are structured to reveal statistical associations between injury severity outcomes and explanatory variables. The results show that the combined models can discover the underlying patterns behind the crash data and quantify the impacts of significant contributing factors on pedestrian injury severity outcomes. Among the major findings are:

- Fatal and severe crashes are closely linked to variables, such as pedestrian alcohol or drugs involvement, older than 65 years old, and adverse weather conditions. The probability of having an injury is also high for pedestrians crossing or entering roads between 6 pm and midnight during the winter.
- Pedestrian behavior plays an important role in pedestrian safety. Crossing, entering road roadway under influence of alcohol or drugs significantly increase the likelihood of fatal crash.
- Contrary to popular belief, there are significant number of pedestrian crashes occur away from urban intersections. As revealed in this study, the probability of pedestrian injury is high when people crossing or entering road away from designated crossing location such as intersection and marked crosswalk, particularly between 6 pm and midnight under dark lighting condition.
- High speeds in rural areas are more likely to cause fatally/seriously injured pedestrian crashes.
- Segmenting the pedestrian crash dataset into homogeneous subsets helps identify important contributing factors that are not evident when using the dataset as a whole.

This study demonstrated that it is critical to develop targeted pedestrian crash countermeasures. Concentrating resources at urban intersection problems is just part of the solution for pedestrian safety. Pedestrian crossing facility design, safety education and enforcement strategies must work together to ensure a safe pedestrian travel environment.

Acknowledgements

The authors would like to thank LADOTD for providing the data. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Akaike, H., 1987. Factor analysis and AIC. Psychome 52, 317–332.
Al-Ghamdi, A., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accid. Anal. Prev. 34 (6), 729–741.
Bedard, M., Guyatt, G., Stones, M., Hirdes, J., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accid. Anal. Prev. 34 (6), 717–727.
Biernacki, C., Govaert, G., 1999. Choosing models in model-based clustering and discriminant analysis. J. Stat. Comput. Simul. 64, 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. Market. 21, 323–340.
Carson, J., Mannering, F., 2001. The effect of ice warning signs on ice-accident

frequencies and severities. Accid. Anal. Prev. 33 (1), 99–109.
Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40 (4), 1257–1266.
Fraley, C., Raftery, A.E., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. 41, 578–588.
Hagenaars, J.A., 1990. Categorical Longitudinal Data - Loglinear Analysis of Panel, Trend and Cohort Data. Sage, Newbury Park.
Haleem, K., Alluri, P., Gan, A., 2015. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accid. Anal. Prev. 81, 14–23.
Hamzehei, A., Chung, E., Miska, M., 2014. Traffic safety risks trends and patterns analysis on motorways. The Transportation Research Board (TRB) 93rd Annual Meeting.
Islam, S., Jones, S.L., 2014. Pedestrian at-fault crashes on rural and urban roadways in Alabama. Accid. Anal. Prev. 72, 267–276.
Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. J. Safety Res. 37 (3), 267–276.
Ivan, J.N., Gårder, P.E., Zajac, S.S., 2000. Finding Strategies to Improve Pedestrian Safety in Rural Areas. Connecticut Transportation Institute, University of Connecticut.
Kashani, A., Mohaymany, A., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf. Sci. 49, 1314–1320.
Kim, K., Yamashita, E., 2007. Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. J. Adv. Transp. 41 (1), 69–89 Winter 2007.
Kim, J.-K., Kim, S., Ulfarsson, G., Porrello, L., 2007. Bicyclist injury severities in bicycle-motor vehicle accidents. Accid. Anal. Prev. 39, 238–251.
Kweon, Y., Kockelman, K., 2003. Driver attitudes and choic- es: seatbelt use, speed limits, alcohol consumption, and crash histories. 82nd Annual Meeting of Transportation Research Board.
Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accid. Anal. Prev. 37 (4), 775–786.
Louisiana Department of Transportation and Development. Crash Data. wwwsp.dotd.la.gov/Inside_LaDOTD/Divisions/Multimodal/Highway_Safety/Pages/Crash_Data.aspx.
McFadden, D., 1981. Econometric models of probabilistic choice. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data With Econometric Applications. MIT Press, Cambridge, MA.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.
Mohamed, G.M., Saunier, N., Miranda-Moreno, L.F., Ukkusuri, S.V., 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.
Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41, 683–691.
Park, B.-J., Lord, D., Hart, J., 2010. Bias properties of Bayesian statistics in finiture mixture of negative regression models for crash data analysis. Accid. Anal. Prev. 42, 741–749.
Raftery, A.E., 1986. A note on Bayes factors for log-linear contingency table models with vague prior information. J. R. Stat. Soc. Ser. B 48, 249–250.
Retting, R., 2017. Pedestrian Traffic Fatalities by State 2016: Preliminary Data. Governors Highway Safety Association, Washington, DC.
Rothman, L., William Howard, A., Camden, A., Macarthur, C., 2012. Pedestrian crossing location influences injury severity in urban areas. Inj. Prev.
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.
Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. J. Safety Res. 27 (3), 183–194.
Smart Growth America, 2014. Dangerous by Design. Washington, DC.
Sze, N.N., Wong, S.C., 2007. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accid. Anal. Prev. 39, 1267–1278.
Traffic Safety Facts, 2016. Data, Pedestrians, 2018. (Report No. DOT HS 812 493). National Highway Traffic Safety Administration, Washington, DC.
Ulfarsson, G., Mannering, F., 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., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors for fatal road traffic accidents in Udine, Italy. Accid. Anal. Prevent. 34 (1), 71–84.
Vermunt, J.K., 1997. Log-linear Models for Event Histories. Thousand Oakes. Sage Publications.
Vermunt, J.K., Magidson, J., 2002. Latent class cluster analysis. In: Hagenaars, J.A., McCutcheon, A.L. (Eds.), Applied Latent Class Analysis. Cambridge University Press, Cambridge, UK, pp. 89–106.
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, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y., 2000. Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. Accid. Anal. Prev. 32 (1), 117–125.