Accident Analysis and Prevention 131 (2019) 284–296

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Modelling severity of pedestrian-injury in pedestrian-vehicle crashes with latent class clustering and partial proportional odds model: A case study of North Carolina

Yang Li[a], Wei (David) Fan[b,*]

[a] USDOT Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE), Department of Civil and Environmental Engineering, University of North Carolina at Charlotte, EPIC Building, Room 3366, 9201 University City Boulevard, Charlotte, NC, 28223-0001, United States
[b] USDOT Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE), Department of Civil and Environmental Engineering, University of North Carolina at Charlotte, EPIC Building, Room 3261, 9201 University City Boulevard, Charlotte, NC, 28223-0001, United States

ARTICLE INFO

Keywords:
Pedestrian
Crash
North Carolina
Latent class clustering
Partial proportional odds

ABSTRACT

There are more than 2000 pedestrians reported to be involved in traffic crashes with vehicles in North Carolina every year. 10%–20% of them are killed or severely injured. Research studies need to be conducted in order to identify the contributing factors and develop countermeasures to improve safety for pedestrians. However, due to the heterogeneity inherent in crash data, which arises from unobservable factors that are not reported by law enforcement agencies and/or cannot be collected from state crash records, it is not easy to identify and evaluate factors that affect the injury severity of pedestrians in such crashes. By taking advantage of the latent class clustering (LCC), this research firstly applies the LCC approach to identify the latent classes and classify the crashes with different distribution characteristics of contributing factors to the pedestrian-vehicle crashes. By considering the inherent ordered nature of the traffic crash severity data, a partial proportional odds (PPO) model is then developed and utilized to explore the major factors that significantly affect the pedestrian injury severities resulting from pedestrian-vehicle crashes for each latent class previously obtained in the LCC. This study uses police reported pedestrian crash data collected from 2007 to 2014 in North Carolina, containing a variety of features of motorist, pedestrian, environmental, roadway characteristics. Parameter estimates and associated marginal effects are mainly used to interpret the models and evaluate the significance of each independent variable. Lastly, policy recommendations are made and future research directions are also given.

1. Introduction

Compared to other entities in transportation, pedestrians are the most vulnerable ones. The number of injuries and deaths in pedestrian-vehicle crashes causes huge impacts both socially and economically. Such issue is particularly critical within the context of continuous traffic safety improvements in US. According to the data released from North Carolina Department of Transportation (NCDOT), more than 2000 pedestrians are involved in crashes with vehicles each year during the past decades in North Carolina. On average, a total of 150–200 pedestrians are annually killed on NC roads, and additional 200–300 pedestrians are severely injured.

Under such situation, a large number of research studies have been conducted to examine the major contributors to pedestrian safety related issues, such as the influence of alcohol, head injuries, demographic and economic characteristics, roadway characteristics, environmental factors, collision types, party at fault, and locations (Kim et al., 2008a,b, 2010; Dai, 2012). It is important to identify the key contributing factors in pedestrian-vehicle crashes by applying and developing proper modelling approaches.

Moreover, due to the heterogeneity inherent in traffic crash data, which arises from unobservable factors that are not reported by law enforcement agencies and cannot be collected from state crash records, it is not easy to identify and evaluate factors that significantly affect injury severity of pedestrians in such crashes. Such heterogeneity might result in biased estimation of parameters and thus drawing potentially incorrect conclusions (Mannering and Bhat, 2014; Shaheed and Gkritza, 2014).

The purpose of this research is to investigate factors that significantly contribute to the severity of pedestrian injuries resulting from pedestrian-vehicle crashes by taking advantage of LCC and PPO logit model. This study uses police reported pedestrian crash data collected

* Corresponding author.
E-mail addresses: yli107@uncc.edu (Y. Li), wfan7@uncc.edu (W.D. Fan).

https://doi.org/10.1016/j.aap.2019.07.008
Received 22 April 2019; Received in revised form 12 July 2019; Accepted 15 July 2019
Available online 24 July 2019
0001-4575/

from 2007 to 2014 in North Carolina. A variety of motorist, pedestrian, environmental, and roadway characteristics are examined. Parameter estimates and associated marginal effects are also calculated and used to interpret the model and evaluate the significance of each independent variable.

2. Literature review

Many research efforts have been made to analyze the injury severity of pedestrians in pedestrian-vehicle crashes. A wide variety of contributing factors has been identified by existing research, such as alcohol, head injuries, demographic and economic characteristics, roadway characteristics, environmental factors, collision types, party at fault, and locations (Lee and Abdel-Aty, 2005; Kim et al., 2008a,b, 2010; Tay et al., 2011; Dai, 2012). Such studies clearly showed that it is not a coincidence that certain pedestrian injury occurs in a traffic crash, and developing effective modeling approach to investigating contributing factors to pedestrian injury severity continues to be a critical need in the field of transportation safety.

In general, traffic crash data are inherently heterogeneous, and such data heterogeneity can cause one to draw incorrect conclusions in many ways (Depaire et al., 2008). Therefore, researchers always seek to reduce such heterogeneity by focusing on building separate models for each traffic accident type (Valent et al., 2002; Islam and Mannering, 2006; Savolainen and Mannering, 2007; Haleem et al., 2015) or a very specific traffic type (Bedard et al., 2002; Ulfarsson and Mannering, 2004; Haleem and Gan, 2015). By applying such methods, results from aforementioned studies show that heterogeneity does seem to be reduced to some extent.

Depaire et al. (2008) pointed out that most beforementioned segmentations of the traffic accident data are built upon expertise, methodological decisions or the will of focusing on particular issues, which sometimes could result in a viable outcome. However, it does not always guarantee the homogenous segmentations of the data. Thus, it is crucial to apply non-trivial and valid process of identifying rational segmentations in traffic crash data. From the standpoint of data analysis, data mining technique could be viewed as and used to conduct a computer-aided automatic exploratory data analysis of large complex data sets (Friedman, 1998). Recently several data mining techniques have found their benefits in the conduct of traffic safety research, such as classification and regression tree (CART) (Chang and Wang, 2006; Kashani and Mohaymany, 2011), k-means clustering (Kim and Yamashita, 2007; Mohamed et al., 2013), and LCA (Depaire et al., 2008; Shaheed and Gkritza, 2014; Yasmin et al., 2014a,b; Sun et al., 2019). Most results indicate that instead of using the full dataset, clustered data could provide more insightful and meaningful information.

In analyzing the crash severity, several statistical methods has been widely deployed, such as binary logit models (Sze and Wong, 2007; Moudon et al., 2011), multinomial logit (MNL) models (Quddus et al., 2009; Aziz et al., 2013; Zhou et al., 2013; Chen and Fan, 2019a), mixed logit models (ML) (Milton et al., 2008; Malyshkina and Mannering, 2010; Chen and Chen, 2011; Haleem et al., 2015; Chen and Fan, 2019b), ordered logit or probit model (Lee and Abdel-Aty, 2005; Yasmin et al., 2014a,b), partial proportional odds (PPO) model (Wang and Abdel-Aty, 2008; Rifaat et al., 2012; Li and Fan, 2018, 2019). Despite the popularity of the MNL model, it would not be valid if unobserved data heterogeneities actually exist, since MNL models assume the same effects of independent variables across crash observations. Thus, by having parameters that are randomly distributed, the ML model is applied to deal with such independence of irrelevant alternatives (IIA). However, both MNL and ML models treat the injury severity levels as non-ordered, which neglects the inherent hierarchical nature of crash injury severities. In the meantime, ordered logit/probit models assume that the parameter estimates are same and constant across severity level, which could be unrealistic and therefore should be relaxed in crash injury severity modelling. Beforementioned discussions could be supported by multiple safety papers with comparison among various models (Mooradian et al., 2013; Yasmin et al., 2014a,b; Eluru and Yasmin, 2015; Gong et al., 2016). To overcome the previously mentioned limitations and follow the LCC process, as well as to consider the inherent ordered nature of the traffic crash severity data, this study applies the partial proportional odds (PPO) model to explore the contributing factors that significantly affect the pedestrian injury severities resulting from pedestrian-vehicle crashes for each latent class that are obtained in the LCC.

3. Data description

Data used in this study are the police reported pedestrian crash data of 2007–2014, which are retrieved from the Division of Bicycle and Pedestrian Transportation in North Carolina Department of Transportation (NCDOT). Categories of explanatory variables include the crash group, injury severity (of both the driver and pedestrian), roadway characteristics, vehicle type, driver and pedestrian demographics, and environmental characteristics, etc. Only cases with pedestrian injury severity in pedestrian-vehicle crashes are examined and analyzed. A total of 10,875 observations are identified, by removing incomplete and obviously incorrect records at the very beginning and during the data examination and cleaning process. The distributions of the 10,875 records at each injury level are shown as follows: Fatal Injury (K): 921(8.47%); Incapacitating Injury (A): 861(7.92%); Non-incapacitating (evident) Injury (B): 4097(37.67%); Possible Injury (C): 4431(40.74%); and No Injury (O): 565(5.2%). The crash injury severity is reorganized into four categories: K, A, B, and C/O. Kim et al. (2008a,b, 2010) and Eluru et al. (2008) applied the similar classification method in their studies in which the injury levels of C and O were combined into one group perhaps because they share most of the crash characteristics. In addition, it is regarded to be more realistic to separate the injury levels of K and A, due to the complete difference in terms of their inherent injury degrees.

Table 1 displays a summary of the data, including detailed information about variables of their definitions and descriptions, as well as the percentage of observed crash frequency at each severity level. Explanatory variables are classified into 10 categories describing the crash time, personal characteristics (driver/vehicle and pedestrian), traffic and road variables (traffic controls, roadway characteristics), environment and land use specifications, etc. for each recorded crash. For each classification variable, dummy variables (0–1) are also created, which are shown in Table 1.

4. Methodologies

4.1. Latent class clustering (LCC)

Latent class clustering is a statistical technique that can be used for finding groups or subtypes of cases in multivariate categorical data with each class being characterized by a pattern of common traits that maximize both the homogeneity of elements within classes and the heterogeneity between classes (Hair et al., 1998). This unsupervised learning algorithm does not require the number of clusters to be predetermined (Vermunt and Magidson, 2002; Depaire et al., 2008; Sun et al., 2019).

Given a sample of crash data, it is assumed that there are C latent

Table 1
Descriptive Statistics of Explanatory Variables.

Variable			Total	C/O[a]	B[b]	A[c]	K[d]
Pedestrian–vehicle crashes	**Label**		10875(100%)	4996(45.94%)	4097(37.67%)	861(7.92%)	921(8.47%)
Motorist characteristics							
Age (in years)	da_24	≤ 24	2173(19.98%)	923(42.48%)	881(40.54%)	172(7.92%)	197(9.07%)
	da25_54	25 - 54	5844(53.74%)	2652(45.38%)	2200(37.65%)	479(8.2%)	513(8.78%)
	da_55	≥ 55*	2858(26.28%)	1421(99.65%)	1016(71.78%)	210(14.62%)	211(13.93%)
Gender	PedSex	Male	6086(55.96%)	2637(43.33%)	2326(38.22%)	522(8.58%)	601(9.88%)
		Female*	4789(44.04%)	2359(49.26%)	1771(36.98%)	339(7.08%)	320(6.68%)
Hit and run?	Hitrun	Yes	290(2.67%)	128(44.14%)	91(31.38%)	35(12.07%	36(12.41)
		No*	10,585(97.33%)	4868(45.99%)	4006(37.85%)	826(7.8%)	885(8.36%)
Has been drinking?	CrashAlcoh	Yes	1806(16.61%)	486(26.91%)	720(39.87%)	248(13.73%)	352(19.49%)
		No*	9069(83.39%)	4510(49.73%)	3377(37.24%)	613(6.76%)	569(6.27%)
Pedestrian characteristics							
Age (in years)	ped_24	< 24	3739(34.38%)	1635(86.21%)	1616(88.17%)	295(15.57%)	630(15.51%)
	ped25_54	25 - 54	5069(46.61%)	2410(47.54%)	1758(34.68%)	400(7.89%)	501(9.88%)
	ped_55	> 54*	2067(19.01%)	951(88.14%)	723(73.39%)	166(15.31%)	227(23.15%)
Gender	DrvrSex	Male	6474(59.53%)	2715(41.94%)	2500(38.62%)	589(9.1%)	670(10.35%)
		Female*	4401(40.47%)	2281(51.83%)	1597(36.29%)	272(6.18%)	251(5.7%)
Time							
Time	c0am_4am	12:01 AM - 4:00 AM	641(5.89%)	165(25.74%)	262(40.87%)	79(12.32%)	135(21.06%)
	c4am_8am	4:01 AM - 8:00 AM	982(9.03%)	387(39.41%)	364(37.07%)	98(9.98%)	133(13.54%)
	c8am_12am	8:01 AM - 12:00 PM	1428(13.13%)	815(57.07%)	495(34.66%)	68(4.76%)	50(3.5%)
	c12pm_4pm	12:01 PM - 4:00 PM	2264(20.82%)	1252(55.3%)	822(36.31%)	127(5.61%)	63(2.78%)
	c4pm_8pm	4:01 PM - 8:00 PM	3228(29.68%)	1521(47.12%)	1266(39.22%)	232(7.19%)	209(6.47%)
	c8pm_12am	8:01 PM - 12:00 AM*	2332(21.44%)	856(36.71%)	888(38.08%)	257(11.02%)	331(14.19%)
Traffic control							
Control type	NConP	No Control Present	6947(63.88%)	3183(45.82%)	2615(37.64%)	549(7.9%)	600(8.64%)
	Hcon	Human Control	127(1.17%)	79(62.2%)	39(30.71%)	5(3.94%)	4(3.15%)
	TraS	Traffic Sign	2060(18.94%)	857(41.6%)	777(37.72%)	185(8.98%)	241(11.7%)
	TraSl	Traffic Signal*	1741(16.01%)	877(50.37%)	666(38.25%)	122(7.01%)	76(4.37%)
Land development							
Land use	Commercial	Commercial	5485(50.44%)	2769(50.48%)	1925(35.1%)	416(7.58%)	375(6.84%)
	FWP	Farms, Woods, Pastures	1171(10.77%)	334(28.52%)	423(36.12%)	143(12.21%)	271(23.14%)
	Industrial	Industrial	65(0.6%)	28(43.08%)	25(38.46%)	7(10.77%)	5(7.69%)
	Institutional	Institutional	377(3.47%)	217(57.56%)	140(37.14%)	12(3.18%)	8(2.12%)
	Residential	Residential*	3777(34.73%)	1648(43.63%)	1584(41.94%)	283(7.49%)	262(6.94%)
Environmental factors							
Weather	FSS	Fog, Smog, Smoke	53(0.49%)	19(35.85%)	21(39.62%)	1(1.89%)	12(22.64%)
	Clear	Clear	8367(76.94%)	3812(45.56%)	3191(38.14%)	652(7.79%)	712(8.51%)
	Cloudy	Cloudy	1530(14.07%)	711(46.47%)	548(35.82%)	131(8.56%)	140(9.15%)
	Rain	Rain	871(8.01%)	425(48.79%)	322(36.97%)	72(8.27%)	52(5.97%)
	SSHFRD	Snow, Sleet, Hail, Freezing Rain/Drizzle	42(0.39%)	24(57.14%)	13(30.95%)	2(4.76%)	3(7.14%)
	OtherWeather	Other*	12(0.11%)	5(41.67%)	2(16.67%)	3(25%)	2(16.67%)
Light Condition	DarkLRd	Dark - Lighted Roadway	2301(21.16%)	927(40.29%)	918(39.9%)	224(9.73%)	232(10.08%)
	DarkNOTLRd	Dark - Roadway Not Lighted	2169(19.94%)	635(29.28%)	769(35.45%)	277(12.77%)	488(22.5%)
	Dawm	Dawn	180(1.66%)	88(48.89%)	69(38.33%)	14(7.78%)	9(5%)
	Daylight	Daylight	5915(54.39%)	3189(53.91%)	2231(37.72%)	322(5.44%)	173(2.92%)
	Dusk	Dusk*	310(2.85%)	157(50.65%)	110(35.48%)	24(7.74%)	19(6.13%)
Vehicle characteristics							
Vehicle type	Car	Car	6088(55.98%)	2865(47.06%)	2362(38.8%)	462(7.59%)	399(6.55%)
	Bus	Bus	86(0.79%)	50(58.14%)	24(27.91%)	6(6.98%)	6(6.98%)
	Motorcycle	Motorcycle	47(0.43%)	15(31.91%)	22(46.81%)	4(8.51%)	6(12.77%)
	Heavy_Vehicle	Heavy Vehicle	227(2.09%)	63(27.75%)	73(32.16%)	24(10.57%)	67(29.52%)
	Van	Van	727(6.69%)	358(49.24%)	253(34.8%)	51(7.02%)	65(8.94%)
	Pickup_Truck	Pickup Truck	1494(13.74%)	648(43.37%)	554(37.08%)	129(8.63%)	163(10.91%)
	Sport_Utility	Sport Utility	1969(18.11%)	872(44.29%)	730(37.07%)	172(8.74%)	195(9.9%)
	Public	Public (Police, etc.)	45(0.41%)	24(53.33%)	14(31.11%)	0(0%)	7(15.56%)
	LiTr	Light Truck (Mini-Van, Panel) *	192(1.77%)	101(52.6%)	65(33.85%)	13(6.77%)	13(6.77%)
Roadway characteristics							
Road class	Freeway	Freeway	152(1.4%)	24(15.79%)	38(25%)	18(11.84%)	72(47.37%)
	Local_Street	Local Street	5810(53.43%)	2670(45.96%)	2398(41.27%)	437(7.52%)	305(5.25%)
	NCSR	North Carolina state route	658(6.05%)	203(30.85%)	231(35.11%)	96(14.59%)	128(19.45%)
	PRdDr	Private Road, Driveway	172(1.58%)	71(41.28%)	76(44.19%)	20(11.63%)	5(2.91%)
	PVehA	Public Vehicular Area	2036(18.72%)	1404(68.96%)	557(27.36%)	61(3%)	14(0.69%)
	NCSSR	North Carolina secondary state route	1272(11.7%)	410(32.23%)	525(41.27%)	135(10.61%)	202(15.88%)
	USRoute	US Route*	775(7.13%)	214(27.61%)	272(35.1%)	94(12.13%)	195(25.16%)

(*continued on next page*)

Table 1 (*continued*)

Variable			Total	C/O[a]	B[b]	A[c]	K[d]
Road geometry	RdCharacte	Straight	10317(94.87%)	4809(46.61%)	3883(37.64%)	790(7.66%)	835(8.09%)
		Curve*	558(5.13%)	187(33.51%)	214(38.35%)	71(12.72%)	86(15.41%)
Road type	one_way	One-Way	780(7.17%)	467(59.87%)	262(33.59%)	29(3.72%)	22(2.82%)
	two_way_D	Two-Way, Divided	2122(19.51%)	746(35.16%)	849(40.01%)	218(10.27%)	309(14.56%)
	two_way_NOTD	Two-Way, Not Divided*	7973(73.31%)	3783(47.45%)	2986(37.45%)	614(7.7%)	590(7.4%)
Crash characteristics							
Crash group (Pedestrian behavior associated with Motorist maneuver)	Midblock	Midblock	250(2.3%)	97(38.8%)	111(44.4%)	22(8.8%)	20(8%)
	PedCro	Pedestrian Crossing	4447(40.89%)	1941(43.65%)	1683(37.85%)	385(8.66%)	438(9.85%)
	PedinRd	Pedestrian in Roadway	1009(9.28%)	310(30.72%)	351(34.79%)	117(11.6%)	231(22.89%)
	Back	Backing Vehicle	1121(10.31%)	750(66.9%)	326(29.08%)	31(2.77%)	14(1.25%)
	WorkinRd	Working in Roadway	197(1.81%)	116(58.88%)	63(31.98%)	11(5.58%)	7(3.55%)
	OffRd	Off Roadway	1296(11.92%)	829(63.97%)	399(30.79%)	53(4.09%)	15(1.16%)
	DD	Dash / Dart-Out	1401(12.88%)	435(31.05%)	723(51.61%)	156(11.13%)	87(6.21%)
	PedWCro	Pedestrian Waiting to Cross	8(0.07%)	5(62.5%)	2(25%)	1(12.5%)	0(0%)
	PedWARd	Pedestrian Walking Along Roadway*	1146(10.54%)	513(44.76%)	439(38.31%)	85(7.42%)	109(9.51%)
Rural or urban	RuralUrban	Urban	7879(72.45%)	3954(50.18%)	2969(37.68%)	532(6.75%)	424(5.38%)
		Rural*	2996(27.55%)	1042(34.78%)	1128(37.65%)	329(10.98%)	497(16.59%)

* Selected as base among each variable category.
[a] C/O - No/Possible injury (base).
[b] B - non-incapacitating injury.
[c] A - incapacitating injury.
[d] K - killed.

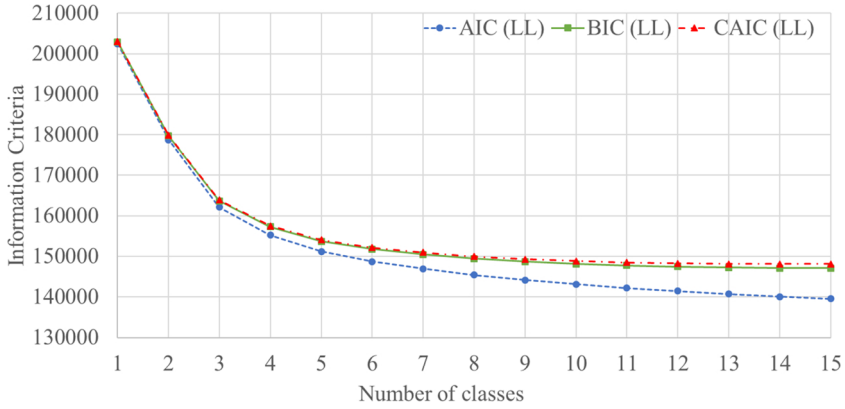

Fig. 1. Evolution of AIC (LL), BIC (LL), CAIC (LL) when adding classes to the model.

classes to be estimated based on J categorical items. Let $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{iJ})$ denote observation i's responses to J categorical items where possible values of Y_{ij} are 1, …, r_j. Let $C_i = 1, 2, \ldots, C$ represent the latent class membership of individual i, and let $I(y_j=k)$ be the indicator function, which equals to 1 if y_j equals to r_j and 0 otherwise. The contribution made by observation i to the likelihood can be shown as:

$$P(\mathbf{Y_i} = \mathbf{y}) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)}$$

where γ_c is the probability of membership in latent class c, and ρ is the item-response probability conditional on latent class membership. The LCC models for this study are built using SAS 9.4, by calling the LCA procedure, which is developed by The Methodology Center, and more details can be referred to (Lanza et al., 2007).

For the goodness-of-fit and the purpose of deciding the appropriate number of classes, the estimation of the p-values by time-consuming parametric bootstrapping and information criteria (i.e., Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and Consistent Akaike Information Criterion (CAIC)) are commonly used. In addition, it is well note that due to the sparse frequency table in such data analysis, asymptotic p-values from the Pearson or the likelihood-ratio chi-squared tests can no longer be trusted (Sun et al., 2019). As such, this study applies the BIC, AIC and CAIC to identify the final LCC models.

4.2. Partial proportional odds model (PPO)

The parameter estimates are assumed to be same and constant across severity levels in ordered logit/probit models. However, such parallel-lines/proportional odds (PO) assumption could be violated in many circumstances. The generalized ordered logit model with a full relaxation of the PO assumption to all variables could be utilized to overcome this problem, as shown below:

$$P(Y_i \geq j) = \frac{\exp(\alpha_j - \mathbf{X'_i}\boldsymbol{\beta_j})}{1 + \exp(\alpha_j - \mathbf{X'_i}\boldsymbol{\beta_j})}$$

Table 2
Results of LCC and the distributions of featured variables (bold) in each class.

Variables		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Total cases		860 (7.91%)	1380 (0.1269%)	2261 (20.79%)	2040 (18.76%)	1918 (17.64%)	2416 (22.22%)
Traffic control							
Traffic control	Traffic Signal	24 (2.79%)	9 (0.65%)	8 (0.35%)	44 (2.16%)	**1193 (62.2%)**	463 (19.16%)
Land development							
Land development	Residential	389 (45.23%)	480 (34.78%)	427 (18.89%)	**1298 (63.63%)**	311 (16.21%)	872 (36.09%)
Environmental factors							
Light condition	Dark - Lighted Roadway	5 (0.58%)	41 (2.97%)	379 (16.76%)	96 (4.71%)	153 (7.98%)	**1627 (67.34%)**
	Dark - Roadway Not Lighted	67 (7.79%)	**1305 (94.57%)**	71 (3.14%)	35 (1.72%)	10 (0.52%)	681 (28.19%)
	Daylight	**724 (84.19%)**	8 (0.58%)	1717 (75.94%)	1762 (86.37%)	1659 (86.5%)	45 (1.86%)
Roadway characteristics							
Road class	Local Street	13 (1.51%)	23 (1.67%)	157 (6.94%)	**1916 (93.92%)**	1702 (88.74%)	1999 (82.74%)
	Public Vehicular Area	4 (0.47%)	3 (0.22%)	**1929 (85.32%)**	24 (1.18%)	49 (2.55%)	27 (1.12%)
	North Carolina secondary state route	**572 (66.51%)**	**654 (47.39%)**	0 (0%)	6 (0.29%)	22 (1.15%)	18 (0.75%)
Crash characteristics							
Crash group (Pedestrian behavior associated with Motorist maneuver)	Pedestrian Crossing	212 (24.65%)	310 (22.46%)	76 (3.36%)	606 (29.71%)	**1779 (92.75%)**	1464 (60.6%)
	Off Roadway	2 (0.23%)	0 (0%)	**1294 (57.23%)**	0 (0%)	0 (0%)	0 (0%)
Rural or Urban	Rural	**840 (97.67%)**	**1329 (96.3%)**	443 (19.59%)	154 (7.55%)	85 (4.43%)	145 (6%)

where j denotes the predefined severity level and Y_i represents the observed crash injury for crash i and j is the severity level. $\mathbf{X_i}$ is a $p \times 1$ vector containing the crash value i on the full set of p explanatory variables, $\boldsymbol{\beta}j$ denotes a $p \times 1$ vector of estimated coefficients, and α_j is the cut-point for the jth cumulative logit. In the generalized ordered logit model, $\boldsymbol{\beta}$ is not fixed across levels while compared to the ordered logit model. However, in most cases, not all variables violate the PO assumption. To resolve the issue, the PPO model relaxes the PO assumption by allowing particular explanatory variables to violate the PO assumption, while other remaining independent variables could stick to the PO assumption. Hence, the PPO model divides the explanatory variables into two sets corresponding to those with/without violating the PO assumption: $\mathbf{X_i}$ and $\mathbf{T_i}$, and then the PPO model with logit function can be expressed as:

$$P(Y_i \geq j) = \frac{\exp[\alpha_j - (\mathbf{X'_i\beta_j} + \mathbf{T'_i\gamma_j})]}{1 + \exp[\alpha_j - (\mathbf{X'_i\beta_j} + \mathbf{T'_i\gamma_j})]}$$

A series of Wald Chi-square tests can be conducted and used to examine whether or not the PO assumption is violated for each individual explanatory variable in the generalized ordered logit model (Wang and Abdel-Aty, 2008; Sasidharan and Menéndez, 2014).

Prior to the development of PPO models, an ordered logit model is firstly built to identify the significance of explanatory variables and then a generalized ordered logit model is used as the base to establish the PPO model for each class. The PPO models for this study are developed using SAS 9.4, by calling the LOGISTIC procedure, which can be referred to (Derr, 2013).

It is noticeable that the sign of $\boldsymbol{\beta}$ does not always denote the direction of the effect of the intermediate outcomes (Wooldridge, 2010; Washington et al., 2010) while interpreting the results of the PPO model. Therefore, the marginal effects are utilized to further explore the impacts of variables. All variables are dummy-coded in this study so that the marginal effects (i.e., a difference of probability change rather than the derivative) are calculated after the development of PPO models for each variable.

5. Modeling results

5.1. Latent class clustering results

All variables mentioned in the Table 1 are used in the LCC. The evolution of AIC, BIC and CAIC information criteria are shown in the Fig. 1. According to Fig. 1, the six-class model is chosen as the final LCC model, since from six classes onwards, the BIC and CAIC barely show further improvements (< 1%). Meanwhile, the quality of the clustering solution is evaluated by using the entropy criterion (Peel and McLachlan, 2000), where the closer this criterion is to 1, the better the clustering. The value is 0.91 in this case, which indicates a good separation between the classes.

Similar to the research studies of (Depaire et al., 2008; Sun et al., 2019), the final six-class models could be described based on the skewed feature distributions of particular variables, which differ between the classes. For example, if one class has 95% of its crashes occurring between midnight to 4am, while other classes have more balanced distributions over the feature of “time”, one can identify this class as “12am to 4am traffic crash” class. It is noticeable that this analysis just tries to identify various traffic crash types within the data, which can be overlapping on some variables.

Table 2 shows the results of the final set of featured variables that are used for profiling the six classes. In class 1, 97.67% crashes happen on rural and 66.51% on North Carolina secondary state routes, particularly 84.19% in daylight condition. Thus, class 1 can be referred to as “Crashes on rural North Carolina secondary route in daylight”. Class 2 overlaps class 1 on “rural” and “road class (North Carolina secondary state route)” with 96.3% and 47.39% respectively, but 94.57% crashes

Table 3a
PPO Model for Class 1 in Pedestrian-Vehicle Crashes.

Class 1		Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Variable		All Level		K[d]		A[c]		B[b]	
Intercept				−2.4075^{++}	0.2429	−1.3645^{++}	0.2271	0.7726^{++}	0.2224
Motorist characteristics									
Hit and run (vs. no)	Yes	1.1415^{+}	0.5785						
Pedestrian characteristics									
Age (in years, base: ≥ 55)	25 - 54	−0.3709^{+}	0.1526						
Time									
Time (base: 8:01 PM - 12:00)	4:01 AM - 8:00 AM	0.5176^{++}	0.1866						
Land development									
Land use (base: Residential)	Farms, Woods, Pastures	0.3194^{+}	0.1351						
	Institutional	−0.8146^{+}	0.3855						
Vehicle characteristics									
Vehicle type (base: Light Truck (Mini-Van, Panel)	Heavy Vehicle			2.1778^{++}	0.3956	1.6291^{++}	0.3862	0.5582	0.4495
Roadway characteristics									
Road class type (base: US Route)	Freeway	1.6656^{++}	0.5915						
Road geometry	Straight	−0.7540^{++}	0.1884						
Crash characteristics									
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)	Midblock	0.7719^{++}	0.2505						
	Pedestrian Crossing	0.7878^{++}	0.1715						
	Dash / Dart-Out	1.0100^{++}	0.1851						

No. of observations, 860.
-2 × Log-likelihood at convergence, 1965.521.
2 × Log-likelihood (constant only), 2103.394.
$^{+}$Level of significance > 95%. $^{++}$Level of significance > 99%.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

occur on dark roadway without lighting. For class 3, 85.32% crashes are on the public vehicular area with pedestrians off roadways. 93.92% of the crashes occur on local streets for class 4 with 63.63% of the crashes in residential area. Class 5 has 92.75% of the crashes happening while pedestrians are crossing the roads, and 62.2% of the crashes occur with traffic signal being present. For the last class, 67.34% of the crashes happen on the lighted roadways under dark light condition. The total sample sizes of each class are also presented in Table 2.

5.2. Partial proportional odds model results

A restricted PPO model including all observations is built initially and its associated log likelihood value at convergence is -10975. Then PPO models for each latent class are also developed, except Class 3, which collapses into ordered logit model with all variables passing the Wald Chi-square tests for the PO assumption. It is found that various explanatory variables have a statistically significant standard deviation in all six datasets and sets of the identified explanatory variables are different between classes. Therefore, five PPO models and one ordered logit model have been obtained. The sum of the log likelihood values of the all sub-models is -10759. Following the procedure of conducting the likelihood ratio test, the value of test statistics equals to 432 with 61 degrees of freedom. The result rejects the hypothesis that all six latent classes are the same at the confidence level of 99.9%. Therefore, it indicates that significant differences of pedestrian injury severity in pedestrian-vehicle crashes exist between different latent classes. Results of all models for each latent class are displayed in Tables 3a–3f, respectively. For variables that violate the PO assumption, coefficients are set to be varied across injury severity levels while other variables that meet the PO assumption have fixed effects.

5.3. Marginal effects results

Obvious differences can be observed among the contributing factors under each crash severity level across all latent classes when comparing sub-models. Cautions are needed while interpreting the results, since the sign(s) of the coefficient(s) does(do) not always denote the direction of the effect of the intermediate outcomes. Hence, marginal effects of each variable are computed to show how contributing factors affect the pedestrian injury severity levels. In this study, marginal effects measure discrete changes on account of the categorical variables, which represent how the predicted probabilities change as the binary independent variables change from 0 to 1. This section will focus on the interpretation of the contributing factors affecting the crash severity of pedestrians and discuss the differences between different latent classes, mainly based on the marginal effects from Tables 4a and 4b.

5.3.1. Human factors

Factors of this category come from both sides of pedestrians and drivers. The influence of alcohol on drivers is significant in class 2, 3, and 6. Differences of such impacts can also be seen between class 2, 3 and 6 across severity levels, where the drunk drivers raise the risks of non-capacity injury (marginal effects +0.1077) for pedestrians in class 3 (i.e., crashes with pedestrians off roadway on the public vehicular area), and raise the risks of fatal injury for class 2 (i.e., crashes on rural North Carolina secondary route in daylight) and class 6 (i.e., crashes on

Table 3b
PPO Model for Class 2 in Pedestrian-Vehicle Crashes.

Class 2		Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Variable		All Level		K[d]		A[c]		B[b]	
Intercept				−1.5981++	0.2004	−1.1424++	0.1908	0.6111++	0.1837
Motorist characteristics									
Has been drinking (vs. no)	Yes	0.3460++	0.1065						
Hit and run (vs. no)	Yes			0.7724+	0.3548	1.2354++	0.3581	0.2814	0.4165
Pedestrian characteristics									
Age (in years, base: ≥ 55)	< 25			−0.9596++	0.2108	−0.7406++	0.1985	−0.3808	0.2068
Gender (base: female)	Male								
Time									
Time (base: 8:01 PM - 12:00)	12:01 AM - 4:00 AM	0.4964++	0.1500						
	4:01 AM - 8:00 AM	0.9489++	0.1848						
Vehicle characteristics									
Vehicle type (base: Light Truck (Mini-Van, Panel)	Heavy Vehicle	0.9855++	0.3297						
Roadway characteristics									
Road class type (base: US Route)	Freeway	1.0784++	0.3086						
Road type (base: Two-Way, Not Divided)	Two-Way, Divided	0.5268++	0.1515						
Crash characteristics									
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)	Pedestrian Crossing			1.0201++	0.1761	1.3675++	0.1623	1.0384++	0.1857
	Pedestrian in Roadway			1.0181++	0.1575	1.072++	0.1433	0.6207++	0.1455
	Dash / Dart-Out	1.0960++	0.2321						

No. of observations, 1380.
-2×Log-likelihood at convergence, 3387.673.
-2×Log-likelihood (constant only), 3674.085.
+Level of significance > 95%. ++Level of significance > 99%.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

Table 3c
PPO Model for Class 3 in Pedestrian-Vehicle Crashes.

Class 3		Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Variable		All Level		K[d]		A[c]		B[b]	
Intercept				−3.8250++	0.2742	−2.1466++	0.1825	0.4354++	0.1618
Motorist characteristics									
Has been drinking (vs. no)	Yes	0.6281++	0.2004						
Pedestrian characteristics									
Age (in years, base: ≥ 55)	< 25	−0.2602+	0.1258						
	25 - 54	−0.5664++	0.1109						
Land development									
Land use (base: Residential)	Commercial	−0.5605++	0.1185						
	Institutional	−0.6159++	0.2054						
Vehicle characteristics									
Vehicle type (base: Light Truck (Mini-Van, Panel)	Heavy Vehicle	1.0183++	0.2626						
Roadway characteristics									
Road class type (base: US Route)	Private Road, Driveway	0.8684++	0.1744						
Crash characteristics									
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)	Pedestrian Crossing	−0.8485++	0.3064						
	Backing Vehicle	−0.3355++	0.0958						
Rural or urban (vs. no)	Urban	−0.4196++	0.1117						

No. of observations, 2261.
-2×Log-likelihood at convergence, 3388.357.
-2×Log-likelihood (constant only), 3556.410.
+Level of significance > 95%. ++Level of significance > 99%.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

Table 3d
PPO Model for Class 4 in Pedestrian-Vehicle Crashes.

Class 4		Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Variable		All Level		K[d]		A[c]		B[b]	
Intercept				−5.1293++	0.4674	−3.7029++	0.3943	−1.2624++	0.3671
Motorist characteristics									
Gender (base: female)	Male	0.2027+	0.0891						
Pedestrian characteristics									
Age (in years, base: ≥ 55)	< 25	−0.8738++	0.1627						
	25 - 54	−0.8197++	0.1675						
Land development									
Land use (base: Residential)	Industrial	2.9886++	0.9701						
Environmental factors									
Weather (base: Other)	Clear	0.5363+	0.2272						
	Cloudy	0.5603+	0.2502						
Vehicle characteristics									
Vehicle type (base: Light Truck (Mini-Van, Panel)	Sport Utility	0.3364++	0.1186						
Roadway characteristics									
Road class type (base: US Route)	Local Street	0.6231++	0.2365						
	North Carolina state route	0.8906+	0.4035						
	North Carolina secondary state route	2.6061++	0.7612						
Road type (base: Two-Way, Not Divided)	Two-Way, Divided	0.4906++	0.1157						
Crash characteristics									
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)	Midblock	0.7233++	0.1972						
	Pedestrian Crossing			0.3732	0.4340	0.4838+	0.2342	0.8923++	0.1366
	Pedestrian in Roadway	0.5689++	0.1853						
	Dash / Dart-Out			0.5988	0.4121	0.6932++	0.2277	1.3088++	0.1414
	Pedestrian Waiting to Cross	3.3831+	1.5248						

No. of observations, 2040.
-2×Log-likelihood at convergence, 3689.59.
-2×Log-likelihood (constant only), 3867.856.
+Level of significance > 95%. ++Level of significance > 99%.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

lighted roadway with dark light condition) differently (marginal effects +0.0614 and +0.0320, respectively). Young pedestrians (< 24) and mid-age pedestrians (25–54) are less likely to suffer fatal injuries, particularly for class 2 (marginal effects -0.1550 and -0.0946) and class 6 (marginal effects -0.1149 and -0.0639). Male pedestrians in class 6 have higher chance to be fatally injured in crashes (marginal effects +0.0223). "Hit and run" has very significant impacts on both severity levels of capacitating and fatal injury for class 1, 2 and 6, but the degrees of effect are different, which can be seen in Tables 4a and 4b. It is found that pedestrians in class 4 (i.e., crashes on residential local street) are more likely to be negatively affected by male drivers, while no statistical relationships are observed for other classes. This comes as no surprise, since previous studies have shown that male drivers are more likely to take risks (Deery, 1999), and engage in aggressive driving.

5.3.2. Vehicle characteristics

Heavy vehicle involved crashes have higher chances to cause severe injury (capacitating and fatal injury) in almost all classes (except for class 4), but this chance is highest for class 1 (marginal effects +0.3400). Similar conclusions were drawn in (Aziz et al., 2013). The only statistical relationship is observed for class 4 with contributing factor being "Sport Utility Vehicle". It is also observed for class 6 and contributing factor of "Public Vehicle" with effect on reducing risk of being fatally injured (marginal effects -0.1032). Crashes occurred involving car only have impacts on class 5 (i.e., crashes with pedestrians crossing the roads and traffic signal being present) and 6.

5.3.3. Crash characteristics

This category is classified as "Crash Groups" in the police-reported crash datasets. Variables in this category are highly related to the behaviors of both drivers and pedestrians. For crashes happening midblock, they only have impacts on class 1 and 4, and the associated marginal effect indicates the highest negative impact on "Fatal Injury" for class 1 (marginal effects +0.0800). For crashes that occurred while pedestrians cross roadways, the risks for being severely injured are higher for almost all classes (except for class 3) with highest marginal effects being on class 2 (marginal effects +0.1933). Similar trends can also be seen for the variable of "Dash / Dart-Out" (marginal effects +0.2157 for class 3, much higher than other classes). The impacts of backing vehicles involved in crashes show significance only in class 3. Such unique impacts are also true for "pedestrian waiting to cross" in class 4, and "working in roadway" in class 6. Differences of impacts of pedestrians in roadway involved in crashes between class 4 and 6 can be obviously seen, especially for the negative effects on increasing risks of fatal injury (marginal effects +0.0127 for class 4 and +0.0802 for class 6).

Table 3e
PPO Model for Class 5 in Pedestrian-Vehicle Crashes.

Class 5		Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Variable		All Level		K[d]		A[c]		B[b]	
Intercept				−5.2185++	0.7243	−3.3551++	0.4764	−1.1239++	0.4204
Pedestrian characteristics									
Age (in years, base: ≥ 55)	25 - 54			−0.7282+	0.3316	−0.8181++	0.2096	−0.1554	0.0959
Traffic control									
Control type (base: Traffic Signal)	Traffic Sign	−0.378++	0.1238						
Environmental factors									
Weather (base: Other)	Clear			1.5541++	0.6005	0.532+	0.2512	0.2343+	0.1115
Vehicle characteristics									
Vehicle type (base: Light Truck (Mini-Van, Panel)	Car			−1.3461++	0.3881	−0.6454++	0.2049	−0.1729	0.0971
	Heavy Vehicle			2.0772++	0.4515	2.0423++	0.3598	0.9688++	0.3441
Roadway characteristics									
Road class type (base: US Route)	Local Street	−0.4132++	0.1604						
	Public Vehicular Area	−0.9787++	0.3709						
Road type (base: Two-Way, Not Divided)	Two-Way, Divided	0.3891++	0.0996						
Crash characteristics									
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)	Pedestrian Crossing	1.0200++	0.3804						
	Dash / Dart-Out	1.4353++	0.4261						

No. of observations, 1918.
-2×Log-likelihood at convergence, 3257.291.
-2×Log-likelihood (constant only), 3396.861.
+Level of significance > 95%. ++Level of significance > 99%.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

5.3.4. Time of day and weather

For the weather variable group, various impacts show the heterogeneities between classes across severity levels, and cloudy and rain only show significances on class 4 and 6, respectively. Clear weather condition has impacts on class 4 and 5. For crashes happening during midnight (i.e.,12:00am to 4:00am), there are higher risks for fatal injury for pedestrians in class 2 (marginal effects +0.0924). Similar trend can be seen for early morning time (4:00am to 8:00am) in class 2 (marginal effects +0.1866). No statistical relationships are observed with variables in crash time group to class 3–5, which confirm the heterogeneities in the data.

5.3.5. Roadway characteristics

Straight roadway shows positive impacts on mitigating the injury severity for pedestrians in class 1, and no other relationships are found for other classes. Freeways involved in crashes will heavily increase the chance of being severely injured for class 1 (marginal effects +0.1044 for capacitating injury and +0.2344 for fatal injury) and class 2 (marginal effects +0.2179 for fatal injury). Local streets show almost positive mitigations of the injury severity for pedestrians, especially for fatal injury of pedestrians in class 6. The significances of "North Carolina state route" and "North Carolina secondary state route" only exhibit on pedestrians in class 4, and the latter one greatly increases the chances of pedestrians suffering from severe injuries (marginal effects +0.2498 for capacitating injury and +0.1711 for fatal injury). Public vehicle areas involved in crashes seem to decrease the probability of being severely injured for pedestrians in only class 5 and 6. Except for class 3, there are no statistical relationships between crash injury severity and private road/driveway. Two-way divided roadway will highly increase the chance for pedestrians being killed in class 2 (marginal effects +0.0994).

5.3.6. Other factors

According to Tables 4a and 4b, one can see that crashes occurred on urban roadways could decrease the probability of severe injuries in only class 3. The reason could come from differences of the speed and speed limit between urban and rural areas. Drivers usually drive slower on urban roadways than on rural roadways due to the traffic conditions. Mueller et al. (1988) also pointed out that pedestrian crashes in rural areas are generally located further from quality emergency care and delays in emergency services would be a key factor. Crashes happening in industrial areas have higher chance to be severely injured for class 4 (marginal effects +0.2775 for capacitating injury and +0.2349 for fatal injury). "No Control Present" could increase the probabilities of being seriously injured for pedestrians in crashes, only for pedestrians in class 6 (marginal effects +0.0336), which matches the fact and logic. On contrast, "Traffic Sign" could mitigate the injury severity, but only shows its significance for class 5.

Compared to the authors' previous work (Li and Fan, 2018), and additionally the variances of the estimations and marginal effects between classes and across the injury severity level do confirm the assumption that performing traffic crash analysis of a large heterogeneous data set can obscure significant relations, which is also found by other researchers as mentioned in the literature review section. Thus, some variables are significant only within specific classes, which can provide additional and more insightful information, leading to more accurate and instructive analyses of the contributing factors for further improving the safety of pedestrians.

Table 3f
PPO Model for Class 6 in Pedestrian-Vehicle Crashes.

Class 6		Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Variable		All Level		K[d]		A[c]		B[b]	
Intercept				−1.5763^{++}	0.2224	−0.8689^{++}	0.2173	0.7859^{++}	0.2164
Motorist characteristics									
Has been drinking (vs. no)	Yes	0.2794^{++}	0.0821						
Hit and run (vs. no)	Yes			0.6002^{+}	0.2895	0.997^{++}	0.2318	0.4998	0.2601
Pedestrian characteristics									
Age (in years, base: ≥ 55)	< 25			−1.2280^{++}	0.1835	−1.0382^{++}	0.1450	−0.5294^{++}	0.1357
	25 - 54	−0.5470^{+}	0.1089						
Gender (base: female)	Male	0.2041^{+}	0.0843						
Time									
Time (base: 8:01 PM - 12:00)	12:01 AM - 4:00 AM	0.3973^{++}	0.1123						
	4:01 AM - 8:00 AM	0.4517^{++}	0.1357						
Traffic control									
Control type (base: Traffic Signal)	No Control Present	0.3111^{++}	0.0856						
Land development									
Land use (base: Residential)	Commercial	0.2189^{++}	0.0834						
Environmental factors									
Weather (base: Other)	Rain	−0.2690^{+}	0.1112						
Vehicle characteristics									
Vehicle type (base: Light Truck (Mini-Van, Panel)	Car			−0.5941^{++}	0.1220	−0.3243^{++}	0.0975	−0.0473	0.0955
	Heavy Vehicle	0.8916^{+}	0.3955						
	Public (Police, etc.)	−1.485^{+}	0.743						
Roadway characteristics									
Road class type (base: US Route)	Local Street	−0.7380^{++}	0.1083						
	Public Vehicular Area	−1.3580^{++}	0.3954						
Road type (base: Two-Way, Not Divided)	One-Way	−0.4740^{+}	0.2393						
Crash characteristics									
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)	Pedestrian Crossing	0.6157^{++}	0.1267						
	Pedestrian in Roadway	0.6779^{++}	0.1564						
	Working in Roadway	−1.0840^{+}	0.4810						
	Dash / Dart-Out	0.6223^{++}	0.1536						

No. of observations, 2416.
-2×Log-likelihood at convergence, 5827.473.
-2×Log-likelihood (constant only), 6143.813.
$^{+}$Level of significance > 95%. $^{++}$Level of significance > 99%.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

6. Summary and conclusions

This research aims at exploring the differences of the factors contributing to pedestrian injury severities between different latent classes in pedestrian-vehicle crashes. Five separate PPO models and one ordered logit model are developed using the police reported pedestrian crash data collected from 2007 to 2014 in North Carolina. The result shows that instead of developing a single discrete choice model using all data, latent class segmented sub-models are more effective and useful in identifying the targeted contributing factors for populations, which is in agreement with the work of Yau (2004); Depaire et al. (2008) and Sun et al. (2019). A variety of motorist, pedestrian, environmental, and roadway characteristics are inspected to identify the contributing factors for each latent class. Marginal effects of variables in all models are also computed and used to illustrate the effects of key factors that significantly influence the injury severity levels in pedestrian-vehicle crashes in North Carolina. Differences between six classes do exist when looking into the interpretations which clearly confirm the superiority of developing models considering potential latent variables.

Back to the results of the marginal effects, though heterogeneities exist, several general major factors contributing to severe crashes are found, such as the heavy vehicle involved. Thus, such vehicle types' relevant flows or movements may need to be restricted at locations (also time periods) with high pedestrian activities. In the meantime, proper trainings for raising awareness among drivers of heavy vehicles could be helpful. On the other hand, installing warning signs as alerts

Table 4a
PPO Average Marginal Effects of Explanatory Variables (Class 1–3).

Variables	Class 1				Class 2				Class 3			
	Crash injury severity				Crash injury severity				Crash injury severity			
	C/O[a]	B[b]	A[c]	K[d]	C/O[a]	B[b]	A[c]	K[d]	C/O[a]	B[b]	A[c]	K[d]
Alcohol involved (vs. no)					−0.0614	−0.0097	0.0097	0.0614	−0.1394	0.1077	0.0249	0.0069
Pedestrian age (base ≥ 55)												
Young (< 25)					0.0699	0.0756	0.0095	−0.1550	0.0526	−0.0426	−0.0078	−0.0021
Mid (25-54)	0.0783	−0.0221	−0.0266	−0.0295	0.0926	0.0149	−0.0129	−0.0946	0.1171	−0.0954	−0.0172	−0.0046
Vehicle type (base: Light Truck (Mini-Van, Panel))												
Heavy vehicle	−0.1056	−0.2252	−0.0093	0.3400	−0.2307	0.1692	0.0478	0.0137	−0.2307	0.1692	0.0478	0.0137
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)												
Midblock	−0.1438	0.0059	0.0579	0.0800								
Pedestrian Crossing	−0.1526	0.0192	0.0572	0.0762	−0.1637	−0.1229	0.0933	0.1933	0.1518	−0.1278	−0.0191	−0.0049
Pedestrian in Roadway					−0.1074	−0.1110	0.0330	0.1854				
Back vehicle									0.0688	−0.0559	−0.0102	−0.0027
Dash/dart-Out	−0.1892	0.0111	0.0745	0.1037	−0.1560	−0.0695	0.0097	0.2157				
Time (base: 8:01 PM - 12:00)												
12:01 AM - 4:00 AM					−0.0824	−0.0210	0.0110	0.0924				
4:01 AM - 8:00 AM	−0.1010	0.0131	0.0389	0.0490	−0.1418	−0.0573	0.0125	0.1866				
Road class (base: US Route)												
Freeway	−0.2459	−0.0930	0.1044	0.2344	−0.1513	−0.0753	0.0086	0.2179				
Private Road, Driveway									−0.1979	0.1514	0.0365	0.0099
Straight road (vs. curved)	0.1420	−0.0085	−0.0573	−0.0762								
Roadway type (base: Two-Way, Not Divided)												
Two-way, divided					−0.0872	−0.0240	0.0117	0.0994				
Urban roadway (vs. rural)									0.0910	−0.0727	−0.0144	−0.0038
Land development (base: Residential)												
Commercial									0.1219	−0.0980	−0.0190	−0.0050
Farms, Woods, Pastures	−0.0651	0.0143	0.0234	0.0274								
Institutional	0.1790	−0.0746	−0.0518	−0.0527					0.1162	−0.0967	−0.0155	−0.0040
Hit and run (vs. no)	−0.1907	−0.0291	0.0818	0.1380	−0.0476	−0.2084	0.1058	0.1501				

[a] C/O - No/Possible injury.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

for pedestrians with heavy vehicle flows are essential. Similar suggestions can also be found in (Aziz et al., 2013). In accordance with (Sun et al., 2019), pedestrian behaviour is another crucial issue, particularly crossing and dash/dart-out that show great increments of the risk being fatally injured. It should also be noted that the compared to older pedestrians (> 55), other pedestrian age groups seem less likely to be severely injured in crashes. Despite these general factors, some variables are significant only within specific classes, which provides additional and more insightful information and leads to more accurate and instructive analyses of the contributing factors in order to further improve the safety of pedestrians.

Overall, this study shows the necessity of identifying targeted contributing factors for pedestrian safety improvements, by conducting latent class clustering analysis for segmenting the large heterogenous dataset. In the meantime, it is well noted that there are several limitations of the current approaches and in that regard, some future research efforts are still needed. The data used in this study range from 2007 to 2014, and it is well understood that temporal instability may exist due to the global recession (Behnood and Mannering, 2016; Mannering, 2018). Thus, the latent class clustering analysis may be picking up temporal homogeneities in nature, which could have impact on the accuracy of the model and therefore may also affect the conclusions derived from the model results. According to (Mannering, 2018), temporal instability can affect different modelling approaches in different ways. Therefore, such issue needs more cautions when interpreting the results and also requires advanced models with the capability of handling temporal heterogeneity in future research studies.

It should be noted that this study applies the latent class clustering and models separately and sequentially. According to the studies of (Xiong and Mannering, 2013; Yasmin et al., 2014a,b), though the sequential procedure is much easier to estimate, there is a possibility to deploy a more efficient approach to modelling the latent classes and the injury severities simultaneously, which is certainly worth exploring in the future. In the meantime, instead of applying the fixed effect PPO model, either the mixed effect generalized ordered model or mixed effect PPO model (Eluru and Yasmin, 2015) could be further utilized and integrated in the latent class analysis to capture the unobserved heterogeneity. Last, the latent variables identified by LCC in this study can be treated as single level variables for partitioning the whole dataset, where the correlations between each variable are not obvious. More elaborate and multidimensional data clustering techniques, such as latent tree analysis (Liu et al., 2014; Zhang and Poon, 2017) could be developed for crash data analysis in the future.

Table 4b
PPO Average Marginal Effects of Explanatory Variables (Class 4–6).

Variables	Class 4				Class 5				Class 6			
	Crash injury severity				Crash injury severity				Crash injury severity			
	C/O[a]	B[b]	A[c]	K[d]	C/O[a]	B[b]	A[c]	K[d]	C/O[a]	B[b]	A[c]	K[d]
Alcohol involved (vs. no)									−0.0549	0.0031	0.0197	0.0320
Male driver (vs. female)	−0.0466	0.0330	0.0100	0.0035								
Pedestrian age (base ≥ 55)												
Young (< 25)	0.1888	−0.1229	−0.0481	−0.0178					0.1083	0.0629	−0.0562	−0.1149
Mid (25-54)	0.1861	−0.1359	−0.0369	−0.0132	0.0358	0.0066	−0.0283	−0.0141	0.1064	−0.0061	−0.0364	−0.0639
Male pedestrian (vs. female)									−0.0409	0.0044	0.0143	0.0222
Vehicle type (base: Light Truck (Mini-Van, Panel)												
Car					0.0400	−0.0044	−0.0104	−0.0252	0.0094	0.0509	0.0096	−0.0699
Heavy vehicle					−0.2295	−0.0058	0.1343	0.1010	−0.1464	−0.0386	0.0559	0.1291
Sport utility	−0.0766	0.0519	0.0181	0.0065								
Public (Police, etc.)									0.3304	−0.1410	−0.0861	−0.1032
Crash group (Pedestrian behavior associated with Motorist maneuver, base: Pedestrian Walking Along Roadway)												
Midblock	−0.1569	0.0941	0.0455	0.0174								
Pedestrian Crossing	−0.1950	0.1594	0.0285	0.0071	−0.2016	0.1602	0.0261	0.0154	−0.1243	0.0157	0.0420	0.0665
Pedestrian in Roadway	−0.1252	0.0787	0.0339	0.0127					−0.1209	−0.0131	0.0450	0.0890
Working in Roadway									0.2411	−0.0870	−0.0681	−0.0860
Dash/dart-Out	−0.2888	0.2374	0.0397	0.0117	−0.3276	0.1969	0.0782	0.0525	−0.1127	−0.0086	0.0411	0.0802
Pedestrian Waiting to Cross	−0.4307	−0.1695	0.2868	0.3133								
Weather (base: Other)												
Clear	−0.1226	0.0909	0.0235	0.0082	−0.0535	0.0275	0.0048	0.0211				
Cloudy	−0.1248	0.0801	0.0326	0.0121								
Rain									0.0551	−0.0080	−0.0187	−0.0284
Time (base: 8:01 PM - 12:00)												
12:01 AM - 4:00 AM									−0.0747	−0.0015	0.0275	0.0487
4:01 AM - 8:00 AM									−0.0834	−0.0045	0.0310	0.0568
Road class (base: US Route)												
Local Street	−0.1426	0.1090	0.0250	0.0086	0.0976	−0.0719	−0.0160	−0.0096	0.1326	0.0154	−0.0516	−0.0964
North Carolina state route	−0.1895	0.1048	0.0607	0.0240								
Public Vehicular Area					0.1960	−0.1587	−0.0237	−0.0136	0.3020	−0.1219	−0.0810	−0.0991
North Carolina secondary state route	−0.3962	−0.0247	0.2498	0.1711								
Roadway type (base: Two-Way, Not Divided)												
One-way									0.1004	−0.0218	−0.0324	−0.0462
Two-way, divided	−0.1116	0.0744	0.0273	0.0099	−0.0915	0.0692	0.0140	0.0083				
Traffic control (base: Traffic Signal)												
No Control Present									−0.0630	0.0077	0.0218	0.0336
Traffic Sign					0.0853	−0.0665	−0.0119	−0.0069				
Urban roadway (vs. rural)					0.0963	−0.0705	−0.0161	−0.0097				
Land development (base: Residential)												
Commercial									−0.0439	0.0043	0.0155	0.0242
Industrial	−0.4160	−0.0963	0.2775	0.2349								
Hit and run (vs. no)									−0.0904	−0.1171	0.1278	0.0797

[a] C/O - No/Possible injury.
[b] B - Non-incapacitating (evident) Injury.
[c] A - Incapacitating Injury.
[d] K - Fatal Injury.

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.

Acknowledgments

The authors want to express their deepest gratitude to the financial support by the United States Department of Transportation, University Transportation Center through the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE) at The University of North Carolina at Charlotte (Grant Number: 69A3551747133).

References

Aziz, H.A., Ukkusuri, S.V., Hasan, S., 2013. Exploring the determinants of pedestrian–vehicle crash severity in New York City. Accid. Anal. Prev. 50, 1298–1309. https://doi.org/10.1016/j.aap.2012.09.034.
Bedard, M., Guyatt, G.H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accid. Anal. Prev. 34 (6), 717–727.
Behnood, A., Mannering, F.L., 2016. An empirical assessment of the effects of economic recessions on pedestrian injury crashes using mixed and latent-class models. Anal. Methods Accid. Res. 12, 1–17.
Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38 (5), 1019–1027.
Chen, F., Chen, S., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accid. Anal. Prev. 43 (5), 1677–1688. https://doi.org/10.1016/j.aap.2011.03.026.
Chen, Z., Fan, W.D., 2019a. A multinomial logit model of pedestrian-vehicle crash severity in North Carolina. Int. J. Transp. Sci. Technol. 8 (1), 43–52.
Chen, Z., Fan, W., 2019b. Modeling pedestrian injury severity in pedestrian-vehicle crashes in rural and urban areas: mixed logit model approach. Transp. Res. Rec

0361198119842825.
Dai, D., 2012. Identifying clusters and risk factors of injuries in pedestrian–vehicle crashes in a GIS environment. J. Transp. Geogr. 24, 206–214. https://doi.org/10.1016/j.jtrangeo.2012.02.005.
Deery, H.A., 1999. Hazard and risk perception among young novice drivers. J. Saf. Res. 30 (4), 225–236.
Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40 (4), 1257–1266.
Derr, B., 2013. Ordinal response modeling with the LOGISTIC procedure. SAS Global Forum. SAS Institute, Inc., Cary, NC, pp. 1–20.
Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accid. Anal. Prev. 40 (3), 1033–1054.
Eluru, N., Yasmin, S., 2015. A note on generalized ordered outcome models. Anal. Methods Accid. Res. 8, 1–6.
Friedman, J.H., 1998. Data Mining and Statistics: What's the connection? Comput. Sci. Stat. 29 (1), 3–9.
Gong, L., Fan, W., Washing, E.M., 2016. Modeling severity of single vehicle run-off-road crashes in rural areas: model comparison and selection. Can. J. Civ. Eng. 43 (6), 493–503.
Hair, J.F., Anderson, R.E., Tatham, R.L., William, C., Black, 1998. Multivariate Data Analysis. 1998. .
Haleem, K., Gan, A., 2015. Contributing factors of crash injury severity at public highway-railroad grade crossings in the US. J. Saf. Res. 53, 23–29.
Haleem, K., Alluri, P., Gan, A., 2015. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accid. Anal. Prev. 81, 14–23.
Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. J. Saf. Res. 37 (3), 267–276.
Kashani, A.T., Mohaymany, A.S., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf. Sci. 49 (10), 1314–1320.
Kim, J.K., Ulfarsson, G.F., Shankar, V.N., Kim, S., 2008a. Age and pedestrian injury severity in motor-vehicle crashes: a heteroskedastic logit analysis. Accid. Anal. Prev. 40 (5), 1695–1702. https://doi.org/10.1016/j.aap.2008.06.005.
Kim, J.K., Ulfarsson, G.F., Shankar, V.N., Kim, S., 2008b. Age and pedestrian injury severity in motor-vehicle crashes: a heteroskedastic logit analysis. Accid. Anal. Prev. 40 (5), 1695–1702.
Kim, J.K., Ulfarsson, G.F., Shankar, V.N., Mannering, F.L., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accid. Anal. Prev. 42 (6), 1751–1758.
Kim, K., Yamashita, E.Y., 2007. Using ak-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. J. Adv. Transp. 41 (1), 69–89.
Lanza, S.T., Collins, L.M., Lemmon, D.R., Schafer, J.L., 2007. PROC LCA: a SAS procedure for latent class analysis. Struct. Eq. Model.: A Multidiscip. J. 14 (4), 671–694.
Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accid. Anal. Prev. 37 (4), 775–786.
Li, Y., Fan, W., 2018. Modelling the severity of pedestrian injury in pedestrian—vehicle crashes in North Carolina: a partial proportional odds logit model approach. J. Transp. Saf. Secur. 1–22.
Li, Y., Fan, W., 2019. Pedestrian injury severities in pedestrian-vehicle crashes and the partial proportional odds logit model: accounting for age difference. Transp. Res. Rec 0361198119842828.
Liu, T., Zhang, N.L., Chen, P., 2014. Hierarchical latent tree analysis for topic detection. September. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, pp. 256–272.
Malyshkina, N.V., Mannering, F.L., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accid. Anal. Prev. 42 (1), 131–139.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40 (1), 260–266.
Mohamed, M.G., 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.
Mooradian, J., Ivan, J.N., Ravishanker, N., Hu, S., 2013. Analysis of driver and passenger crash injury severity using partial proportional odds models. Accid. Anal. Prev. 58, 53–58.
Moudon, A.V., Lin, L., Jiao, J., Hurvitz, P., Reeves, P., 2011. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accid. Anal. Prev. 43 (1), 11–24.
Mueller, B.A., Rivara, F.P., Bergman, A.B., 1988. Urban-rural location and the risk of dying in a pedestrian-vehicle collision. J. Trauma 28 (1), 91–94.
Peel, D., McLachlan, G.J., 2000. Robust mixture modelling using the t distribution. Stat. Comput. 10 (4), 339–348.
Quddus, M.A., Wang, C., Ison, S.G., 2009. Road traffic congestion and crash severity: econometric analysis using ordered response models. J. Transp. Eng. 136 (5), 424–435. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000044.
Rifaat, S.M., Tay, R., De Barros, A., 2012. Severity of motorcycle crashes in Calgary. Accid. Anal. Prev. 49, 44–49.
Sasidharan, L., Menéndez, M., 2014. Partial proportional odds model—an alternate choice for analyzing pedestrian crash injury severities. Accid. Anal. Prev. 72, 330–340.
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.
Sun, M., Sun, X., Shan, D., 2019. Pedestrian crash analysis with latent class clustering method. Accid. Anal. Prev. 124, 50–57.
Sze, N.N., Wong, S.C., 2007. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accid. Anal. Prev. 39 (6), 1267–1278.
Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian–vehicle crash severity. Int. J. Sustain. Transp. 5 (4), 233–249. https://doi.org/10.1080/15568318.2010.497547.
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., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors for fatal road traffic accidents in Udine, Italy. Accid. Anal. Prev. 34 (1), 71–84.
Vermunt, J.K., Magidson, J., 2002. Latent class cluster analysis. Appl. Latent Class Anal. 11, 89–106.
Wang, X., Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Accid. Anal. Prev. 40 (5), 1674–1682. https://doi.org/10.1016/j.aap.2008.06.001.
Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall/CRC.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press.
Xiong, Y., Mannering, F., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: a finite-mixture random-parameters approach. Transp. Res. Part B Methodol. 49, 39–54.
Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014a. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal. Methods Accid. Res. 1, 23–38.
Yasmin, S., Eluru, N., Ukkusuri, S.V., 2014b. Alternative ordered response frameworks for examining pedestrian injury severity in New York City. J. Transp. Saf. Secur. 6 (4), 275–300.
Yau, K.K., 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accid. Anal. Prev. 36 (3), 333–340.
Zhang, N.L., Poon, L.K., 2017. Latent tree analysis. February. Thirty-First AAAI Conference on Artificial Intelligence.
Zhou, Z.P., Liu, Y.S., Wang, W., Zhang, Y., 2013. Multinomial logit model of pedestrian crossing behaviors at signalized intersections. Discrete Dyn. Nat. Soc. 2013. https://doi.org/10.1155/2013/172726.