Safety Science 54 (2013) 27–37

Contents lists available at SciVerse ScienceDirect

Safety Science

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

A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada

Mohamed Gomaa Mohamed [a,*], Nicolas Saunier [a,1], Luis F. Miranda-Moreno [b,2], Satish V. Ukkusuri [c,3]

[a] Department of Civil, Geological and Mining Engineering, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3A7
[b] Department of Civil Engineering and Applied Mechanics, McGill University, Room 268, Macdonald Engineering Building, 817 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6
[c] School of Civil Engineering, Purdue University, West Lafayette, IN 47907-2051, United States

ARTICLE INFO

Article history:
Received 1 March 2012
Received in revised form 8 November 2012
Accepted 10 November 2012
Available online 21 December 2012

Keywords:
Pedestrian safety
Contributing factors
Latent class
Clustering injury severity
Pedestrian–driver characteristics
Built environmental

ABSTRACT

Understanding the underlying relationship between pedestrian injury severity outcomes and factors leading to more severe injuries is very important in addressing the problem of pedestrian safety. This research combines data mining and statistical regression methods to identify the main factors associated with the levels of pedestrian injury severity outcomes. This work relies on the analysis of two unique pedestrian injury severity datasets from New York City, US (2002–2006) and the City of Montreal, Canada (2003–2006). General injury severity models were estimated for each dataset and for sub-populations obtained through clustering analysis. This paper shows how the segmentation of the accident datasets helps to better understand the complex relationship between the injury severity outcomes and the contribution of geometric, built environment and socio-demographic factors. While using the same methodology for the two datasets, different techniques were tested. Within the New York dataset, a latent class with ordered probit method provides the best results. However, for Montreal, K-means with a multinomial logit model proves most appropriate. Among other results, it was found that pedestrian age, location type, driver age, vehicle type, driver alcohol involvement, lighting conditions, and several built environment characteristics influence the likelihood of fatal crashes. Finally, the research provides recommendations for policy makers, traffic engineers, and law enforcement in order to reduce the severity of pedestrian–vehicle collisions.

1. Introduction

Pedestrian safety is a vital transportation issue when promoting active transportation. Pedestrians are vulnerable road users often suffering serious consequences when involved in motor-vehicle crashes. Therefore, it is important to understand the factors associated with pedestrian injury severity levels. This will help traffic engineers, planners and decision makers to target the injury-related factors through various engineering counter-measures (such as improvements to motorized vehicles, pedestrian facility designs, and built environment and road geometric design), as well as education and enforcement actions (referred to as the 3-E approach).

This paper combines the use of regression modeling techniques with clustering analysis to identify the main contributing factors associated with pedestrian–vehicle injury severity levels in two case study locations: New York City, US and Montreal, Canada. The relationship of injury severity levels and a large set of factors (covering built environment, geometric design, and vehicle–pedestrian characteristics) is investigated.

The paper is organized into five sections. The following section provides a review of previous studies on injury severity modeling. The methodologies used in this research are described in the third section. The fourth section presents the data, to which a clustering algorithm and injury severity regression model are applied. The fifth section reports and analyzes the results of the different methods and the final section concludes the work.

2. Related work

Many researchers have attempted to establish crash consequence models to determine the injury severity of pedestrians involved in motor-vehicle accidents. Eluru et al. (2008) categorized the risk factors considered in earlier studies into the following six categories: (1) pedestrian characteristics (e.g. age, gender, state of soberness), (2) motorized vehicle driver characteristics (e.g. state

* Corresponding author. Tel.: +1 514 340 5121x4210.
E-mail addresses: mohamed.gomaa@polymtl.ca (M.G. Mohamed), nicolas.saunier@polymtl.ca (N. Saunier), luis.miranda-moreno@mcgill.ca (L.F. Miranda-Moreno), sukkusur@purdue.edu (S.V. Ukkusuri).
[1] Tel.: +1 514 340 4711x4962.
[2] Tel.: +1 514 398 6589; fax: +1 514 398 7361.
[3] Tel.: +1 765 494 2296.

0925-7535/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ssci.2012.11.001

of soberness, age), (3) motorized vehicle characteristics (e.g. vehicle type, speed), (4) roadway characteristics (e.g. speed limit, road system) (5) environmental factors (e.g. time, weather conditions), and (6) crash characteristics (e.g. vehicle motion prior to accident).

In addition to these variables, researchers recently started looking into characteristics of the built environment (Aziz et al., 2012; Clifton et al., 2009; Ukkusuri et al., 2012; Zahabi et al., 2011). Clifton et al. (2009) studied the effect of built environment and other characteristics on pedestrian–vehicle crashes. Regarding the individual and behavioral variables, they found that older individuals are more likely to be fatally injured. With respect to characteristics of the built environment, although they examined many built environment variables, only network connectivity and transit access had a significant influence in non-fatal injury and were negatively associated with sustaining minor injury. They concluded that built environmental characteristics should be considered when evaluating and planning for pedestrian safety. Zahabi et al. (2011) estimated the effects of road design, built environment, speed limit, and other factors on the injury severity levels of pedestrians and cyclists involved in a collision with a motorized vehicle. Their research found that factors significantly increasing pedestrian collision severity include presence of a major road, vehicle straight movements, darkness, median income, transit access, mixed land use, and park presence within 10 meters. Furthermore, they found that accidents occurring at an intersection and near a school have a lower pedestrian severity. In another study (Sze and Wong, 2007), the authors explored the contributing factors that lead to mortality and severe injury in crashes involving pedestrians in Hong Kong during the period of 1991–2004. They considered the effect of demographic, crash, environmental, geometric, and traffic characteristics. They found that the factors that increase the probability of fatal and severe injury include elderly people above 65, head injuries, a speed limit above 50 km/h, and if a crash is at either a crossing or close to a crosswalk, at a signalized intersection, or on a road with two or more lanes. In contrast, some factors that are associated with lower injury severity include male, time of day, and if the footpath is obstructed or overcrowded.

Recently, Ukkusuri et al. (2012) investigated the link between the frequency of pedestrian–vehicle accidents classified by injury severy types and built environment variables, including land use patterns, demographics, transit charactristics and road network characteristics. The authors used the same accident dataset from New York City (NYC) as in this paper. The analysis was conducted at the zip code and census tract levels. The results showed the effect of built environment on pedestrian safety. For example, multilane roads increase the likelihood of fatal and total pedestrian crashes. In addition, land use patterns affect the likelihood of pedestrian crashes; commercial, industrial and open land use types increase the likelihood for crashes while residential land use has opposite effect. A borough level analysis using the same NYC dataset was conducted by (Aziz et al., 2012). They divided the dataset into five separate datasets depending on the borough of the accident location. Then, they explored the contributing factors associated with the levels of pedestrian injury severity outcomes in each borough. The findings showed the importance of using separate models for each borough instead of analyzing the whole dataset as one. Consequently, the suggested countermeasures are different in each borough.

There are several statistical methods that can be used for analyzing the crash severity, such as ordered logit or probit models (Lee and Abdel-Aty, 2005; Zahabi et al., 2011), generalized logit models (Clifton et al., 2009), multinomial logit models (Tay et al., 2011), and binary logit models (Sze and Wong, 2007). Data mining has been used for data exploration and analysis in many scientific areas for years. Among the data mining techniques, classification methods such as decision trees, non-linear regression, and clustering techniques such as latent class (LC), K-means have been the most popular data mining techniques. In the field of safety analysis, some researchers trained a decision tree to analyze the injury severity (Chang and Wang, 2006; Prato et al., 2010) and reported satisfying results in prediction and classification. Other researchers analyzed accidents by clustering using K-means (Kim and Yamashita, 2007; Prato et al., 2010) and LC (Depaire et al., 2008). Finally, some researchers have recommended combining data mining and statistical techniques. Kuhnert et al. (2000) combined a nonparametric model like Classification And Regression Trees (CARTs) and Multivariate Adaptive Regression Splines (MARSs) with logistic regression to analyze motor vehicle injury data. They suggested that CART and MARS can be used as a precursor to a more detailed logistic regression analysis. Depaire et al. (2008) used LC as a preliminary analysis to identify hidden relationships between severity outcomes and contributing factors, and then applied the multinomial logit model to injury analysis. They found that this methodology is more powerful compared to applying only a multinomial logit model to the whole dataset. More recently, Eluru et al. (2012) have used a latent segmentation based ordered logit model for identifying vehicle driver injury severity factors at highway-railway crossings.

3. Methodology

While each of the models used in the safety literature has its advantages, it appears that the injury severity regression model is the most common technique used to identify the relationship between the dependent and independent variables. Also, it calculates the significance level of each variable, although there may be hidden significant variables that must be considered in specific cases. However, the effect of a particular factor might vary across collision subgroups. To address this issue, one solution is to classify homogeneous accidents into clusters that can make other relationships appear.

3.1. Clustering analysis

Clustering means to classify the data into groups (clusters) with similar characteristics. It is a category of unsupervised learning methods developed in the discipline of machine learning that has been applied to data mining, pattern recognition, and image processing. There are many clustering algorithms. The most popular clustering algorithms are hierarchical, partitioning, density based, and grid based. For further reading, the readers are referred to (Berkhin, 2002; Xu and Wunsch, 2005). In this study, we focus on partitioning clustering, which divides the data into k clusters with no hierarchical relationship. There are two approaches for clustering:

- The first approach relies on a distance between the dataset elements. The algorithm attempts to maximize the similarity within each cluster and the dissimilarity between clusters. The best known algorithm in this category is K-means.
- The second approach is probabilistic. It considers that the data comes from a mixture model of several probability distributions.

Both approaches, in the form of K-means and latent class (LC), are used in this study. LC is known as a finite mixture model and theoretically is similar to fuzzy clustering as it considers each element class membership uncertainty. The main difference is that in fuzzy clustering, the membership levels are the estimated parameters, while in LC, each element cluster membership is computed

from the estimated model parameters. LC analysis has become more common for clustering over the last few years as faster computers make the computations manageable. Among the available packages for LC analysis, one can mention the software Latent GOLD 4.5, which was used in this study. The basic LC cluster form is (Vermunt and Magidson, 2002):

$$f(\mathbf{z}_i|\theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{z}_i|\theta_k) \quad (1)$$

where z_i is a vector of observed variables from the ith crash outcome, K is the number of clusters, π_k denotes the prior probability of membership in latent class or cluster k, θ_k is the cluster model parameters and $f_k(\mathbf{z}_i|\theta_k)$ is the mixture probability density.

LC parameter estimation is based on maximum likelihood (ML). Since ML solutions cannot be obtained analytically, the expectation–maximization algorithm is used for iterative estimation (Berkhin, 2002; Xu and Wunsch, 2005). LC deals with model selection (number of clusters) by trying multiple models and computing various information criteria such as the Bayesian Information Criteria (BIC), Akaike Information Criterion (AIC), and Consistent Akaike Information Criterion (CAIC). The appropriate number of clusters is the one that minimizes the score of these criteria. LC is advantageous to traditional partitioning clustering methods, such as K-means, in that it does not depend on a distance between the elements; there is no need to normalize or standardize the data before processing. Consequently, variables of different types (ordinal, count, nominal, continuous) can be included in the analysis without special processing (Depaire et al., 2008).

3.2. Injury severity models

The ordered probit (OP) regression model is commonly used for analyzing datasets that include categorical and ordered dependent variables such as the pedestrian injury severity levels. The structural model can be written as (Borooah, 2002; Jackman, 2000):

$$y_i^* = \sum_{k=1}^{k} \beta_k X_{ki} + \varepsilon_i \quad (2)$$

where y_i^* is the injury severity risk, which is an unobserved continuous variable called latent variable ranging from $-\infty$ to ∞, and is mapped to an observed variable y_i, x_{ki} is a row vector of independent variables such as pedestrian, driver, vehicle, road and built environment characteristics (not including a constant). Moreover, β is a vector of parameters to be estimated from the data and ε_i is the error term, which is assumed to be normally distributed For example, the value of the dependent variables y_i in the case of three categories is then determined as:

$$y_i = \begin{cases} \text{category 1} & \text{if } y_i^* \leqslant \tau_1 \\ \text{category 2} & \text{if } \tau_i^* \leqslant y_i^* \leqslant \tau_2 \\ \text{category 3} & \text{if } y_i^* \geqslant \tau_2 \end{cases} \quad (3)$$

where the τ_1 and τ_2 values are called the thresholds or cut-off points of the categories. The threshold values are parameters to be estimated from the data. According to the measurement model, the probability that the ith crash has a severity level of m (m = 1–3) is the probability that the injury risk y^* takes a value between two cut-off points (see (Borooah, 2002; Jackman, 2000)).

As an alternative technique, the multinomial logit (MNL) model can be used instead of OP model when considering three or more severity outcomes (Washington et al., 2010). In some cases, the multinomial model can be more flexible and allows for estimating the effect of independent variables in each severity category relative to the base outcome case (Tay et al., 2011). In other words, any contributing factor may be significant in one category but not significant in other categories or in the whole dataset, making it easier to interpret the results. The probability of pedestrian k being injured with severity category i is expressed as

$$P_k(i) = \frac{e^{\beta_k x_{ki}}}{\sum_{k=1} e^{\beta_k x_{ki}}} \quad (4)$$

Finally, a common measure of overall model fit used for both models is the ρ^2 statistic, with $\rho^2 = 1 - LL(\beta)/LL(0)$ (Washington et al., 2010), with $LL(\beta)$ being the log likelihood at convergence with parameter vector β and $LL(0)$ being the initial log likelihood (with all coefficients set to zero). The estimation of both model parameters was carried out through maximum likelihood approach using SPSS software.

4. Context and data

The analyzed pedestrian–vehicle collision datasets were built combining different sources of information for the Cities of New York and Montreal. The NYC dataset is the main data in this study as it contains more contributing variables. The NYC dataset was obtained from New York City Department of Transportation (NYCDOT), and was processed by CUBRC, Buffalo, NY. The data includes the information reported by the police officers for each accident from 2002 to 2006. This information contains important variables describing the characteristics of the accident and injury severity. Complementary information was added from three sources. The source of most variables is the New York State Department of Transportation (NYSDOT) – Safety Information Management System (SMS). To examine the built environment and design characteristics, two other sources of data were used: (1) the Primary Land Use Tax Lot Output (PLUTLO™) data files, which provided land use variables, and (2) the New York City Department of Transportation (NYCDOT), which provided: travel lane, park lane, road width, existence of a truck route within 50 feet, bus route, subway station, metered parking, and bike on street.

In the NYC dataset, the accidents with a fatal or severe injury outcome were analyzed. We removed the accidents with property damage only as they represent a small share of the dataset and this category of accident is known to be largely under-reported. A total of 6896 pedestrian–vehicle accidents were used for injury severity analysis. The dependent variable is the crash injury severity, while the potential contributing factors are summarized in Table 1. All possible values for nominal variables were used in the clustering process but only the values that represented more than 1% of the whole dataset (not marked as italics in Table 1) were used in the regression model. Fatal pedestrian crashes accounted for 9.6% of accidents and 90.4% were classified as an injury.

For Montreal, the primary source of the secondary dataset is the Quebec's auto insurance company (SAAQ – *Société de l'Assurance Automobile du Québec* for the years 2003–2006). This dataset was previously used by (Zahabi et al., 2011). This source provided the following variables: road type (local, major, highway), accident location at intersection (yes/no), type of movement (straight, left turn, right turn, reverse), vehicle type (automobile, van/truck/bus (VTB), motorcyclist, emergency vehicle), environmental condition (after dark, bad weather), visibility (bad due to weather, bad due to object) and built environment (population density, transit accessibility, network connectivity, land use mix, school presence, park presence, hospital presence, etc.). Again, for more details one can refer to (Zahabi et al., 2011).

A total of 5,820 pedestrian–vehicle collisions were observed in this dataset. There are three categories of outcome: no injury, minor injury, and fatal crash. Their proportions are 6.1%, 81.6% and 12.3%, respectively. It is important to note that many variables

Table 1
Independent variables.

Variable	Values[a]
1-Pedestrian characteristics	
Gender	Male, female, *unknown*
Age	Under 5, 5–15, 15–25, 25–40,40–65, over 65, *unknown*
Location	At intersection, *not at intersection, unknown*
Pedestrian action prior to accident	Crossing with signal, crossing against signal, crossing, no signal, marked crosswalk, crossing, no signal or crosswalk, along highway with traffic, along highway against traffic, emerged behind parked vehicle, *child getting on/off school bus, getting on/off vehicle*, working in roadway, playing on roadway, other action in roadway, not in roadway, *unknown*
2-Vehicle and driver characteristics	
Gender	Male, female, *unknown*
Age	Under 26, 26–50, 50–65, over 65, *unknown*
Vehicle type	Moto, car/van/pick up, truck, bus, *other*
Location	First event occurs on road, *off road, unknown*
Vehicle movement prior to accident	Going straight ahead, making right turn, making left turn, making u-turn, starting from parking, starting in traffic, slowed or stopped, *stopped in traffic*, entering parked position, *parked, avoiding object in roadway*, changing lanes, overtaking, merging, backing, *making right turn on red, making left turn on red, police pursuit, other, unknown*
Primary factors of accident	Alcohol involvement, backing unsafely, driver inattention, driver inexperience, drug (illegal), failure to yield right of way, *fell asleep, following too closely, illness, lost consciousness, passenger distraction*, passing or lane usage improperly, pedestrian's error/confusion, *physical disability, prescription medication*, traffic control devices disregarded, *turning improper*, unsafe speed, *unsafe lane changing, cell phone(hand held), cell phone(hands free), other electronic device, outside car distraction, reaction to other uninvolved vehicle, failure to keep right*, aggressive driving/road rage, *other (human), animal's action,* glare, *obstruction/debris, pavement defective*, pavement slippery, *traffic control device improper/non-working*, view obstructed/limited, *other (environmental), unknown*
3-Environmental condition	
Weekday (Monday to Friday)	Weekday = 1, weekend = 0
Season	Winter (December–January–February), Autumn (September–October–November), Summer (June–July–August), Spring (March–April–May)
Accident time	7 a.m. to 9:59 a.m., 10 a.m. to 3:59 p.m., 4 p.m. to 6:59 p.m., 7 p.m. to 6:59 a.m., unknown
Borough	*Bronx, Brooklyn, Manhattan, Queens, Staten Island*
Road surface	Dry, wet, *muddy*, snow/ice, slush, *flooded water, other, unknown*
Weather	Clear, cloudy, rain, snow, sleet/hail/freezing rain, *fog/smog/smoke, other, unknown*
Light condition	Daylight, dawn, dusk, dark lighted, dark unlighted, *unknown*
4-Built environmental variable	
Land use	Single or double family residential, multi-family residential, mixed residential and commercial, commercial/office, industrial/manufacturing, transportation/utility, public facilities and institutions, open space, parking facilities, vacant land, misc. lots, unknown
Special features (within 50 feet)	Truck route, bus route, near subway station, metered parking, on street bicycle lanes
5-Network variables	
Road system	*State, country*, town, city street, parkway, parking lot, other non-traffic, *interstate unknown*
Road characteristics	Straight and level, straight/grade, straight at hillcrest, curve and level, curve and grade, *curve and hillcrest, unknown*
Traffic control	None, traffic signal, stop sign, *flashing light, yield sign, officer/flagman/guard, no passing zone, rr crossing sign, rr crossing flash light, stopped school bus with red light flash, highway work area (construction), maintenance work area, utility work area, police/fire emergency, school zone, other, unknown*
No. of travel lanes	Zero lane, One lane, Two lane, Multi lane
Park lane	Existing park lane = 1, Other = 0
Road width[b]	*Less than 10 feet, 10–20, 20–30, 30–42, 42–65, More than 65 feet*

[a] In clustering analysis, all values were used. In regression, those values marked in italics were excluded.
[b] The road width variable was excluded from regression because it is correlated with the number of travel lanes.

available in the NYC dataset are not available in the Montreal dataset such as age and gender of pedestrians and drivers. Nevertheless, it will be useful for examining the proposed methodology and exploring the shared contributing variables in injury severity.

5. Results and discussion

5.1. New York case study

5.1.1. Latent class analysis

Vehicle–pedestrian crashes were clustered by using all the available variable values in Table 1. To select the appropriate number of clusters in the final model, different numbers of clusters were tested, from one to eleven. The BIC, AIC, and CAIC criteria were used to select the final number of clusters. As shown in Fig. 1, BIC decreases until seven clusters and increases for eight clusters; for nine clusters the lowest score is observed, and then increases again. On the other hand, AIC decreases monotonically as the number of clusters increases. BIC is more reliable than AIC, especially for large datasets (Vermunt and Magidson, 2002). CAIC has its lowest score for seven clusters. Furthermore, the quality of the clustering solution was assessed by calculating the entropy R squared criterion. The closer the criterion is to 1, the better the clustering. The entropy R squared is equal to 0.9344 and 0.9308 for seven and nine clusters, respectively, both of which are quite high. Based on the BIC and CAIC, it is preferred to use seven clusters.

The final model was described by the proportion of each variable in each cluster. Similarly to the work of (Depaire et al., 2008), the clusters were analyzed and named based on their variable distributions. For example, if one cluster has 95% at 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 presented in Table 2. For cluster 1, the variables are traffic control, pedestrian location before the accident, and lighting conditions. With respect to traffic control, signalized

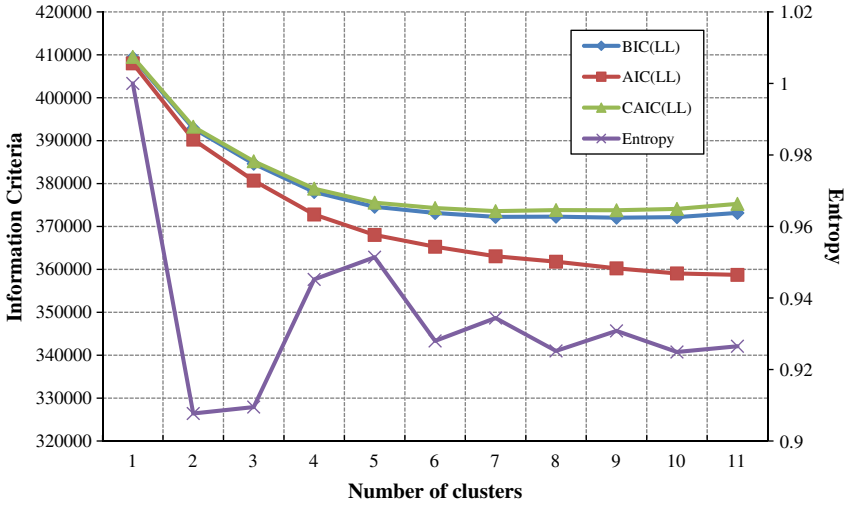

Fig. 1. Variation of BIC, AIC, CAIC and Entropy values for model selection.

Table 2
Summary of interesting variables and their distribution in each cluster.[a].

Variables	Whole data	C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)	C6 (%)	C7 (%)
Fatal crash	9.6	11.0	9.4	7.7	7.0	13.3	9.7	6.5
Injury crash	90.4	89.0	90.6	92.3	93.0	86.7	90.3	93.5
Pedestrian location at intersection	71.8	97.5	95.7	79.9	53.9	44.0	39.1	67.7
Pedestrian action unknown	13.6	14.5	12.7	14.9	8.0	6.8	11.3	84.2
Road surface unknown	3.1	0.0	0.2	0.8	0.4	0.2	1.0	99.4
Weather unknown	3.2	0.2	0.4	0.8	0.7	0.7	1.3	97.6
Road characteristics								
Dry	76.3	94.3	92.8	91.6	91.1	89.9	84.6	0.9
Unknown	3.1	0.2	0.0	1.1	0.5	0.6	1.2	98.5
Traffic control								
Non-signalized	39.8	4.7	4.3	32.5	75.1	82.4	73.1	1.7
Signalized	51.4	92.0	92.2	57.6	12.6	13.3	19.5	7.6
Unknown	4.8	1.3	1.6	4.1	1.7	2.4	4.3	90.2
Light condition								
Daylight	53.9	97.5	5.0	43.9	67.6	54.8	58.1	2.1
Dark with light	34.9	0.2	80.7	45.9	24.7	35.4	31.1	1.6
Unknown	3.6	0.5	0.1	2.4	0.8	1.0	1.6	96.2
Travel lane number								
Zero lane	12.2	0.1	0.2	0.2	0.0	0.0	99.9	18.4
One lane	25.3	20.7	20.4	31.2	69.5	4.8	0.1	21.4
Two lane	35.7	45.6	41.8	39.7	30.4	43.7	0.0	35
Multi lane	26.9	33.5	37.7	28.8	0.1	51.5	0.0	25.3
Park existence	73.4	79.2	77.1	82.4	97.2	84.5	0.0	63.9
Road width under 10 feet	11.8	0.0	0.0	0.0	0.0	0.0	97.6	17.8
Land use = parking facilities	15.4	2.9	3.6	3.5	4.1	6.0	98.6	22.1
Vehicle type								
Car/pickup/van	72.0	80.4	83	25.8	91	83.3	70	60.3
Other	21.4	7.9	12.2	71.2	5.0	7.3	24	35.9
Motion prior accident								
Straight	59.6	46.6	57.6	58.8	68.5	75.6	64.9	14.4
Unknown	6.6	2.1	2.6	13.4	2.7	3.0	5.2	75.3
Driver age unknown	20.9	0.0	0.0	100.0	0.0	0.0	25.1	44.6
Driver sex								
Male	63.7	79.2	86.3	0.0	77.5	79.6	59.3	45.8
Unknown	20.9	0.0	0.0	100.0	0.0	0.0	25.0	44.6
Primary factor unknown	49.7	40.9	46.3	51.3	49.9	54.2	53.4	87.6

[a] For the complete results, contact the authors.

traffic control represents approximately 92.0% of the crashes in this cluster. For the pedestrian location, the accident occurs at an intersection in 97.5% of the cases. The lighting condition in this cluster is daylight for approximately 97.5% of the cases. Consequently, we referred to cluster 1 as "Accidents at signalized intersections in daylight". The other clusters were classified similarly. Cluster 2 is similar to cluster 1 for signalized intersections but distinguishes itself by an over-representation of dark conditions. Cluster 3 reveals the

missing values with regards to driver characteristics and vehicle type, data which are missing in many collision reports, such as the Montreal dataset in our case study. The special features of cluster 4 are the number of travel lanes and the existence of a parking lane. In addition, the involved vehicle is a car/van/pickup in 91% of the cases in this cluster. Analysis of accidents based on vehicle type was recommended by (Depaire et al., 2008; Yau, 2004).

Three variables are specific to cluster 5: traffic control is non-signalized (82%), the vehicle motion before the accident is straight (75.6%) and the number of travel lanes is two or more (95.2%). Cluster 6 describes the accidents that occur in a part of the road network that are less than 10 feet wide (98%) and have no travel lane (99.9%), which corresponds to parking facilities (99%). Finally, cluster 7 contains only about 2.7% 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 in more homogeneous groups and to identify the higher order variables that may have an influence on injury severity. Table 3 shows an overview of the cluster descriptions and the size of each cluster.

5.1.2. Injury severity analysis using OP

As the goal of this study is to explore the variables influencing the occurrence of fatal crashes, an OP model was applied in which the severity output was considered as the dependent variable. For that purpose, the values of categorical variables were converted into binary variables (“dummies”). Seven models were built, one for the whole dataset and one for each cluster except cluster 7, for which too many values are missing. Because each cluster describes a specific accident category, the independent variables that characterize those categories were excluded from the regression analysis. For example, cluster 1 describes signalized intersections in daylight, Hence, traffic control and light condition variables were eliminated from the cluster 1 regression analysis. The estimated coefficients, their significance level and the log likelihood of the model are shown in Table 4. The examination of results depended on the statistical significance of the coefficients of the independent variables. The significance level used in this study is 10%. We built the model considering an injury crash as the base case. Therefore, a positive coefficient sign means a higher probability of a fatal crash.

Table 3
Cluster descriptions and accident categories.

Cluster No.	Category	Cluster label	Proportion of whole dataset
Cluster 1	Accidents happening at signalized intersections in daylight	SigDay	20.6% (1420 cases)
Cluster 2	Accidents happening at signalized intersections in dark conditions with light	SigNit	17.7% (1223 cases)
Cluster 3	Missing driver information	MissDri	16.8% (1160 cases)
Cluster 4	Accidents involving a car/van/pickup, traveling in one or two lanes with a park lane	CVP	15.5% (1072 cases)
Cluster 5	Accidents involving a straight movement and happening in two or more travel lanes in non-signalized parts of the road system	StrNSI	15.0% (1037 cases)
Cluster 6	Accidents taking place at parking facilities	Park	11.6% (798 cases)
Cluster 7	Multiple missing values	MissVal	2.7% (186 cases)

5.1.3. General logistic regression analysis

With respect to the pedestrian characteristics, pedestrians aged 40–65 and more than 65 were more likely to be involved in fatal crashes. Focusing on pedestrian actions prior to the accident, the dataset suggests that crossing without a signal or crosswalk, and actions on roadway (different action types on roadway except playing and working) increase the risk of fatal crashes. On the other hand, if the pedestrian crosses at an intersection, the probability of death is decreased. These results are likely due to most drivers paying attention and reducing their speed when they are at an intersection. In addition, “crossing while respecting a signal” is expected to lower chances of a fatal collision.

With respect to vehicle and driver characteristics, male drivers show a significant effect in increasing the risk of a fatal crash. As expected, if the involved vehicle is a truck or a bus, the probability of a fatal crash increases significantly. Alcohol involvement, backing unsafely, failure to yield right of way, disregard of traffic control, unsafe speed, and obstructed or limited views are statistically significant in increasing the risk of a fatal crash. Vehicles being in reverse prior to the accident result in the opposite effect. The reason may be that the drivers in reverse drive more slowly and pay more attention.

In terms of environmental conditions, winter and autumn seasons, as well as dawn and dark (lighted or unlighted) time periods increase the probability of a fatal accident. The coefficient for dark unlighted is 1.5 times the coefficient for dark lighted. In this perspective, when roads are lighted, fatal crashes are reduced with respect to unlighted roads. Both clear and bad weather, such as cloudy, rainy and snowy, reduce the probability of a fatal crash. The reason behind reductions in fatalities under bad weather may be that drivers travel more cautiously.

By examining the built environment variables, only the accident location near a metered parking facility was found to have a significant effect, reducing the risk of a fatal crash. Usually, metered parking facilities are located in commercial areas where speeds tend to be lower.

Regarding the network variable, the results showed that town and city streets, parking lots, and other non-traffic road system facilities significantly decrease the likelihood of fatal crashes. In addition, fatality probability increases when the number of lanes increases. Interestingly, these variables have a direct link with vehicle speeds and the speed limit.

5.1.4. Cluster-based logistic regression analysis

In this section, the results of the injury risk analysis are finally reported by cluster. Comparing the overall model with each cluster model, three different situations arise for each variable:

- Case A: the variable is significant only within each accident category (cluster), which will provide additional information.
- Case B: the variable is significant in both the overall model and the cluster model.
- Case C: the variable is significant in the overall model but not significant in the cluster model.

Cases A and B are particularly interesting since they show the information provided by the clustering. Variables corresponding to cases A and B are presented for each cluster in Table 6. The results were interpreted systematically for each cluster, and they are explored for cluster 1 as an example. Cluster 1 is the category of collisions at signalized intersections in daylight, and several variables belong to case A; pedestrians less than 5 years old, driver age and sex, vehicle movement prior to accident, built environment

Table 4
Ordered probit model results for whole dataset and each cluster.[a].

	Injury outcome is the base case													
	Whole dataset		SigDay		SigNit		MissDri		CVP		StrNSI		Park	
Model characteristics														
Constant	−2.581		−1.475		−10.892		−3.135		−4.234		−4.963		−6.344	
Log Likelihood at zero coefficient	4356.856		990.016		772.536		619.619		530.099		803.336		506.215	
Log Likelihood at convergence	3586.109		739.825		633.477		473.241		382.986		606.986		317.806	
ρ^2	0.177		0.253		0.180		0.236		0.278		0.244		0.372	
Variables	Whole dataset		SigDay		SigNit		MissDri		CVP		StrNSI		Park	
	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.
Pedestrian characteristics														
Gender														
Male					0.247	0.049			−0.293	0.057				
Pedestrian age														
Under 5 years			1.008	0.086									1.113	0.037
Between 5 and 15 years							−0.780	0.040						
Between 15 and 25 years							−0.828	0.012						
Between 40 and 65 years	0.369	0.000	0.723	0.034	0.508	0.029					0.592	0.009		
Over 65 years	1.014	0.000	1.604	0.000	0.794	0.002			0.942	0.003	1.245	0.000	0.727	0.032
Pedestrian location														
Pedestrian at intersection	−.442	0.000					−0.615	0.000	−0.773	0.000	−0.527	0.000		
Pedestrian action prior to be involved in the accident														
Crossing with signal	−0.189	0.041	−0.439	0.005			−0.534	0.012						
Crossing against signal			−0.262	0.097	0.473	0.013								
Crossing, no signal or crosswalk	0.165	0.073			0.512	0.093					0.420	0.027		
Along highway with traffic									1.399	0.080				
Playing on roadway													1.393	0.019
Other action in roadway	0.274	0.013												
Vehicle and driver characteristics														
Gender														
Male	0.128	0.068	0.249	0.091										
Driver age														
Under 26					0.296	0.043								
Between 26 and 50 years			0.306	0.079									0.744	0.039
More than 65											0.967	0.000	1.199	0.014
Vehicle type														
Moto											1.128	0.047		
Car/van/pickup													−0.524	0.044
Truck	0.857	0.000	1.348	0.000	1.163	0.001					1.151	0.001		
Bus	0.724	0.000	1.030	0.000	1.802	0.000					0.940	0.013		
Location														
First event occurred on road									−.597	.094				
Vehicle movement prior accident														
Going straight ahead			0.678	0.019			−0.416	0.008						
Making right turn			0.527	0.100			−0.679	0.056						
Making left turn			0.700	0.018			−0.643	0.052						
Starting from parking													0.808	0.072
Backing	−0.552	0.002							−0.721	0.048				
Primary factors of accident														
Alcohol involvement or drug (illegal)	0.660	0.000	0.994	0.063	0.701	0.004			1.577	0.000	0.529	0.082		
Backing unsafely	0.336	0.078												
Driver inattention			0.274	0.052							−0.611	0.008		
Failure to yield right of way	0.288	0.002	0.476	0.002			0.407	0.084						
Pedestrian's error/confusion					0.345	0.059								
Traffic control devices disregarded	0.440	0.038					0.802	0.015						
Unsafe speed	0.593	0.000			0.857	0.002	0.472	0.067			0.760	0.030		
View obstructed/limited	0.294	0.069							1.047	0.005			1.068	0.014
Environmental condition														
Weekday (Monday to Friday)					−0.265	0.032								
Season														
Winter (December–January–February)	0.153	0.028							0.495	0.035				
Autumn (September–October–November)	0.202	0.003			0.408	0.017			0.561	0.009				
Summer (June–July–August)					0.389	0.043								
Accident time														
7 a.m. To 9:59 a.m.									0.716	0.021				
4 p.m. To 6:59 p.m.											−0.461	0.032		
7 p.m. To 6:59 a.m.									0.516	0.043				

(*continued on next page*)

Table 4 (*continued*)

Variables	Whole dataset		SigDay		SigNit		MissDri		CVP		StrNSI		Park	
	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.
Weather														
Clear	−0.688	0.002									−1.113	0.021		
Cloudy	−0.519	0.025									−1.296	0.011		
Rain	−0.592	0.016									−1.312	0.011		
Snow	−1.233	0.003												
Light condition														
Dawn	0.625	0.036					1.041	0.089						
Dark lighted	0.598	0.025												
Dark unlighted	0.979	0.002					1.149	0.074						
Built environmental variables														
Land use														
1 and 2 Family residential			−0.752	0.050										
Mixed residential and commercial							0.867	0.037						
Public facilities and institutions							0.845	0.063						
Parking facilities							1.152	0.015						
Special features														
Located on bus route (or within 50 feet)			−0.314	0.050										
Located near metered parking (within 50 feet)	−0.172	0.003					−0.283	0.073	−0.673	0.017	−0.246	0.068		
Network variables														
Road system														
Town	−1.550	0.004												
City street	−1.222	0.000											−1.580	0.000
Parking lot. Other non-traffic	−1.101	0.004											−1.209	0.012
Traffic control														
None									−0.616	0.083				
No. of travel lanes														
One lane	0.453	0.007												
Two lane	0.570	0.001												
Multi lane	0.639	0.000												

[a] Only significant variables are shown in these tables: contact the authors for complete results.

variables and driver inattention have an influence on the probability of fatal crash. The following variables were also significant, in this cluster and in the whole dataset (case B): pedestrians aged over 40, crossing with signal, heavy vehicle, alcohol involvement, and failure to yield right of way.

The effect of some variables changes direction between certain clusters and the reason is unclear. It is, for example, not clear why being a male pedestrian increases the probability of a fatal crash at a signalized intersection with dark lighting condition (cluster 2) and decreases for accidents involving a car/van/pickup which happen on roads with one or two lanes and a park lane (cluster 4). Furthermore, driver inattention increases the probability of fatal crashes at signalized intersections (cluster 1) and has the opposite effect at non-signalized road sections for accidents involving straight movements (cluster 5). A change in sign was also observed for vehicle movement prior to accidents in clusters 1 and 3 and for pedestrian crossing against signal in clusters 1 and 2. These opposite effects show the interaction between pedestrian crashes and different network variables. They cannot be simply explained and may indicate the need to validate some observations more closely.

5.2. Montreal case study

5.2.1. Clustering analysis

K-means was preferred for the Montreal dataset. LC put about 90% of the dataset in the first two clusters, regardless of the selected number of clusters, and it was difficult to describe the accidents in each cluster. K-means classified the data into 5 clusters relying on type of movement and environmental conditions. Cluster 1 contains the accidents related to vehicles in reverse (11%). Cluster 2 contains accidents occurring in bad weather and dark lighting conditions (21.5%). Cluster 3 contains the accidents with left turn movements at intersections (23.4%). Cluster 4 contains collisions involving a straight movement (32.4%). Cluster 5 contains the collisions involving a right turn (11.7%).

5.2.2. Injury severity using MNL

Since there are three categories of injury severity, the MNL model is more appropriate for analyzing this dataset. A model of the whole dataset and five models for each cluster were examined. Crashes without injuries were selected as a reference (base) case for the dependent variables. Consequently, the estimated coefficients show the effects of a contributing factor on the probability of a fatal or minor injury relative to a no-injury crash. Table 5 summarizes the coefficient estimation for the Montreal dataset.

Focusing on the whole dataset, variables that significantly increase the probability of fatal crash are straight movement, right turn, VTB, after dark, median income, transit access, mixed use and park presence. Conversely, variables that significantly decrease the probability of fatal collisions are accidents at intersection and connectivity factor. On the other hand, significant variables that increase the probability of minor injury are after dark, bad visibility due to objects and median income.

For the cluster-based analysis, Table 6 summarizes the variables contributing to fatal and minor injury for each cluster corresponding to case A and case B. Similar to the NYC dataset, bad visibility increases the likelihood of fatality. An important finding is that the presence of a hospital reduces fatal crashes. It was unexpected

Table 5
MNL model estimation for the Montreal dataset.[a].

Base case : no injury	Whole dataset		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
Variables	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.	Coeff.	*P* val.
Fatal crash												
Intercept	1.868		2.056		7.044		0.765		−3.774		0.068	
Type of road (Ref. local road)												
Highway											1.407	0.068
Accident at intersection	−0.359	0.022			−1.077	0.011			−0.659	0.009		
Type of vehicle movement at accident (Ref. other)												
Straight	0.808	0.002			1.698	0.012						
Right turn	0.673	0.041										
Type of vehicle dummy categories (automobile category is the base case)												
Vans, trucks, buses (VTB)	0.286	0.069	0.789	0.105			0.561	0.070				
Environmental condition												
After dark	0.738	0.000							1.112	0.001		
Visibility (Ref. good vis.)												
Visibility obstructed due to bad weather					0.923	0.009						
Visibility obstructed due to an object											2.284	0.033
Built environmental characteristics												
Median Income (in 1000$)	0.012	0.029	−0.031	0.075					0.019	0.023		
Population density (in 1000 capita/km^2)							0.000	0.100				
Transit access	0.022	0.024							0.049	0.005	0.061	0.059
Connectivity	−0.512	0.082										
Mixed-use (HHI/1000)	0.049	0.008							0.098	0.002	0.211	0.001
Park present in 10 m distance	0.473	0.072										
Hospital Presence									−2.754	0.029		
Minor injury												
Intercept	2.869		2.224		4.177		1.674		11.318		1.629	
Type of road (Ref. local road)												
Major road					0.587	0.041						
Highway					1.014	0.106						
Accident at intersection							0.662	0.030	−0.360	0.103		
Environmental condition												
After dark	0.346	0.011							0.605	0.041	−0.636	0.101
Visibility (Ref. good vis.)												
Visibility obstructed due to bad weather					0.563	0.068			−2.050	0.002		
Visibility obstructed due to an object	0.571	0.003									1.737	0.091
Built environmental characteristics												
Median Income (in 1000$)	0.01	0.033			0.023	0.070	0.025	0.043				
Mixed-use (HHI/1000)							−0.072	0.026			0.095	0.041
Log Likelihood at zero coefficient	6636.130		633.049		1455.638		1435.271		2321.007		718.895	
Log Likelihood at convergence	6452.763		602.512		1364.872		1389.738		2234.546		660.600	
ρ^2	0.028		0.048		0.062		0.031		0.037		0.081	

[a] Only significant variables are shown in this table: contact the authors for complete results.

that mixed use in cluster 3 and after dark variable in cluster 5 would reduce the chance of minor injury.

6. Conclusion

This paper investigates the link between pedestrian injury severity outcomes and a rich set of factors associated to the built environment, geometric design, demographics, vehicle characteristics and pedestrian and driver features. For this purpose, a cluster-based regression model was implemented. Clustering analysis yielded clusters based on crash characteristics such as traffic control, lighting conditions, vehicle type, land use, type of movement, environmental conditions, and missing attributes. Once the dataset was segmented, specific types of accidents (clusters) were separately analyzed. The clustering and parameters explain different features of the models, which complement each other to provide a more detailed analysis.

By clustering the dataset, this work confirms that segmenting the traffic accident dataset into homogeneous subsets helps identify important contributing factors that would be hidden if the whole dataset was used. Thus, it is recommended that clustering be used not only for descriptive analysis, but also as a preliminary segmentation tool for a more detailed, standard statistical analysis.

In terms of the contributing factors, several variables were common in the two case studies their effect was confirmed in both cities. Heavy vehicles, dark lighting conditions, mixed land use, and major roads increase the probability of fatal crashes. In addition, crossing at intersections lowers the severity. These results support the following recommendations. Truck flows or movements at intersections with high pedestrian activity should be restricted, or it should be attempted to concentrate truck traffic at times of low pedestrian activity. This would be part of a general strategy of reducing exposure of pedestrians to heavy vehicles traffic. In

Table 6
Contributing variables for each cluster in NYC and Montreal case studies.

New York case study			
Cluster #	Impact on fatality probability	Case A	Case B
Cluster 1	Increase	Pedestrians aged under 5; male driver; driver aged 26–50 years; straight motion; right turn; and left turn; driver inattention	Pedestrians aged 40–65; over than 65; heavy vehicle (truck, bus); alcohol involvement; failure to yield right of way
	Decrease	Single or double family residential land use; bus route existence within 50 feet	Crossing with signal
Cluster 2	Increase	Male pedestrian; crossing against signal; driver aged under 26; summer season; primary factor concerning pedestrian's error/ confusion	Pedestrians aged 40–65; more than 65; crossing no signal or sidewalk; heavy vehicle (truck, bus); alcohol involvement; unsafe speed; and winter and autumn season
	Decrease	Accident happening in weekday	
Cluster 3	Increase	Mixed residential and commercial; public facilities and institutions; parking facilities	Failure to yield right of way; traffic control devices disregarded; Unsafe speed; dawn; dark unlighted
	Decrease	Pedestrian aged 5–15; 15–25; motion prior accident either straight; right turn; left turn	Accident at intersection; crossing with signal; effect of existence of metered parking near the accident
Cluster 4	Increase	Crossing along highway with traffic; time of accident 7 a.m. to 9:59a.m and 7 p.m. to 6:59a.m.	Pedestrian over 65 years; alcohol involvement; obstructed/limited view; winter and autumn season
	Decrease	Male pedestrian; first event happen on road; none signalize traffic control	Accident at intersection; backing; effect of existence of metered parking near the accident
Cluster 5	Increase	Driver aged more than 65; motorcyclist	Crossing without signal or crosswalk; truck and bus; pedestrian aged 40–65 and over 65; alcohol involvement; and unsafe speed
	Decrease	Time from 4 p.m to 7 p.m; driver inattention	Accident at intersection; effect of existence of metered parking near the accident; weather (clear; cloudy; rain)
Cluster 6	Increase	Pedestrian aged fewer than five; driver aged 26–50 years; over 65 years; motion prior accident if it is starting from parking; Playing on roadway	Pedestrian aged over 65; obstructed/limited view
	Decrease	Car/van/pickup	City street; parking lot or non-traffic road system

Cluster #	Impact on probability	Fatal crash		Minor injury	
		Case A	Case B	Case A	Case B
Montreal case study					
Cluster 1	Increase		Van/truck/bus		
	Decrease		Median income		
Cluster 2	Increase	Bad visibility due to bad weather.	Straight	Major road; highway; bad visibility due to bad weather	Median income
	Decrease		Accident at intersection		
Cluster 3	Increase		Van/truck/bus	Accident at intersection	Median income
	Decrease	Population density		Mixed use	
Cluster 4	Increase		After dark; Median income; transit access; mixed use		After dark
	Decrease	Presence of hospital	Accident at intersection	Accident at intersection; bad visibility due to bad weather	
Cluster 5	Increase	Highway; Bad visibility due to object	Transit access; mixed use	Mixed use	Bad visibility due to object
	Decrease				After dark

addition, a training program can be provided to trucks drivers to raise their awareness of urban areas with high pedestrian activities. Warning signs can be installed for pedestrians in areas of high truck traffic. These recommendations were also suggested by Aziz et al. (2012). Another recommendation is the retrofitting of major roads into complete streets and the improvement of road lighting to increase visibility during night and adverse weather conditions. Also, the type of land use and intersections should be considered in the design of roads to improve safety, in particular in areas that have high pedestrian activity such as mixed residential and commercial zones, and public institutions.

Other contributing variables influencing crash severity were found in the analysis of the NYC dataset. With respect to the pedestrian characteristics, older pedestrians are the most prone to fatal injuries in pedestrian–vehicle crashes. This is in accordance with the current literature. Pedestrians under 5 years old are also more likely to be involved in fatal crashes. Moreover, pedestrians crossing in the absence of a signal or crosswalk increase the likelihood of fatal crash. This suggests there should be pedestrian signals at most signalized intersections where it is warranted by pedestrian volume. In terms of vehicle and driver characteristics, disregard of traffic control devices and bad visibility increase the likelihood of fatal accidents. Hence, it is important that traffic engineers ensure good visibility of traffic devices and law enforcement ensure that traffic regulations are respected. This could be implemented by targeting intersections with a high number of infractions. Also, pedestrian error or confusion is considered as one of the reasons for fatal crashes at signalized intersections in dimly-lit conditions. When examining the built environment, the existence of a bus route and on-street bike lane at signalized intersections, and metered parking reduce the risk of fatal crashes. These build environment features seem related to denser and more urbanized areas where pedestrians are more numerous and vehicle speeds are lower which may explain the observed association.

Future research should examine different built environment characteristics to identify more countermeasures to help policy makers, planners, and traffic engineers improve safety. The contradicting coefficients between different clusters for the same variable should be further studied. The link between observed operating speeds and injury levels should also be investigated. This could help in understanding crash injury severity mechanisms. Also, the effect of both data period and number of years used in the analysis should be investigated. These two factors can also affect the outcome (parameter estimates). Despite that 5 years of data (2002–2006) were used in this research and that these years are relatively recent, the results should still be validated using more recent years.

Acknowledgments

We acknowledge the NYCDOT Pedestrian Safety Project along with CUBRC who assisted in putting together the NYC Pedestrian database. The authors also thank SAAQ for contributing the Montreal Dataset, Tom Nosal, Paul St-Aubin, Shaun Burns and Sabrina Chan for their help in proofreading the text. We thank the anonymous journal reviewers for their helpful comments.

References

Aziz, H., Ukkusuri, S.V., Hasan, S., Exploring the determinants of pedestrian–vehicle crash severity in New York City. Accid. Anal. Prev. (2012), in press http://dx.doi.org/10.1016/j.aap.2012.09.034.

Berkhin, P., 2002. Survey of clustering data mining techniques. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.3739 (last visit: 26 Nov. 2012).

Borooah, V.K., 2002. Logit and Probit: Ordered and Multinomial Models. Sage Publication, Thousand Oaks, CA.

Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.

Clifton, K.J., Burnier, C.V., Akar, G., 2009. Severity of injury resulting from pedestrian–vehicle crashes: what can we learn from examining the built environment? Transportation Research Part D: Transport and Environment 14, 425–436.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40, 1257–1266.

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

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. Accident Analysis and Prevention 40, 1033–1054.

Jackman, S., 2000. Models for ordered outcomes. Political Science 200C.

Kim, K., Yamashita, E.Y., 2007. Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. Journal of Advanced Transportation 41, 69–89.

Kuhnert, P.M., Do, K.-A., McClure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics and Data Analysis 34, 371–386.

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis and Prevention 37, 775–786.

Prato, C.G., Bekhor, S., Galtzur, A.M., D., Prashker, J.N., 2010. Exploring the potential of data mining techniques for the analysis of accident patterns. In: 12th WCTR, Lisbon, Portugal.

Sze, N.N., Wong, S.C., 2007. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident Analysis and Prevention 39, 1267–1278.

Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian–vehicle crash severity. International Journal of Sustainable Transportation 5, 233–249.

Ukkusuri, S.V., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. Safety Science 50, 1141–1151.

Vermunt, J.K., Magidson, J., 2002. Latent Class Cluster Analysis. Applied Latent Class Analysis. Cambridge University Press, Cambridge.

Washington, S., Karlaftis, M., Mannering, F., 2010. Statistical and econometric methods for transportation data analysis, second ed. CRC Press/Routledge.

Xu, R., Wunsch II, D., 2005. Survey of clustering algorithms. IEEE Transactions on Neural Networks 16, 645–678.

Yau, K.K.W., 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention 36, 333–340.

Zahabi, S., Strauss, J., Manaugh, K., Miranda-Moreno, L., 2011. Estimating potential effect of speed limits, built environment, and other factors on severity of pedestrian and cyclist injuries in crashes. Transportation Research Record: Journal of the Transportation Research Board 2247, 81–90.