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

Mohamed Gomaa Mohamed, Ph.D student (corresponding author)
Department of civil, geological and mining engineering
École Polytechnique de Montréal, C.P. 6079, succ. Centre-Ville
Montréal (Québec) Canada H3C 3A7
Phone: +1 (514) 340-5121 ext. 4210
Email: mohamed.gomaa@polymtl.ca

Nicolas Saunier, Ph.D., Assistant professor
Department of civil, geological and mining engineering
École Polytechnique de Montréal, C.P. 6079, succ. Centre-Ville
Montréal (Québec) Canada H3C 3A7
Phone: +1 (514) 340-4711 ext. 4962
Email: nicolas.saunier@polymtl.ca

Luis F. Miranda-Moreno, Ph.D., Assistant Professor
Department of Civil Engineering and Applied Mechanics, McGill University
Room 268, Macdonald Engineering Building, 817 Sherbrooke Street West
Montreal, Quebec H3A 2K6
Tel: (514) 398-6589
Fax: (514) 398-7361
Email:luis.miranda-moreno@mcgill.ca

Satish Ukkusuri, Ph.D., Assistant professor
School of Civil Engineering, Purdue University
West Lafayette, IN 47907-2051
Tel: (765) 494-2296
Email: sukkusur@purdue.edu

Word count

Text	5814
Tables (6 X 250)	1500
Figures (1 X 250)	250
Total	*7564*

Date of submission: **November15th, 2011**

ABSTRACT

Understanding the underlying relationship between pedestrian injury severity outcomes and factors leading to more severe injuries is very important in dealing with the problem of pedestrian safety. To investigate injury severity outcomes, many previous works relied on statistical regression models. There has also been some interest for data mining techniques, in particular for clustering techniques which segment the data into more homogeneous subsets. This research combines these two approaches (data mining and statistical regression methods) to identify the main contributing factors associated with the levels of pedestrian injury severity outcomes. This work relies on the analysis of two unique pedestrian injury severity datasets from the City of New York, US (2002-2006) and the City of Montreal, Canada (2003-2006). General injury severity models were estimated for the whole datasets and for sub-populations obtained through clustering analysis. This paper shows how the segmentation of the accident datasets help to better understand the complex relationship between the injury severity outcomes and the contributing geometric, built environment and socio-demographic factors. While using the same methodology for the two datasets, different techniques were tested. For instance, for New York, latent class with ordered probit method provides the best results. However, for Montreal, the K-means with multinomial logit model is identified as the most appropriate technique. The results show the power of using clustering with regression to provide a complementary and more detailed analysis. Among other results, it was found that pedestrian age, location at intersection, actions prior to accident, driver age, vehicle type, vehicle movement, driver alcohol involvement and lighting conditions have an influence on the likelihood of a fatal crash. Moreover, several features within the built environment are shown to have an effect. Finally, the research provides recommendations for policy makers, traffic engineers, and law enforcement to reduce the severity of pedestrian-vehicle collisions.

KEYWORDS:

Pedestrian safety, regression, latent class, clustering, severity, built environmental, land use variables

INTRODUCTION

Road user safety is a primary concern, not only for traffic safety specialists and traffic engineers, but for educators and law enforcement as well. Most importantly, pedestrian safety is a vital traffic issue as all road users are pedestrians at one point or another. Since pedestrians are vulnerable road users and suffer more in road crashes, it is important to understand the factors affecting pedestrian injury severity levels. In this way, traffic engineers, planners, decision makers and law enforcement will be able to precisely target these factors through various counter-measures, such as improvements to motorized vehicles, roadway and pedestrian facility design, control strategies at conflict locations, and driver and pedestrian education programs.

This paper examines the integration of regression modeling techniques with clustering analysis to identify the main contributing factors associated with pedestrian-vehicle injury severity levels in two case studies in New York City and Montreal. The effect of a rich set of factors (built environmental, geometric designs, and socio-demographic) on pedestrian safety is investigated.

The paper is organized as follows: 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: a clustering algorithm and injury severity regression model are applied to the whole dataset and to each cluster. The fourth section presents the dataset. The fifth section reports and analyzes the results of the different methods and the final section concludes the work.

RELATED WORK

Many researchers have attempted to establish crash consequence models to determine the injury severity of pedestrian casualties. Eluru et al, (1) categorized the risk factors considered in earlier studies into six following categories variables:{1} pedestrian personal characteristics (e.g. age, gender, alcohol consumption), {2} motorized vehicle driver characteristics (e.g. state of soberness and age), {3}motorized vehicle attributes (e.g. vehicle type and 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 (2), (3). Zahabi et al (2) 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 the pedestrian collision severity are; presence of a major road, vehicle straight movements, darkness, median income, transit access, mixed land use, and park presence within 10 meters. On the other hand, they found that accidents occurring at an intersection and near a school have a lower pedestrian severity. Clifton et al.(3) studied the effect of personal and environmental characteristics on pedestrian-vehicle crashes. Regarding the personal and behavior 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 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.

Lee and Abdel-Aty (4) analyzed vehicle-pedestrian crashes at intersections in Florida. First, they identified correlations between the group of drivers and pedestrians, and traffic and environmental characteristics of locations with high pedestrian crashes using log-linear models. Second, they analyzed the injury severity using the ordered probit (OP) model. They found that older pedestrians, females, pedestrians' alcohol/drug use, vehicle speed, heavy vehicle, adverse weather conditions, dark lighting, and rural areas are contributing factors in increasing the severity. Also, they concluded that the influence of rural areas where there are fewer medical facilities than in urban areas.

Sze and Wong, (5) explored the contributing factors that lead to mortality and severe injury in crashes involving pedestrians in Hong Kong during the period of 1991 to 2004. They considered the effect of demographic, crash, environmental, geometric, and traffic characteristics. They found that the contributing factors increasing the probability of fatal and severe injury include: elderly people above 65, head injuries, crash at a crossing or close to a crosswalk, at a signalized intersection, on a road with two or more lanes, and a speed limit above 50 km/hr. In contrast, male pedestrians, with an age below 15, an accident happening in daytime, and overcrowded or obstructed footpath lower the risk of fatal and severe injury.

There are several statistical methods that can be used for analyzing the crash severity such as ordered logit or probit models (2),(4), generalized logit model (3), multinomial logit model (6), binary logit model(5). 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 (7), (8) and reported satisfying results in prediction and classification. Other researchers analyzed the accidents by clustering using k-means (8),(9) and LC(10). Finally, some researchers have recommended combining data mining and statistical techniques. Kuhner et al (11) combined a non-parametric model like CART and Multivariate Adaptive Regression Splines (MARS) 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(10) used LC as a preliminary analysis to expose the hidden relationships 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 on the whole dataset.

METHODOLOGY

Each of the models covered in this brief literature review has its advantages and disadvantages. Among them, it appears that the injury severity regression model is the most common technique 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. Moreover, the effect of a particular factor might vary across collision subgroups. Then, one solution is to classify homogeneous accidents into clusters that can make other relationships appear.

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 (12) and (13). 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 algorithms here are k-means and k-medoids.
- The second approach is probabilistic based. It considers that the data come from a mixture model of several probability distributions.

Both approaches, respectively in the form of k-means and latent class (LC), will be used in this study. LC is known as a finite mixture model. It is theoretically 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 years as faster computers make the computation manageable. Among the available packages for LC analysis, we chose the software Latent GOLD, version 4.5.

The basic LC cluster form is:

$$f(y_i|\theta) = \sum_{k=1}^{K} \pi_k f_k(y_i|\theta_k)$$

Where y_i is a vector of the i^{th} observation of the observed variables, 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 $f_k(y|\theta)$ 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 (12)(13). 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 has the advantage over traditional partitioning clustering methods such as k-means 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 (10).

Injury severity models

The OP model is commonly used for analyzing datasets that include categorical and ordered dependent variables. In our case, the pedestrian injury severity is a categorical variable. The crash injury severity is related to a number of factors named independent variables including pedestrian, vehicle, driver characteristics, environmental condition, etc. The structural model can be written as (14), (15)

$$y_i^* = \sum_{k=1}^{k} \beta_k x_{ki} + \varepsilon_i$$

where y_i^* is the injury risk, which is an unobserved continuous variable called latent variable ranging from $-\infty$ to ∞, and is mapped to an observed variable y. x_{ki} is a row vector of independent variables, with an intercept value of 1 in the first column and the i^{th} observation for variable k in the $(k+1)^{th}$ column. β is a vector of parameters to be estimated and ε_i is the error term, which is assumed to be normally distributed.

The value of the dependent variables y_i in the case for example of three categories is then determined as:

$$Y_i = \begin{cases} category1 \; if y_i^* \leq \tau_1 \\ category2 \; if \tau_1 \leq y_i^* \leq \tau_2 \\ category3 \; if y_i^* \geq \tau_2 \end{cases}$$

The τ values are called the thresholds or cut-off points of the categories. The threshold values are parameters to be estimated. According to the measurement model, the probability that the i^{th} crash has a severity level of m (m = 1 to 3) is the probability that the injury risk y^* takes a value between two cut-off points. That is,

$$\text{Prob}(\text{y}_\text{i} = 1) = \boldsymbol{\Phi}(\tau_1 - \boldsymbol{\beta}_i \boldsymbol{x_i})$$

$$\text{Prob}(\text{y}_\text{i} = 2) = \boldsymbol{\Phi}(\tau_2 - \boldsymbol{\beta}_i \boldsymbol{x_i}) - \boldsymbol{\Phi}(\tau_1 - \boldsymbol{\beta}_i \boldsymbol{x_i})$$

$$\text{Prob}(\text{y}_\text{i} = 3) = 1 - \boldsymbol{\Phi}(\tau_2 - \boldsymbol{\beta}_i \boldsymbol{x_i})$$

Where, Φ(-) is the cumulative standard normal distribution function.

Multinomial logit (MNL) model is used instead of OP model when considering three or more severity outcomes (16). The multinomial model is more flexible and allows for estimating the effect of independent variables in each severity category relative to the base outcome case (6).In other words, any contributing factor may be significant in one category but not significant in other categories or in the whole data, thus the interpretation of the results can be easier. MNL model is used for the Montreal dataset. The probability of pedestrian k being injured with severity category i is expressed as the following:

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

Finally, a common measure of overall model fit used for both models is the ρ^2 statistic. It is expressed as (16):

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)}$$

Where, *LL(β)* is the log likelihood at convergence with parameter vector *β* and *LL(0)* is 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.

CONTEXT AND DATA

The analyzed pedestrian-vehicle collision datasets were collected in the Cities of New York and Montreal. The New York City (NYC) dataset is the main data in this study as it contains more contributing variables. The primary source of collision attributes comes from the New York State Department of Transportation (NYSDOT). The data were obtained by NYC which included the information reported by the police officer for each accident from 2002 to 2006. This information contains important variables describing the characteristics of the accident and injury severity. To examine the built environment and design characteristics, two other sources of data were used:

- The Primary Land Use Tax Lot Output (PLUTLO™) data files to get the land use variables,
- The New York City Department of Transportation (NYCDOT) to get the following variables: travel lane, park lane, road width, existence of a truck route within 50 feet, bus route, subway station, metered park, and bike on street.

In the NYC dataset, the accidents with a fatal or 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 outcome (or injury severity), while the independent variables for each crash are summarized in Table 1.All possible variable values were used in the clustering process but only the values that represented more than 1 % of the whole dataset were used in the regression model. 9.6 % of pedestrian crashes were classified as fatal and 90.4 % were classified as an injury.

Table 1 : Independent variables

Variable	**Values***
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 (Mon. to Fri.)	Weekday = 1 , Weekend =0
Season	Winter (Dec-Jan-Feb), Autumn(Sep-Oct-Nov), Summer (Jun-Jul-Aug), Spring (Mar-Apr-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	Truck route, Bus route, Near subway station, Metered parking, On street bicycle

feet)	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**	*less than 10 feet, 10-20, 20-30, 30-42, 42-65, More than 65 feet*

* In clustering analysis, all values were used. In regression, those values marked in italics were excluded.

** The road width variable was excluded from regression because it is correlated with the number of travel lanes.

The primary source of the secondary dataset collected in Montreal from 2003 to 2006 is the Société de l'Assurance Automobile du Québec (SAAQ, Québec's public auto insurance and licensing body). It was used previously in (2) and readers are referred to this publication for more details because of the space constraints. The variables are the following:

- Road type: local, major, highway.
- Accident location at intersection: binary (Yes/No).
- Type of movement: straight, left turn, right turn, reverse.
- Vehicle type: automobile, van/truck/bus, motorcyclist, emergency vehicle.
- Environmental condition, after dark, bad weather.
- Visibility: bad due to weather, bad due to object.
- Built environment variables:
    - Median income: continuous.
    - Population density: continuous.
    - Transit access: continuous.
    - Connectivity: continuous.
    - Mixed use: continuous.
    - School presence: binary.
    - Park presence: binary.
    - Hospital presence: binary.

A total of 5820 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 available in the NYC dataset are not available in the Montreal dataset. Nevertheless, it will be useful for examining the proposed methodology and exploring the shared contributing variables in injury severity.

RESULTS AND DISCUSSION

Case study 1: New York City, US

Latent Class Analysis

The 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 cluster were tested, from one to eleven. The BIC, AIC, and CAIC criteria were used to select the final number of clusters. As shown in Figure 1, BIC decreases until seven clusters, 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 (17). 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, which is quite high. Based on the BIC and CAIC, it is preferred to use seven groups for clustering.

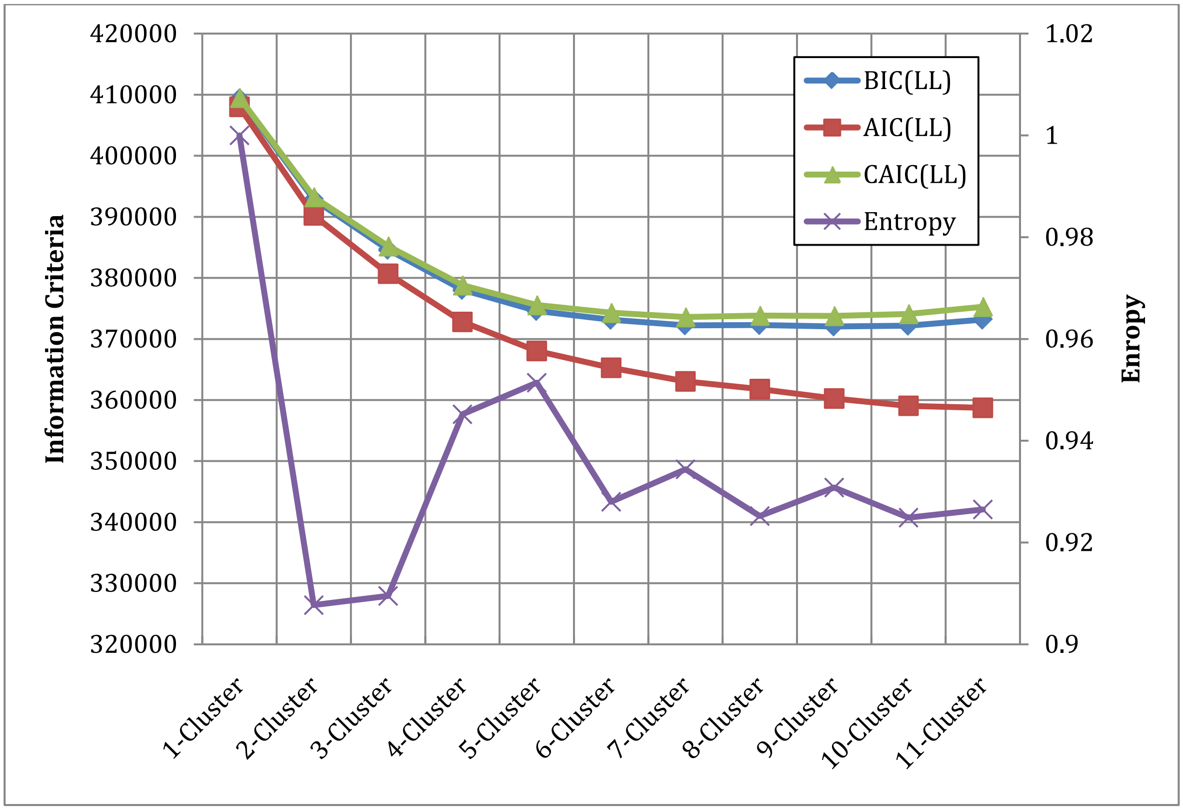

Figure 1 : Variation of BIC, AIC, CAIC and Entropy values for model selection

The final model was described by the proportion of each variable in each cluster. Similarly to (10), the clusters were analyzed and named based on the variable distribution in all the clusters. 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 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 with light. Cluster 3 reveals the missing values in the driver characteristics and vehicle type which presents a type of missing data in many collision reports such as Montreal dataset in our case study. The special features of cluster 4 are the number of travel lanes and the existence of a park 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 (10) and (18).

Table 2: Summary of interesting variables and their distribution in each cluster*

Variables	*Whole*	*C1*	*C2*	*C3*	*C4*	*C5*	*Cr6*	*C7*
	data	*(%)*	*(%)*	*(%)*	*(%)*	*(%)*	*(%)*	*(%)*
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 ft	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

* For the complete results, contact the authors

To categorize cluster 5, three variables are specific to this cluster. 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 multi lanes (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.

Table 3 : Cluster descriptions and accident categories

Cluster no.	*Category*	*Proportion of whole dataset*
Cluster 1	Accidents happening at signalized intersections in daylight	20.6% -1420 cases
Cluster 2	Accidents happening at signalized intersections in dark conditions with light	17.7% - 1223 cases
Cluster 3	Missing driver information	16.8% - 1160 cases
Cluster 4	Accidents involving a car/van/pickup, traveling in one or two lanes with a park lane	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	15.0% - 1037 cases
Cluster 6	Accidents taking place at parking facilities	11.6% - 798 cases
Cluster 7	Multiple missing values	2.7% - 186 cases

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 where too many values are missing. As each cluster describes a specific accident category, the independent variables that characterize this category 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 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 taken into consideration in this study is 10 %.We built the model considering injury crash as the base case. Therefore, a positive coefficient sign means a higher probability of a fatal crash.

Table 4 : Ordered probit model results for whole dataset and each cluster[1]

Table 4-1: Ordered probit model results: Model characteristics

Variables	injury outcome is the base case						
	Whole dataset	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Constant	-2.581	-1.475	-10.892	-3.135			
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

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

Table 4-2: Ordered probit model results: Pedestrian characteristics

Variables	injury outcome is the base case													
	Whole dataset		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val.	Coeff.	P.val.	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val
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												

Table 4-3: Ordered probit model results: Vehicle and driver characteristics

Variables	injury outcome is the base case													
	Whole dataset		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val.	Coeff.	P.val.	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val
Gender														
Male	0.128	0.068	0.249	0.091										
Driver Age														
Under 26					0.296	0.043								
Between 26 to 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

Table 4-4: Ordered probit model results: Environmental condition

Variables	injury outcome is the base case													
	Whole dataset		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val.	Coeff.	P.val.	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val
Weekday (Monday to Friday)					-0.265	0.032								
Season														
Winter (Dec-Jan-Feb)	0.153	0.028							0.495	0.035				
Autumn(Sep-Oct-Nov)	0.202	0.003			0.408	0.017			0.561	0.009				
Summer (Jun-Jul-Aug)					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				
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						

Table 4-5: Ordered probit model results: Built environmental variables

Variables	injury outcome is the base case													
	Whole dataset		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val.	Coeff.	P.val.	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val
Land Use														
1 & 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		

Table 4-6: Ordered probit model results: Network variables

Variables	injury outcome is the base case													
	Whole dataset		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val.	Coeff.	P.val.	Coeff.	P.val	Coeff.	P.val	Coeff.	P.val
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												

General logistic regression analysis

With respect to the pedestrian characteristics, pedestrians aged 40 to 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 sidewalk, 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. The results make sense because we can provide a likely explanation or mechanism that most drivers pay attention and reduce the 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 primary factors in the accidents that are statistically significant in increasing the risk of a fatal crash. Vehicles 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, dawn, dark (lighted or unlighted) increase the probability of 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 types of weather, either clear or bad such as cloudy, rain and snow, had negative signs that mean they reduce the probability of a fatal crash. The reason behind reductions in fatalities under bad weather may be that drivers travel more slowly.

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

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

Cluster-based logistic regression analysis

In this section, we report the results of the injury risk analysis per 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. The results were 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: pedestrian aged under 5, driver age and sex, vehicle movement prior to accident, built environment 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.

Another finding is that the effect of some variables changes direction between certain clusters and the reasons are 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 1) 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 similar finding was done 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 some observations to be validated more closely.

Case study 2: Montreal Canada

Clustering Analysis

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

Injury Severity using MNL

Since there are three categories of injury severity, the MNL model is more appropriate to analyze this dataset. A model of the whole dataset and 5 models for each cluster were examined. No injury crashes were selected as a reference (base) case for the dependent variables. Consequently, the estimated coefficients show the effects of a contributing factor to a fatal or minor injury relative to 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 collision are: accidents at intersection and connectivity. 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 contributing variables in fatal and minor injury for each cluster corresponding to case A and case B. Similar to the NYC dataset, bad visibility has positive effects in increasing the fatality. The effect of hospital presence for reducing fatal crashes is an important finding in this analysis. The influence of mixed use in cluster 3 and after dark variable in cluster 5 for reducing minor injury is unexpected.

Table 5: MNL model estimation for the Montreal dataset[2]

Base case : no injury	Overall data		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	

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

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 to 50 years; straight motion; right turn; and left turn; driver inattention	Pedestrians aged 40 to 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 50ft.	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 to 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 to 15; 15 to 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 to 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 to 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.

Montreal Case Study					
Cluster #	***Impact on probability***	**Fatal**		**Minor Injury**	
		Case A	**Case B**	**Case A**	**Case B**
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

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 different clusters based on some crash characteristics such as traffic control, lighting conditions, vehicle type, land use, type of movement, environmental condition and the missing events. Once the dataset was segmented, specific types of accidents (clusters) were separately analyzed. Although the clustering and parameters explain different features of the models, they in fact complement each other to provide a more detailed analysis.

By clustering the dataset, this work confirms the hypothesis that segmenting the traffic accident dataset into homogeneous subsets helps in the identification of important contributing factors that will be hidden if the whole dataset was used. Thus, it is recommended to use clustering not only for descriptive analysis but also as a preliminary tool for a more detailed analysis using well-known statistical methods.

In terms of exploring the contributing factors in fatal crashes, comprehensive analyses were done using two different case studies. Interestingly, several variables were common and their effect was confirmed in both case studies. Heavy vehicles, dark lighting conditions, mixed land use, and major road increase the probability of fatal crashes. In addition, crossing at intersections lowers the severity. These support the following recommendations:

- Restrict truck flows or movements at intersections with high pedestrian activity, or to require smaller vehicles for local deliveries. This will be a complementary strategy to a policy of reducing overall traffic volume. Retrofit major roads into complete streets or improve road lighting to increase visibility;
- Traffic engineers need to pay attention to land use to improve safety, in particular in an area that has many pedestrian activities such as mixed residential and commercial zones, and public institutions.

Secondly, other contributing variables influencing crash severity were found in the analysis of the NYC dataset. The complete results with the effect on fatal crash were presented in Table 6. The most interesting findings are the following:

- With respect to the pedestrian characteristics,
    - ➢ Older pedestrians are the most prone to fatal injuries in pedestrian-vehicle crashes. The reason may be that their speed and reaction time is relatively low so they are less likely to avoid an accident. A solution could be to increase the signal timing, based on lower walking speeds.
    - ➢ Child pedestrians aged less than 5 are also more likely to be involved in fatal crashes. At this age, most children are accompanied by an adult, who has this responsibility.
    - ➢ Pedestrians crossing in absence of a signal or crosswalk increase the likelihood of fatal crash. This suggests there should be pedestrian signals at most signalized intersections.
- In term of vehicle and driver characteristics,
    - ➢ Disregard of traffic control devices and bad visibility increase the likelihood of fatal accidents. Hence, it is important that both traffic engineers and law enforcement ensure good visibility of traffic devices and enforce their respect. Also, pedestrians' error/confusion is considered as one of the reasons for fatal crashes at signalized intersections in dark lighting conditions with light.
- 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. This sheds light in taking these special features into consideration while designing or retro-fitting roads or intersections. It is important to consider also that some of these factors might have an inverse effect on crash frequency.

For future research, it is recommended to examine different types of built environment characteristics to propose more countermeasures to help policy makers, planners, and traffic engineers to 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 shall also be investigated. This could help learn more about crash injury severity mechanisms.

ACKNOWLEDGMENT

We acknowledge the NYCDOT Pedestrian Safety Project along with CUBRC who assisted in putting together the NYC Pedestrian database. The authors thank also SAAQ for contributing Montreal Dataset and Paul St-Aubin for his help in proofreading the text.

REFERENCES

1. *A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes.* **Eluru N., Bhat C.R., Hensher D.A.** s.l. : Accident Analysis & prevention, 2008. Vol. 40, pp. 1033-1054.
2. *Estimating the Potential Effect of Speed Limits, Built Environment and Other Factors on the Pedestrian and Cyclist Injury Severity Levels in Traffic Crashes.* **Zahabi, S.A., Strauss, J., Miranda-Moreno ,L., Manaugh, K.** s.l. : Transportation Research Board 90th Annual Meeting, 2011.
3. *Severity of injury resulting from pedestrian–vehicle crashes: What can we learn from examining the built environment?* **Clifton, K.J, Burnier, C.V., Akar, G.** s.l. : Transportation Research , 2009. Vol. Part D 14, pp. 425–436.
4. *Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida.* **Lee, C., Abdel-Aty, M.** 4, s.l. : Accident Analysis and Prevention, 2005, Vol. 37, pp. 775-786.
5. *Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes.* **Sze, N.N., Wong, S.C.** 6, s.l. : Accident Analysis and Prevention, 2007, Vol. 39, pp. 1267-1278.
6. *A Multinomial Logit Model of Pedestrian-Vehicle Crash Severity",, 5: 4, 233- 249.* **Tay, R., Choi, J., Kattan, L. and Khan, A.** 4, s.l. : International Journal of Sustainable Transportation, 2011, Vol. 5, pp. 233-249.
7. *Analysis of traffic injury severity: An application of non-parametric classification tree techniques* . **Chang, L. Y. , Wang, H. W.** s.l. : Accident Analysis and Prevention, 2006, Vol. 38, pp. 1019–1027.
8. *Exploring the potential of data mining techniques for the analysis of accident patterns.* **Prato, C. G., Bekhor, S., Galtzur, A., Mahalel, D., Prashker, J.N.** Lisbon, Portugal : 12th WCTR, 2010.
9. *Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawai.* **Kim, K., Yamashita, E. Y.** 1, s.l. : Journal of advanced transportation, 2007, Vol. 41, pp. 69-89.
10. *Traffic accident segmentation by means of latent class clustering.* **Depaire, B., Wets, G., Vanhoof, K.** s.l. : Accident Analysis and Prevention, 2008, Vol. 40, pp. 1257–1266.
11. *Combining non-parametric models with logistic regression: an application to motor vehicle injury data.* **Kuhnert, P. M. and Do, K-A, McClure, R.** s.l. : Computational Statistics & Data Analysis, 2000, Vol. 34, pp. 371-386.
12. **P., Berkhin.***Survey of Clustering Data Mining Techniques.* s.l. : Accrue Software Inc., 2002.
13. *Survey of Clustering Algorithms.* **Xu, R.** 3, s.l. : IEEE Transactions on Neural Network, 2005, Vol. 16.
14. *Models for Ordered Outcomes.* **S., Jackman.** s.l. : Political Science 200C, 2000.
15. **Borooah, V. K.***Logit and Probit: Ordered and Multinomial Models.* Thousand Oaks, CA : Sage Publication, 2002.

16. **Washington, S., P., Karlaftis, M. G., Mannering F. L.***Statistical and econometric methods for transportation data analysis.* Boca Raton, FL : Chapman & Hall/CRC , Taylor & Francis group, 2011.

17. *Latent class cluster analysis.* **Vermunt J.K., Magidson, J.** s.l. : Applied latent class analysis, Cambridge: Cambridge University Press, 2002, pp. 89-106.

18. *Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong.* **Yau, K.K.W.** 3, s.l. : Accident Analysis and Prevention, 2004, Vol. 36, pp. 333–340.