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