Accident Analysis and Prevention 85 (2015) 219–228

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland

Lekshmi Sasidharan[a,*], Kun-Feng Wu[b], Monica Menendez[a]

[a] Swiss Federal Institute of Technology, ETH Zurich, Stefano-Franscini-Platz 5, CH-8093 Zurich, Switzerland
[b] Department of Civil Engineering, National Chiao Tung University, Taiwan

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form
24 September 2015
Accepted 24 September 2015

Keywords:
Latent class
Cluster analysis
Pedestrian
Severity
Binary logit
Receiver operating characteristic (ROC) curve
Switzerland

ABSTRACT

One of the major challenges in traffic safety analyses is the heterogeneous nature of safety data, due to the sundry factors involved in it. This heterogeneity often leads to difficulties in interpreting results and conclusions due to unrevealed relationships. Understanding the underlying relationship between injury severities and influential factors is critical for the selection of appropriate safety countermeasures. A method commonly employed to address systematic heterogeneity is to focus on any subgroup of data based on the research purpose. However, this need not ensure homogeneity in the data. In this paper, latent class cluster analysis is applied to identify homogenous subgroups for a specific crash type-pedestrian crashes. The manuscript employs data from police reported pedestrian (2009–2012) crashes in Switzerland. The analyses demonstrate that dividing pedestrian severity data into seven clusters helps in reducing the systematic heterogeneity of the data and to understand the hidden relationships between crash severity levels and socio-demographic, environmental, vehicle, temporal, traffic factors, and main reason for the crash. The pedestrian crash injury severity models were developed for the whole data and individual clusters, and were compared using receiver operating characteristics curve, for which results favored clustering. Overall, the study suggests that latent class clustered regression approach is suitable for reducing heterogeneity and revealing important hidden relationships in traffic safety analyses.

1. Introduction

Identification of possible risk factors and implementation of appropriate countermeasures to reduce the risk of crashes is one of the most conventional and practically adopted traffic safety improvement strategies. Researchers depend on traffic safety data analyses to identify the alleviating and aggravating factors influencing the frequency and severity of crashes. Nevertheless, these crashes might have occurred under different conditions, which make the traffic safety data highly heterogeneous in nature, thereby making it difficult to identify some hidden relationships. These relationships may include different effects of the same factor under different conditions. As such, researchers focus on narrow crash variables – crashes due to specific movements (left turn or right turn movements at signalized intersections; Wang and Abdel-Aty, 2008), crashes associated with specific vehicle type (sports utility vehicle, pickups, minivan and passenger cars; Ulfarsson and Mannering, 2004), motorcycles (Shankar and Mannering, 1996; Quddus et al., 2002), crashes involving particular age group of people or gender (Zhang et al., 2000; Ulfarsson and Mannering, 2004; Sasidharan and Menendez, 2014), specific accident type (Islam and Mannering, 2006; Ulfarsson and Mannering, 2004; Savolainen and Mannering, 2007) and so on. Although this approach helps to reduce heterogeneity, it does not guarantee homogenous group of crashes in the dataset. Some previous safety studies (Yau, 2004; Depaire et al., 2008; de Oña et al., 2013) suggest that data mining techniques such as cluster analysis aids in reducing the heterogeneity in the data. Unlike these studies in which the main focus was on vehicle–vehicle crashes and information on different crash types (rear end, angle, head-on, side-swipe, etc.) were available and used in segmentation, the current study focuses on a specific crash type–pedestrian crashes. Another approach that takes into account of the unobserved heterogeneity in analyzing pedestrian injury severities in pedestrian–vehicle crashes includes using a mixed logit model (Kim et al., 2010). Previous pedestrian safety studies have utilized the narratives obtained from police accident report to classify pedestrian crashes into different groups such mid-block dart crashes, crashes due to pedestrian error, turning vehicle crashes, and crashes involving driver's failure to grant right of way for pedestrians (e.g. Fontaine and Gourlet, 1997; Preusser et al., 2002). These studies suggested that pedestrian crash typologies are

* Corresponding author.
E-mail addresses: lekshmi.sasidharan@ivt.baug.ethz.ch, lechu18181@yahoo.com (L. Sasidharan).

http://dx.doi.org/10.1016/j.aap.2015.09.020
0001-4575/

associated with crash severities and using the whole data involving all crashes without distinguishing the typology makes the data heterogeneous. One of the implications is that safety countermeasures needed for different pedestrian crash types are different. Accurate estimation of the effects of different factors influencing pedestrian injury severities under different conditions is vital as traffic engineers, policy makers and planners rely on this information for identifying appropriate safety countermeasures which includes geometric improvement, traffic control measures, dedicated pedestrian facilities, modifying land use, educational and enforcement actions. Therefore, an attempt is made to accommodate the systematic heterogeneity in pedestrian crashes in Switzerland using a latent class clustering approach. Subsequently, a binary logit model is used for each of the identified latent clusters to identify the effect of different crash contributing factors.

1.1. The heterogeneity in crash data

As discussed above, heterogeneity in the crash data is unavoidable and highly undesirable. The three main problems of heterogeneous data include: (1) certain crash contributing factors will remain hidden, e.g. a crash contributing factor that is highly influential for crashes involving specific vehicle types may not be significant in the whole data analysis (Valent et al., 2002; Yau, 2004; Depaire et al., 2008); (2) the magnitude of the effect of certain crash contributing factors may be different for different conditions, e.g. different effects of injury severities for males and females in different age groups (Ulfarsson and Mannering, 2004; Islam and Mannering, 2006); and (3) the increase or decrease in severity levels for a crash contributing factor may be different for different crash types, e.g. an increase in the probability of less injury accidents for male drivers and increase in probability of severe injury accidents for female drivers for crashes involving guardrails (Ulfarsson and Mannering, 2004).

One way to account for the heterogeneous nature of the data is to divide the data into different homogenous subgroups based on exogenous variables (crash location, crash type, speed limit, roadway geometry, traffic conditions, cause of accidents and so on) and analyze each of the subgroups separately to identify the effect of influential factors for that subgroup. However, analysis involving 'dividing the dataset based on all the exogenous variables' is unrealistic as the number of subgroups can be very large and the sample size in some of the subgroups can be very low, thereby restricting the application of severity models. For example, considering all possible subgroups in a study with 8 binary variables will results in $2^8 = 256$ subgroups, which definitely is not practically feasible to analyze and interpret separately. Previous studies show that researchers generally try to divide the data into subgroups based on the objective of research, methodologies used, or on expert domain knowledge (Ulfarsson and Mannering, 2004; Islam and Mannering, 2006; Savolainen and Mannering, 2007). However, some studies suggests that even though the above said factors can result in a workable segmentation of the crash data, one cannot guarantee that each of the subgroups comprise of homogeneous group of crashes (Depaire et al., 2008; de Oña et al., 2013).

1.2. The latent class clustering analysis

To address the heterogeneity issue, a data mining technique, such as cluster analysis can be used to aid in the crash segmentation process (Depaire et al., 2008; de Oña et al., 2013; Kim and Yamashita, 2007; Mohamed et al., 2013). Cluster analysis is a descriptive data mining technique, which can divide a heterogeneous data set into homogenous subgroups or clusters (Berry and Linoff, 1997). It is an unsupervised learning technique that divides data into subgroups or clusters with the goal to maximize both the homogeneity of elements within the cluster and heterogeneity between clusters (Hair et al., 1998). Traffic safety researchers have used different types of cluster analysis in the past to meet different research objectives. Some have used a partitioning method called k-means clustering, a distance based clustering algorithm to identify homogenous crash clusters (Kim and Yamashita, 2007; Mohamed et al., 2013; Hamzehei et al., 2014). Another approach used is hierarchical clustering, which uses methods like Ward's linkage, Single linkage, average linkage, median linkage, and centroid linkage (Depaire et al., 2008). However, the statistical properties of these methods are mostly unknown (Fraley and Raftery, 1998; Vermunt and Magidson, 2002; Depaire et al., 2008). These methods also require the researcher to specify the number of clusters in advance and assign individual observations to one cluster or the other. To avoid these problems, a probability-model based clustering technique known as latent class clustering analysis (LCA) is proposed to identify homogenous subgroups.

LCA posits that there exists an unobserved or latent categorical variable that divides the data into mutually exclusive and exhaustive latent classes (Collins and Lanza, 2010; Goodman, 1974; Lanza and Rhoades, 2013). Even though the class memberships of individual crashes are unknown, it can be inferred from the observed variables. In LCA, the probabilities of each crash to be in different clusters are estimated based on different models developed for different values of clusters specified. LCA do not put one crash into any cluster based on any one property, instead it assigns probability of that crash to be in different clusters and assigns a best index cluster for the cluster with the highest probability of accommodating that crash.

1.3. Pedestrian safety in Switzerland

Pedestrian safety is very important in Switzerland because of the large share of walking commuters and the high injury severity levels associated with crashes involving this group. In Switzerland, people chose to walk for more than 40% of their total trip time as part of their daily routine (Microcensus, 2010), which points to the importance of ensuring the safety of pedestrians in this country. In addition, Switzerland has a very unique unexplored crash database. The year 2010 in Switzerland witnessed 2418 pedestrian crashes of which 69 were fatal crashes (Swiss council for accident prevention, 2013). A pedestrian safety study in Switzerland showed that the factors influencing the injury severity levels in old and young pedestrian crashes in Switzerland and its effect are very different when compared to all pedestrian crashes, which clearly points to the need to segment the pedestrian data before analysis (Sasidharan and Menendez, 2014).

Previous studies have identified that safety countermeasures needed for different pedestrian crash types are different. A study conducted in France suggests that a typology based on pedestrian crashes can assist in an in-depth analysis to identify influential factors and determined four groups based on a correspondence analysis for pedestrian fatal injury crashes – elderly pedestrians who were crossing a road in an urban area; children involved in daytime crashes in urban areas while playing or running; pedestrians under influence in nighttime crashes while walking on road; and pedestrians involved in secondary crashes (Fontaine and Gourlet, 1997). Similarly, another study (Preusser et al., 2002) based on pedestrian crashes in two urban areas in the United States identified crash groups for mid-block dart crashes, crashes due to pedestrian error, turning vehicle crashes, and crashes involving driver's failure to grant right of way for pedestrians. Although there is a need to distinguish different types of pedestrian crashes for more accurate crash data analysis, such information is often not directly documented and can only be assembled using the narratives in the