Analyzing Motorcycle Crash Pattern and Riders' Fault Status at a National Level: A Case Study from Iran

Ali Tavakoli Kashani[1], Mohammad Mehdi Besharati[2], Ahmad Mohamadian[3]

Received: 16. 11. 2016 *Accepted: 03. 05. 2017*

Abstract

Motorcycle crashes constitute a significant proportion of traffic accidents all over the world. The aim of this paper was to examine the motorcycle crash patterns and rider fault status across the provinces of Iran. For this purpose, 6638 motorcycle crashes occurred in Iran through 2009-2012 were used as the analysis data and a two-step clustering approach was adopted as the analysis framework. Firstly, hierarchical clustering (HC) was applied to group the provinces into homogenous clusters, based on the distribution of crash characteristics in each province. In the second step, the latent class clustering (LCC) was employed to investigate the crash patterns and rider fault status among the provinces. The provincial groupings were found to be an influential factor in the final crash clusters implying the effectiveness of the proposed framework. Results of LCC also indicated that Cluster 8 with the highest percentages of not wearing helmet, unlicensed and under 21 years old riders, had the highest percentage of fatal crashes. In addition, the motorcyclists seemed to be less responsible in the pedestrian-motorcycle crashes. Accordingly, training programs for the riders in the license issuance process about the risk of pedestrian-motorcycle crashes could help mitigate this type of crashes. Generally, analyzing the culpability in pedestrian-motorcycle crashes might be a good topic for future research. Further discussions on the crash patterns are provided. Finally, the combined use of HC and LCC should not be regarded as an alternative to the other more qualitative predictive methods, but as a preliminary analysis tool to provide insights over the road safety condition at the national level.

Keywords: Hierarchical clustering, latent class clustering, motorcycle crashes, motorcyclists' fault status

Corresponding author E-mail: alitavakoli@iust.ac.ir

[1] Assistant Professor, School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran, alitavakoli@iust.ac.ir

[2] Ph.D. Candidate, School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran,

[3] M.Sc. in Transportation Engineering, School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran,

1. Introduction

Motorcyclists are among the most vulnerable road user groups, overrepresented in road traffic crashes, especially in the developing countries [Haque, Chin, and Huang, 2010; Tavakoli Kashani, Rabieyan, and Besharati, 2016; Vlahogianni, Yannis, and Golias, 2012]. This is the same in Iran, where motorcycle is a popular transportation mode and is, unfortunately, involved in a significant proportion of fatal crashes [Tavakoli Kashani, Rabieyan, and Besharati, 2014]. According to the Iran Forensic Medicine Organization report, motorcyclists comprise 25.7% of the total traffic crash fatalities occurred during 2006 to 2010. On the other hand, previous studies have showed that at-fault motorcyclists are more likely to be involved in or injured due to a traffic crash. For example, Savolainen and Mannering (2007) showed that motorcyclists who are at-fault in crashes are also more likely to die in the event of a crash. In addition, results of previous researches suggest that at-fault and not-at-fault motorcycle crashes might have different causes [Haque, Chin, and Huang, 2009; Savolainen and Mannering, 2007]. This has inclined the researchers to study the various factors that might be associated with motorcyclists' responsibility in crash causation [Haque et al., 2009; Kim and Boski, 2001; Kim and Li, 1996].

In this regard, in a study on two-vehicle motorcycle crashes, Schneider, Savolainen, Van Boxel, and Beverley (2012), examined such factors as alcohol consumption, rider age, helmet use, car insurance as well as driving experience and found them effective on the at-fault and not-at-fault crashes. In terms of rider's age, previous researches generally suggest that younger riders have a stronger propensity of risky behavior [Harrison and Christie, 2005; Rutter and Quine, 1996; Schneider et al., 2012].

Zhang, Yau, and Zhang (2014), studied motorcycle-pedestrians, and concluded that motorcyclists are more likely to be at-fault in these crashes.

In another study, Haque et al. (2009) found that motorcyclists were more likely to be at-fault when the crash occurred on an expressway or where the speed limits were higher, when the motorcycle had a larger engine size, under wet pavement conditions, and in collision with pedestrians. They also showed that young and older riders as well as those carrying pillion passengers were more likely to be at-fault in crashes.

In terms of analytical methods, previous studies have employed a variety of methodological approaches including regression models to investigate the factors contributing to traffic crash frequencies and the injury severity of different road users [Besharati and Kashani, 2017; Jalayer and Zhou, 2017; Mannering and Bhat, 2014; Ali Tavakoli Kashani and Mohammad Mehdi Besharati, 2016]. More recently, several data mining techniques have been used by researchers in the road safety domain [Chang and Chen, 2005; Kumar and Toshniwal, 2016; Tavakoli Kashani et al., 2014].

On the other hand, a number of previous studies have tried to employ exploratory data analysis techniques such as multiple correspondence analysis [Jalayer and Zhou, 2017] in order to identify patterns in large crash datasets and provide a general overview on the key factors affecting crash occurrence and injury severity. Also, clustering analysis is a popular descriptive data mining technique that has been widely used in recent years in the road safety domain [Depaire, Wets, and Vanhoof, 2008; Kaplan and Prato, 2013; Ali Tavakoli Kashani and Mohammad Mehdi Besharati, 2016]. Results of previous studies have showed that applying clustering techniques as a preliminary analysis tool can reveal hidden relationships and help the traffic safety researcher to group traffic crashes into more homogenous clusters [Depaire et al., 2008; Kaplan and Prato, 2013].

Although previous studies have explored the association of several contributory factors with the motorcyclists' fault status, a review of the literature reveals that very few researches have been conducted to investigate the pattern of culpable riders in several types of motorcycle crashes at a national level. Therefore, the current study employed a data mining framework as an exploratory data analysis tool to conduct a preliminary analysis over the motorcycle crash patterns and motorcyclists' fault status at a national level.

It seems that addressing the issue of motorcyclists' fault status can help provide a better understanding of the patterns and causes of motorcycle crashes, which in turn, might increase the effectiveness of preventive measures and improve the motorcycle safety [Haque et al. 2009].

2. Methodology

2.1 Crash Data

In this study, Iran crash data recorded by the Information and Technology Department of the Iranian Traffic Police, from 2009 to 2012 were used. Since the scope of the present study was to identify the factors influencing riders' fault in motorcycle crashes, Motorcycle crash data were extracted from the main database which contains all different sort of crashes. After cleaning the data, finally 6638 data records were identified for analysis.

These data are obtained from the Traffic Crash Record form, KAM 114, which contains important information about the crashes. The information covers different aspects of a traffic crash such as cause of crash, collision type, vehicle type, location type, lighting condition, weather condition, Road surface condition, shoulder type, and characteristics of the riders involved such as age, gender, helmet status, type of drivers' license and its issuance date, and so on.

After cleaning the data, finally 6638 data records were prepared for analysis. Ten variables were considered in the current study. Table 1 presents the study variables and subcategories of each variable.

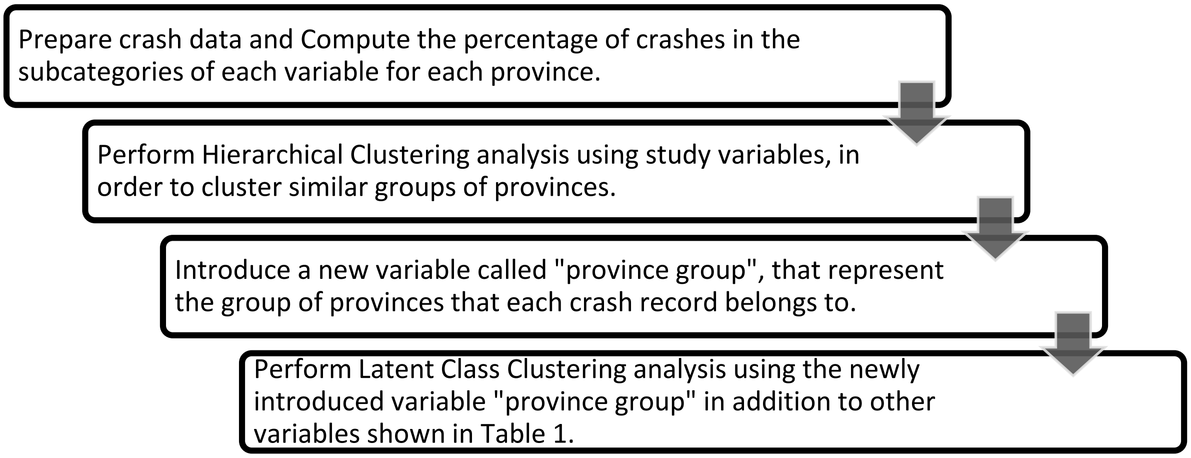
2.2 Analysis Procedure

Since the provinces under study have significantly unequal crash frequencies and fatalities (due to difference in population), it was not possible to simply group the provinces according to raw frequencies of crashes in each province. For example in Iran, from 2009 to 2012 more than 50% of the crashes have taken place in the Tehran province. Using an unsupervised clustering method such as latent cluster analysis, without pre-classifying the provinces with the hierarchical methods, would lead to an unbalanced clustering, in which, for example, half of the clusters would only have incorporated the crashes occurred in Tehran province. In such case, examining the crash patterns among the provinces was not possible, because of the unbalanced distribution of the data across the provinces.

The analysis framework adopted in this study is represented in Figure 1. As shown in this figure, first the percentage of crashes in the subcategories of each variable in each province were calculated. The provinces were then clustered into homogeneous groups using hierarchical clustering analysis and a new variable called "province group" was introduced and added to the crash database. This variable represent the group of provinces that each data record belonged to. In the next step, latent class clustering was performed using the newly introduced variable as well as other variables in Table 1, aiming to group motorcycle crashes into homogeneous clusters and explore hidden patterns of the motorcyclists' injury severity among the clusters.

Variable	Category	Freq.	Variable	Category	Freq.
Area type	Rural	448	Riding license	Unlicensed	463
	Urban	6190		Licensed	6175
Job type	Self-employment	5323	License issuance date	1958/06/28– 2005/09/01	465
	Unemployed	439		2005/09/04-2007/05/31	1208
	Student	84		2007/06/03-2008/11/02	1540
	University student	24		2008/11/03-2009/12/01	1105
	Labor	241		2009/12/02-2010/11/10	1190
	Officeholder	285		2010/11/11-2012/06/20	1130
	Military	195	Age group	15-21	1221
	Driver	47		22-23	954
Education	Illiterate	162		24-26	1253
	Primary school	349		27-30	968
	Guidance school	1244		31-39	1143
	High school	171		40-89	1099
	Diploma	4471	Collision type	Pedestrian-motorcycle	989
	Associate degree	137		Car-motorcycle	5462
	Bachelor's degree	104		Fixed object collision	68
Terrain type	Level	6573		overturn	119
	Rolling	29	Helmet usage	Used	2365
	Mountainous	36		Not-used	4273
Rider fault status	Not at fault	3856			
	At fault	2782	**Total**		**6638**

Table 1. Variable description


Prepare crash data and Compute the percentage of crashes in the subcategories of each variable for each province.

Perform Hierarchical Clustering analysis using study variables, in order to cluster similar groups of provinces.

Introduce a new variable called "province group", that represent the group of provinces that each crash record belongs to.

Perform Latent Class Clustering analysis using the newly introduced variable "province group" in addition to other variables shown in Table 1.

Figure 1. Analytical framework of this study

2.3 Cluster Analysis

Cluster analysis (CA) is an unsupervised learning technique of descriptive data mining, that is employed to group the data into clusters in such a manner that maximize both the homogeneity within each cluster and the dissimilarity among different clusters. This is done without the benefit of prior knowledge about the groups and their characteristics, and it distinguishes clustering models from the other modeling techniques in that there is no predefined output or target field for the model to predict. For this reason, there are no right or wrong answers for these models, and their merit is determined by their ability to capture interesting groupings in the data and provide useful descriptions of those groupings.

Generally, clustering approaches include 1) Probability-based; and 2) Distance-based approaches (such as *Partitioning Clustering* (e.g., K-means, k-mediods), and Hierarchical Clustering (e.g., Ward's method)) [Jain, Murty, and Flynn, 1999].

2.3.1 Hierarchical Clustering

Hierarchical Clustering has two types of strategies, divisive and cumulative. Divisive methods are "top down" approaches in which, all records start in one cluster, and splits are performed recursively as one moves down the hierarchy. Cumulative methods are "bottom up" approaches in which, initially each data is assigned to its own cluster and then the groups that are close to each other are combined so that finally all the groups combine into a single group [Jain et al. 1999]. Cumulative methods include: single linkage, average linkage, complete linkage and Ward linkage methods. In the current study, Ward linkage algorithm which has widely been used in the similar studies [O'brien, Cheshire, and Batty, 2014], was employed to group the provinces.

Ward's method states that the distance between two clusters, A and B, is how much the sum of squares will increase when we merge them:

$$\Delta(\mathrm{A},\ \mathrm{B}) = \sum_{i \in A \cup B} \|\mathrm{xi} - \vec{m}\mathrm{A} \cup \mathrm{B}\|^2 - \sum_{i \in A} \|\mathrm{xi} - \vec{m}\mathrm{A}\|^2 - \sum_{i \in B} \|\mathrm{xi} - \vec{m}\mathrm{B}\|^2 \qquad (1)$$

$$= \frac{\mathrm{n}_A \mathrm{n}_B}{\mathrm{n}_A + \mathrm{n}_B} \|\vec{m}\mathrm{A} - \vec{m}\mathrm{B}\|^2 \qquad (2)$$

Where $\vec{m}$j is the center of cluster j, n_j is the number of points in it, and Δ is called the merging cost of combining the clusters A and B.

With hierarchical clustering, the sum of squares starts out at zero (because every point is in its own cluster) and then grows as we merge clusters. Ward's method keeps this growth as small as possible [Ward Jr, 1963].

As described in Figure 1, the output of the hierarchical clustering was added to the database as a new variable called "province group", which showed the group of provinces that each data record belonged to. The newly introduced variable as well as the other study variables were then imported to the Latent Class Clustering analysis.

2.3.2 Latent Class Clustering

Latent class clustering is one of the unsupervised clustering methods which was conceived more than 4 decades ago. However, its application was limited until the last decade, when renewed interest and advances in computational capabilities led to its widespread application in a variety of social science studies [Lanza, Collins, Lemmon, and Schafer, 2007; Vermunt and Magidson, 2002].

The main advantages of LCC over alternative clustering algorithms (such as k-means clustering, 2-stage clustering, Kohonen networks) is the ability to represent overlap across clusters rather than only independent or nested clusters, the existence of an underlying statistical model that allows calculating cluster probabilities for new cases, and the provision of several goodness-of-fit criteria that facilitate the decision regarding the number of clusters [Depaire et al., 2008; Vermunt and Magidson, 2002].

LCC is defined as the classification of similar objects into C latent classes, where uncertainty is involved in the class membership and the number of clusters and their size is unknown Assume a vector of N observations characterized by a vector of M variables ($y_i = y_1, \ldots, y_M$), and let Y_i ($Y_i = Y_{i1}, \ldots, Y_{iM}$) be the vector of values of observation i for the M items. Then, the latent class model is formulated as follows (Kaplan and Prato, 2013),

$$P(Y_i \mid \theta) = \sum_{k=1}^{K} P(C_k).P(Y_i \mid C_k, \theta_k) \quad (3)$$

where k (k = 1, . . . , K) indicates a latent class, K is the number of latent classes, $P(C_k)$ denotes the prevalence of latent class C_k in the data set, $p(Y_i \mid C_k, \theta_k)$ denotes the conditional multivariate probability that an observation in class C_k would be characterized by Y_i, and θ_k is a vector of variables to be estimated. The model formulation is very flexible in not implying any assumption regarding the nature of the variables (i.e., discrete or continuous), their underlying distributions, and the correlation patterns across observations and variables the mixture probability density for the whole data set can be expressed as

$$P\ (Y|\theta) = \sum_{z=1}^{k}[P(C_z) \prod_{j=1}^{m} p(y_i|C_z, \theta_z)] \quad (4)$$

After estimation of the variable vector θ, the underlying statistical model assigns a set of posterior probabilities p_{ik} which indicate the probability of belonging to cluster C_k for each data element Y_i. In this regard, LCC resembles fuzzy clustering [Höppner, 1999]. One of the common problems in estimating models by means of clustering analysis, is that the model may converge to a local maximum. To avoid this problem, modeling was performed with 50 random starting points. Finally Convergence was achieved when the maximum absolute deviation was less than 1E-06. Since previous studies indicated the superiority of the BIC compared to the other information criteria in terms of consistency and accuracy (Nylund, Asparouhov, and Muthén, 2007), the BIC criterion was used in the current study to determine the number of clusters. In addition, the entropy criterion was used as another measure to find the best clustering model. The entropy criterion take values between 0 and 1, where 1 indicates the highest certainty in the classification and 0 indicates the worst quality in the clustering [Depaire et al., 2008]. Furthermore, the R^2, which indicates how well an indicator is explained by the model, was calculated to identify significant variables [Vermunt and Magidson, 2005].

In the first step, the study variables were used to cluster the provinces. Due to egregious differences in the population of provinces, it was not possible to use the raw frequencies of crashes in the subcategories of variable for clustering the provinces. Therefore, the *percentage* of each subcategory in each variable was used to group the provinces. For example, as shown in Table 2, for the "Helmet usage" variable, with 2 subcategories of "used" and "not used", the percentages of each subcategory in each province was used in hierarchical clustering process. Subcategories of the other variables described in Table 1 were also included in the HC analysis in order to group the provinces. Provinces were finally clustered based on the distribution of study variables in crashes of each province. Figure 2, shows the dendrogram of the provinces as a result of clustering by Ward algorithm. As shown in this figure, the provinces were divided into 5 groups based on the results of the HC. These clusters are;

- **Group A**: Ardebil, Hormozgan, Tehran, Bushehr, Mazandaran, Qazvin
- **Group B**: Chaharmahal and Bakhtiari, Khorasan-Razavi, Gilan, Semnan, Lorestan, Golestan, Fars, Great Tehran, Kordestan, Hamedan
- **Group C**: Ilam, Kohkilouye and boyerahmad, West Azerbaijan, Khuzestan, Kermanshah, Sistan and Balouchestan
- **Group D**: Zanjan, Yazd, Qom
- **Group E**: North Khorasan, South Khorasan, Markazi, Isfahan, Kerman, East Azerbaijan

After clustering the provinces into homogeneous groups, the variable of "province group", was added to the database to represent the group of provinces that each record belonged to. Next, the variable of "province group" as well as other variables were inputted into the LCC analysis. Table 3 shows the corresponding R^2 for each of the variables used in LCC.

Table 2. Univariate distribution of Helmet usage across the provinces

Province	Helmet usage (%)	
	Used	Not used
Ardebil	38.9	61.1
West Azerbaijan	34.6	65.4
East Azerbaijan	25.7	74.3
Bushehr	33.3	66.7
ChaharmahalandBakhtiari	25.0	75.0
Isfahan	27.7	72.3
Fars	26.5	73.5
Qazvin	27.0	73.0
Qom	18.3	81.7
Gilan	31.2	68.8
Golestan	16.0	84.0
Hamedan	47.1	52.9
Hormozgan	54.6	45.5
Ilam	10.5	89.5
Kerman	15.8	84.3
Kermanshah	22.6	77.4
Khuzestan	24.7	75.4
South-Khorasan	29.4	70.6
Khorasan-Razavi	20.0	80.0
North-Khorasan	21.6	78.4
KohkilouyeandBoyerahmad	16.0	84.0
Kordestan	50.8	49.2
Lorestan	23.9	76.1
Markazi	17.8	82.2
Mazandaran	36.1	63.9
Semnan	22.5	77.5
SistanandBalouchestan	18.1	81.9
Tehran	60.5	39.5
Great Tehran	48.9	51.1
Yazd	11.7	88.3
Zanjan	37.9	62.1

3. Results

The BIC values for 2 to 14 cluster models are presented in Figure 4. As there is an apparent elbow in the BIC curve at 8 number of clusters, the 8-cluster solution was selected as the best model. Moreover, the entropy criterion of the model was equal to 0.74, indicating a reasonably high certainty in the clustering (see, e.g., (Depaire et al., 2008; Kaplan and Prato, 2013) for more details about acceptable ranges of entropy).

The next step was to characterize the clusters based on the proportion of each subcategory of variables in each cluster. The variables that were selected to characterize the clusters are shown in Table 4. This table also shows the proportion of each subcategory of variables in each one of the 8 clusters. Note that only significant subcategories of each variable are presented in this table.

Similar to previous works (de Oña, López, Mujalli, and Calvo, 2013; Depaire et al., 2008; Mohamed, Saunier, Miranda-Moreno, and Ukkusuri, 2013), the clusters were analyzed and named based on their variable distributions. Since, only the differences between the clusters where important, the subcategories that were dominant in all clusters were not included in the naming process. For example, “diploma” level of education was dominant in all the clusters, and therefore was not entered in the naming process.

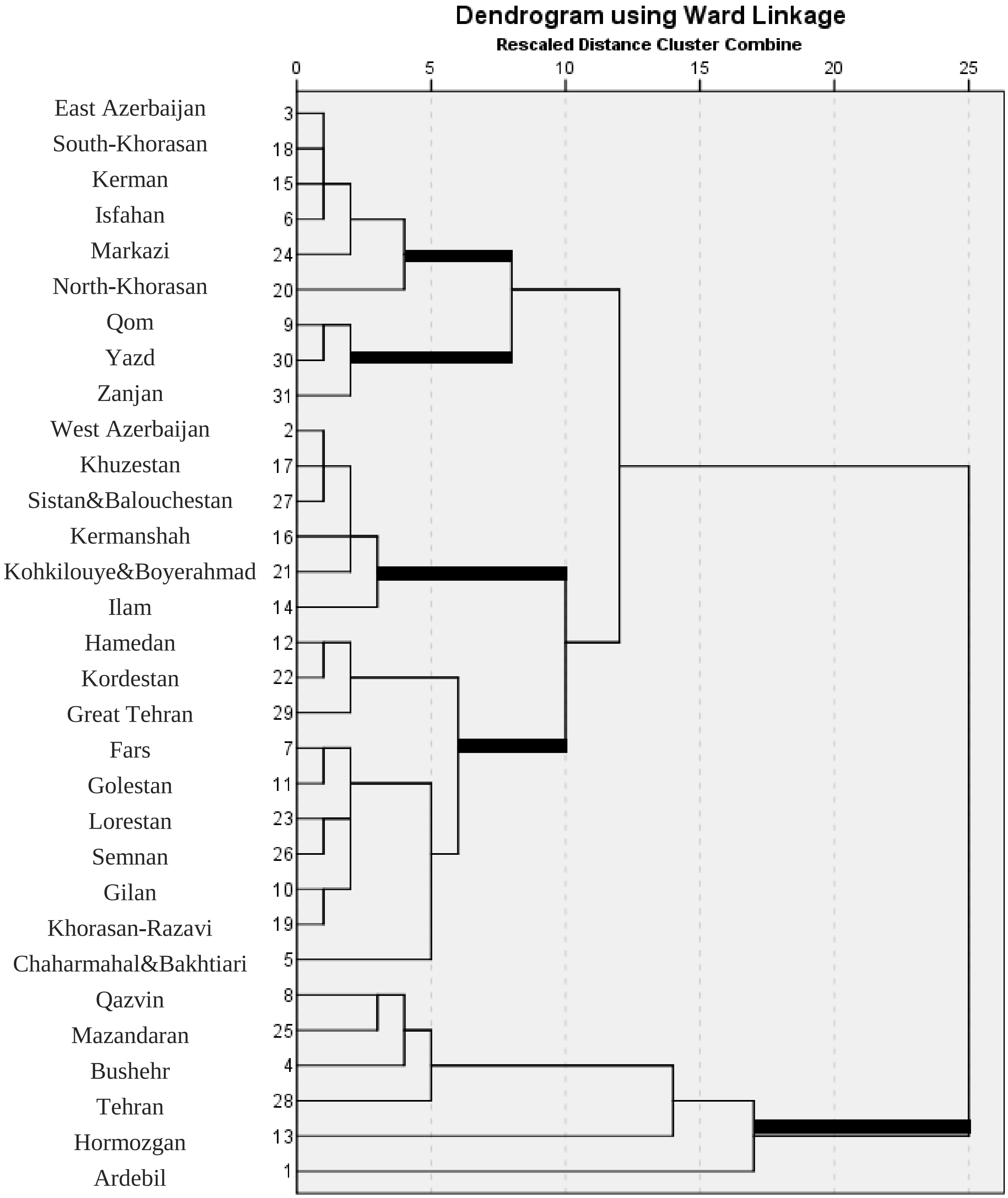


Figure 2. Hierarchical clustering output by Ward method

Table 3. R-squared of variables used in the LCC analysis

Variables	R^2
Province group	0.07
Helmet usage	0.06
License issuance date	0.12
Riding license	0.71
Rider age group	0.35
Job type	0.11
Education	0.10
Area type	0.14
Collision type	0.37
Crash severity	0.33

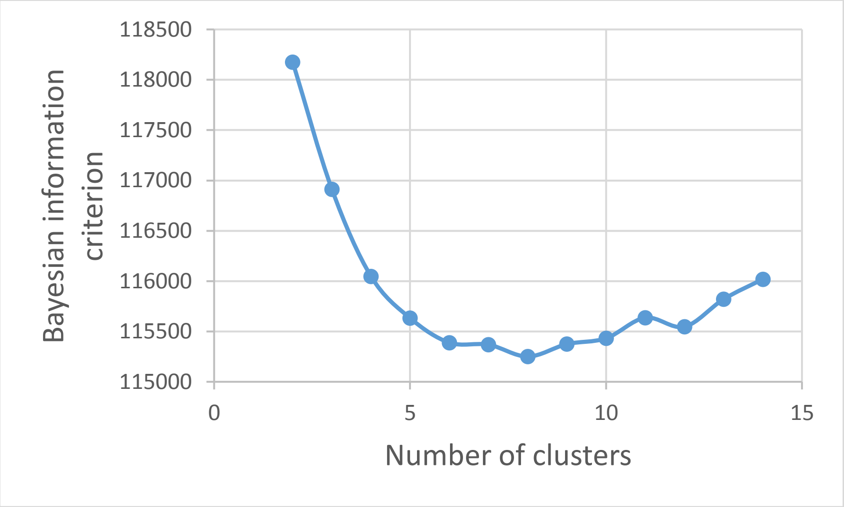


Figure 4. BIC values for several number of clusters

Table 4. Variables distribution in the 8 Latent Class Clusters

Cluster number	1	2	3	4	5	6	7	8
Cluster Size	0.37	0.19	0.14	0.09	0.08	0.06	0.04	0.04
Province group								
A	0.03	0.06	0.05	0.35	0.02	0.01	0.01	0.02
B	0.70	0.56	0.63	0.41	0.46	0.74	0.17	0.23
C	0.07	0.05	0.07	0.12	0.11	0.05	0.39	0.40
D	0.03	0.09	0.08	0.02	0.10	0.04	0.02	0.03
E	0.17	0.24	0.17	0.11	0.31	0.16	0.40	0.31
Helmet usage								
used	0.42	0.27	0.39	0.51	0.26	0.42	0.05	0.06

Not used	0.58	0.73	0.61	0.49	0.74	0.58	0.95	0.94
License issuance date								
1958/06/28– 2005/09/01	0.08	0.00	0.08	0.06	0.12	0.65	0.99	0.93
2005/09/04-2007/05/31	0.24	0.00	0.26	0.19	0.25	0.04	0.00	0.00
2007/06/03-2008/11/02	0.21	0.14	0.19	0.20	0.16	0.06	0.00	0.01
2008/11/03-2009/12/01	0.17	0.25	0.16	0.20	0.15	0.05	0.00	0.03
2009/12/02-2010/11/10	0.13	0.30	0.17	0.18	0.16	0.15	0.00	0.01
2010/11/11-2012/06/20	0.16	0.31	0.14	0.16	0.16	0.05	0.01	0.02
Riding license								
Unlicensed	0.01	0.01	0.00	0.04	0.03	0.08	0.91	0.99
Licensed	0.99	0.99	1.00	0.96	0.97	0.92	0.09	0.01
Rider age group								
15-21	0.03	0.72	0.03	0.03	0.01	0.01	0.02	0.90
22-23	0.14	0.24	0.14	0.13	0.04	0.05	0.12	0.10
24-26	0.27	0.04	0.26	0.25	0.10	0.12	0.23	0.01
27-30	0.21	0.00	0.20	0.20	0.11	0.13	0.19	0.00
31-39	0.22	0.00	0.22	0.23	0.21	0.22	0.23	0.00
40-89	0.14	0.00	0.15	0.16	0.54	0.47	0.19	0.00
Job Type								
Self-employment	0.89	0.81	0.85	0.81	0.79	0.33	0.79	0.48
Unemployed	0.05	0.17	0.05	0.07	0.00	0.00	0.00	0.03
University student	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33
Student	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.04
Laborer	0.00	0.00	0.03	0.03	0.16	0.13	0.13	0.08
Officeholder	0.02	0.00	0.04	0.04	0.04	0.38	0.05	0.01
Military	0.04	0.01	0.03	0.05	0.00	0.05	0.01	0.02
Driver	0.00	0.00	0.00	0.00	0.00	0.10	0.01	0.00
Education								
illiterate	0.00	0.00	0.02	0.02	0.21	0.00	0.06	0.02
diploma	0.77	0.87	0.69	0.60	0.24	0.46	0.45	0.41
primary school	0.02	0.00	0.08	0.09	0.29	0.00	0.02	0.01
guidance school	0.17	0.12	0.18	0.23	0.24	0.17	0.34	0.40
high school	0.00	0.00	0.00	0.00	0.02	0.22	0.12	0.14
Area type								
Rural	0.01	0.04	0.08	0.36	0.08	0.01	0.09	0.08
Urban	0.99	0.96	0.92	0.64	0.92	0.99	0.91	0.92
Collision type								
Fixed object collision	0.09	0.23	0.91	0.03	0.10	0.27	0.15	0.14
Car-motorcycle	0.02	0.04	0.02	0.12	0.01	0.05	0.03	0.03
Pedestrian-motorcycle	0.90	0.73	0.00	0.81	0.88	0.66	0.81	0.82
Crash severity								

Fatal	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.02
Injury	0.94	0.90	1.00	0.32	0.99	0.94	0.93	0.93
Property Damage only	0.06	0.09	0.00	0.67	0.01	0.06	0.06	0.05

As shown in Table 4, Clusters 1, 2 and 3, with more than 70% of the total data account for the largest share of crashes. In addition, 65% of riders are in the age range of 22-40 years.

Cluster 1 with more than 37% of all the crashes contains the largest amount of data among the clusters. About 99% of the riders involved in crashes of this cluster have held a motorcycle license. In addition, the percentage of helmet usage is relatively high in this cluster (42%) compared to the other clusters. In cluster 2, about 96% of riders were under 23 years old and 73% not used helmet.

Most of the motorcyclists in Cluster 3, are between 24-39 years old. Group B is the dominant group of provinces in this cluster. Further, about 91% of the crashes in this cluster are caused due to collision with fixed objects. Moreover, almost all the crashes in this cluster have caused injury.

Cluster 4 with more than 34% of rural crashes, has the largest share of rural crashes among the 8 clusters. This cluster has also the highest percentage of helmet usage (51%) compared to other clusters and is the only cluster with the domination of Group of provinces A. Although the rural crashes are expected to be more severe, but nearly 67% of crashes in this cluster were Property damage only. This might be explained by the high percentage of helmet usage among the riders of this cluster.

Clusters 5 and 6 had the highest percentage of above 30 years old motorcyclists (75% and 69%, respectively). Also, Cluster 5 with an average rider age of 39 years has the highest average age among the 8 clusters. In addition, a considerable percentage of crashes in cluster 5 have occurred in province groups B and E with all the riders being licensed in this cluster. Furthermore, 74% of crashes of cluster 6 have occurred in province group B.

The main feature of clusters 7 and 8 is the percentage of unlicensed riders in these clusters. In cluster 7, 90% of riders were unlicensed and 95% had not used helmet. Similarly, 99% of motorcyclists in cluster 8 were unlicensed and 94% had not used helmet. However, the share of riders' age groups in each of these two clusters are significantly different. Approximately, 90% of riders involved in crashes of cluster 8 were under 21 years old. Additionally, 2% of crashes in this cluster were fatal; the highest percentage among the 8 clusters. This might be attributed to the co-existence of not wearing helmet, unlicensed and under 21 year old riders in this cluster.

In addition, the share of at-fault riders in each of the 8 clusters is shown in Table 5. According to this table, Cluster 3, which contain more than 14 % of crashes in the database (Table 4), have the highest percentage of at-fault riders (99.8%) among all the clusters.

Cluster 6 is the second critical cluster with more than 45 % of at-fault riders. This cluster has the highest percentage of riders of provinces Group B among the 8 clusters (Table 4). Further, although clusters 5 and 6 are similar to each other, but the share of at-fault riders is significantly lower in cluster 5. This might be attributed to the higher share of pedestrian crashes in this cluster. In addition, crashes of Cluster 1 have similar condition, where 90% of crashes are pedestrian-motorcycle collisions and in more than 74% of crashes, the motorcyclists were identified as not-at-fault. Thus, on might conclude that the motorcyclists seem to be less responsible in the pedestrian-motorcycle crashes.

Table 5. Distribution of Rider status variable in clusters

Cluster description		Not-at-fault	At-fault
1	Group of provinces B- Licensed- collision with pedestrian	74.4	25.6
2	Group of provinces B , E-Licensed- No helmet- under 23 years old- collision with pedestrian	57.5	42.5
3	Group of provinces B- No helmet- 30 years old – Injury - Collision with fixed object	0.2	99.8
4	Group of provinces B,A- Licensed- Rural areas	61.0	39.0
5	Group of provinces E,B- Licensed- No helmet- Above 40 years old- Collision with pedestrian	73.5	26.5
6	Group of provinces B- Licensed- Above 40 years old	54.3	45.7
7	Group of provinces E,C- Unlicensed- No helmet	64.7	35.3
8	Group of provinces E,C,B- Unlicensed- No helmet- Under 21 years old	67.9	32.1

4. Discussion and Conclusion

The current study contributes to the literature about motorcycle crashes by providing a holistic view over the crash pattern and fault status of motorcyclists at the national level. For this purpose, two clustering techniques were used in combination. In the first step, the Hierarchical clustering technique was applied to group the provinces according to distribution of crashes in each province. Next, the crash patterns and culpability of riders in each cluster was investigated using the LCC technique.

The LCC analysis produced 8 crash clusters with different share of at-fault motorcyclists. Clusters 2, 3 and 6 have the highest share of at-fault riders. A significant proportion of crashes of these clusters have occurred in province group B. In addition, the share of at-fault motorcyclists were lower in the clusters with domination of pedestrian-motorcycle crashes. Therefore, informing the motorcyclists in these provinces and instructing special training programs for the riders in the license issuance process about the risk of pedestrian-motorcycle crashes could help mitigate this type of crashes in these provinces.

Cluster 4 has the fourth most percentage of at-fault riders. Since this cluster is the only cluster with domination of rural crashes and the only cluster with the domination of province group A; thus, paying more attention to the facilities related to motorcycle safety on the rural roads in these provinces can have a significant impact in reducing crashes.

Clusters 1 and 5 have the lowest percentage of at-fault riders among all clusters. Having riders with an average age of more than 30 years (a relative high age for motorcycle riders) is a common feature among these clusters. This result is in accordance with previous studies (Harrison and Christie, 2005; Rutter and Quine, 1996). Moreover, under 21 years old motorcyclists were involved in more than 90% of crashes in cluster 8. This cluster also had the highest percentage of fatal crashes (2 percent) among the 8 clusters. This issue highlights the influence of inexperience and young riders in the motorcycle crash severity.

Results of this study showed that the idea of combined use of two clustering technique could be helpful to cluster large crash database and provide more homogenous crash groups in order to conduct descriptive preliminary analysis on the crash data from large jurisdictions. The Province group variable resulting from the hierarchical clustering was one of the effective variables in the LCC analysis. This imply the effectiveness of the proposed framework. Results also confirmed that the combined use of HC and LCC can help reveal the pattern of motorcyclists' fault status at a national-level. Since the unbalanced nature of crash data across subnational regions is not peculiar to Iran, the framework adopted in the current study could be used in other similar researches for macro-analysis of crash patterns in other countries or states. Finally, it worth mentioning that this approach might only be regarded as a preliminary analysis tool to provide a holistic view over crash patterns and identify most critical problems at a national level, which in turn might facilitate decision making about road safety issues. Therefore, this approach could complement other more qualitative analytical methods.

5. Limitations

The study variables were limited to only those variables that existed in our crash database. Definitely, there might be other factors that contribute to the motorcycle crashes and the riders' fault status. Furthermore, the injury severity levels that were considered in this study include fatal, injury and no-injury. However, if the injury levels were recorded in more details (i.e., possible, non-incapacitating, and incapacitating injuries), the analysis results could be more helpful.

6. References

-Besharati, M. M. and Kashani, A. T. (2017) "Factors contributing to intercity commercial bus drivers' crash involvement risk". Archives of Environmental and Occupational Health. Available at: http://dx.doi.org/10.1080/19338244.2017.1306478

-Chang, L.-Y. and Chen, W.-C. (2005) "Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research", Vol.36, No.4, pp.365-375.

-de Oña, J., López, G., Mujalli, R. and Calvo, F. J. (2013) "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks", Accident Analysis and Prevention, Vol.51, No.2, pp. 1-10.

-Depaire, B., Wets, G. and Vanhoof, K. (2008) "Traffic accident segmentation by means of latent class clustering", Accident Analysis and Prevention, Vol. 40, No.4, pp.1257-1266.

-Haque, M. M., Chin, H. C. and Huang, H. (2009) "Modeling fault among motorcyclists involved in crashes". Accident Analysis and Prevention, Vol.41, No.2, pp.327-335.

-Haque, M. M., Chin, H. C. and Huang, H. (2010) "Applying Bayesian hierarchical models to examine motorcycle crashes at signalized intersections". Accident Analysis and Prevention, Vol.42, No.1, pp.203-212.

-Harrison, W. A. and Christie, R. (2005) "Exposure survey of motorcyclists in New South Wales", Accident Analysis and Prevention, Vol.37, No.3, pp.441-451.

-Höppner, F. (1999) "Fuzzy cluster analysis: methods for classification, data analysis and image recognition", John Wiley and Sons.

-Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) "Data clustering: a review", ACM Computing Surveys (CSUR), Vol.31, No.3, pp.264-323.

-Jalayer, M. and Zhou, H. (2017) "A multiple correspondence analysis of at-fault motorcycle-involved crashes in Alabama", Journal of Advanced Transportation. Available at: http://dx.doi.org/10.1002/atr.1447

-Kaplan, S. and Prato, C. G. (2013) "Cyclist–motorist crash patterns in Denmark: A latent class clustering approach". Traffic Injury Prevention, Vol.14, No.7, pp.725-733.

-Kim, K. and Boski, J. (2001) "Finding fault in motorcycle crashes in Hawaii: Environmental, temporal, spatial, and human factors". Transportation Research Record, Vol.1779, No.1, pp. 182-188.

-Kim, K. and Li, L. (1996) "Modeling fault among bicyclists and drivers involved in collisions in Hawaii, 1986-1991". Transportation Research Record, Vol.1538, No.1, pp.75-80.

-Kumar, S. and Toshniwal, D. (2016) "A data mining approach to characterize road accident locations". Journal of Modern Transportation, Vol.24, No.1, pp.62-72.

-Lanza, S. T., Collins, L. M., Lemmon, D. R. and Schafer, J. L. (2007) "PROC LCA: A SAS procedure for latent class analysis", Structural equation modeling, Vol.14, No.4, pp.671-694.

-Mannering, F. L. and Bhat, C. R. (2014) "Analytic methods in accident research: Methodological frontier and future directions", Analytic Methods , Accident Research, Vol.1, No.1, pp. 1-22.

-Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F. and Ukkusuri, S. V. (2013) "A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada". Safety science, Vol.54, No.1, pp. 27-37.

-Nylund, K. L., Asparouhov, T. and Muthén, B. O. (2007) "Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study", Structural Equation Modeling, Vol.14, No.4, pp. 535-569.

-O'brien, O., Cheshire, J. and Batty, M. (2014) "Mining bicycle sharing data for generating insights into sustainable transport systems". Journal of Transport Geography, Vol.34, No.1, pp. 262-273.

-Rutter, D. R. and Quine, L. (1996) "Age and experience in motorcycling safety", Accident Analysis and Prevention, Vol.28, No.1, pp.15-21.

-Savolainen, P. and Mannering, F. (2007) "Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes", Accident Analysis and Prevention, Vol.39, No.5, pp. 955-963.

-Schneider, W. H., Savolainen, P. T., Van Boxel, D. and Beverley, R. (2012) "Examination of factors determining fault in two-vehicle motorcycle crashes", Accident Analysis and Prevention, Vol.45, No.1, pp. 669-676.

-Tavakoli Kashani, A. and Besharati, M. M. (2016) "An Analysis of vehicle occupants' injury severity in crashes occurred on rural freeways and multilane highways in Iran", International Journal of Transportation Engineering, Vol.4, No.2, pp. 137-146.

-Tavakoli Kashani, A. and Besharati, M. M. (2016) "Fatality rate of pedestrians and fatal crash involvement rate of drivers in pedestrian crashes: a case study of Iran", International Journal of Injury Control and Safety Promotion, Vol.24, No.2, pp. 222-231.

-Tavakoli Kashani, A., Rabieyan, R. and Besharati, M. M. (2014) "A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers". Journal of safety research, Vol.51, No.1, pp. 93-98.

-Tavakoli Kashani, A., Rabieyan, R. and Besharati, M. M. (2016) "Modeling the effect of operator and passenger characteristics on the fatality risk of motorcycle crashes", Journal of Injury And Violence Research, Vol. 8, No.1, pp. 35.

-Vermunt, J. K., and Magidson, J. (2002) "Latent class cluster analysis", Applied Latent Class Analysis, Vol.11, No.1, pp. 89-106.

-Vermunt, J. K. and Magidson, J. (2005) "Latent GOLD 4.0 user's guide", Tilburg University, Netherlands: TS Social and Behavioral Sciences.

-Vlahogianni, E. I., Yannis, G. and Golias, J. C. (2012) "Overview of critical risk factors in power-two-wheeler safety", Accident Analysis and Prevention, Vol.49, No.1, pp. 12-22.

-Ward Jr, J. H. (1963) "Hierarchical grouping to optimize an objective function", Journal of the American Statistical Association, Vol.58, No.1, pp. 236-244.

-Zhang, G., Yau, K. K., and Zhang, X. (2014) "Analyzing fault and severity in pedestrian–motor vehicle accidents in China", Accident Analysis and Prevention, Vol.73, No.1, pp. 141-150.