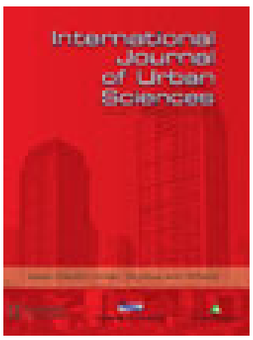

International Journal of Urban Sciences

ISSN: 1226-5934 (Print) 2161-6779 (Online) Journal homepage: http://www.tandfonline.com/loi/rjus20

Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures

Subasish Das, Anandi Dutta, Raul Avelar, Karen Dixon, Xiaoduan Sun & Mohammad Jalayer

To cite this article: Subasish Das, Anandi Dutta, Raul Avelar, Karen Dixon, Xiaoduan Sun & Mohammad Jalayer (2018): Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures, International Journal of Urban Sciences, DOI: 10.1080/12265934.2018.1431146

To link to this article: https://doi.org/10.1080/12265934.2018.1431146

INTERNATIONAL JOURNAL OF URBAN SCIENCES, 2018
https://doi.org/10.1080/12265934.2018.1431146

Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures

Subasish Das [a], Anandi Dutta[b], Raul Avelar[a], Karen Dixon[a], Xiaoduan Sun[c] and Mohammad Jalayer[d]

[a]Texas A&M Transportation Institute (TTI), College Station, TX, USA; [b]Computer Science and Engineering Department, Texas A&M University, College Station, TX, USA; [c]Civil Engineering Department, University of Louisiana at Lafayette, Lafayette, LA, USA; [d]CAIT Central Administration, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

ABSTRACT

In 2011, 4,432 pedestrians were killed (14% of total traffic crash fatalities), and 69,000 pedestrians were injured in vehicle-pedestrian crashes in the United States. Particularly in Louisiana, vehicle-pedestrian crashes have become a key concern because of the high percentage of fatalities in recent years. In 2012, pedestrians were accounted for 17% of all fatalities due to traffic crashes in Louisiana. Alcohol was involved in nearly 40% of these fatalities. This research utilized '*a priori*' algorithm of supervised association mining technique to discover patterns from the vehicle-pedestrian crash database. By using association rules mining, this study aims to discover vehicle-pedestrian crash patterns using eight years of Louisiana crash data (2004–2011). The results indicated that roadway lighting at night helped in alleviating pedestrian crash severity. In addition, a few groups of interest were identified from this study: male pedestrians' greater propensity towards severe and fatal crashes, younger female drivers (15–24) being more crash-prone than other age groups, vulnerable impaired pedestrians even on roadways with lighting at night, middle-aged male pedestrians (35–54) being inclined towards crash occurrence, and dominance of single vehicle crashes. Based on the recognized patterns, this study recommends several countermeasures to alleviate the safety concerns. The findings of this study will help traffic safety professionals in understanding significant patterns and relevant countermeasures to raise awareness and improvements for the potential decrease of pedestrian crashes.

ARTICLE HISTORY
Received 17 August 2017
Accepted 16 January 2018

KEYWORDS
Pedestrian safety; data mining; association rules; pattern identification; rules visualization

1. Introduction

Vehicle-pedestrian crashes constitute a significant concern for roadway traffic safety in the United States. In the twenty-first century, more and more policies are being made to encourage the safer and more effective travel for all roadway users to make the

CONTACT Subasish Das s-das@tti.tamu.edu Texas A&M Transportation Institute, 3135 TAMU, College Station, TX 77843-3135, USA

transportation systems more sustainable and efficient. In 2011, 4,432 pedestrians were killed (14% of total traffic crash fatalities), and 69,000 pedestrians were injured in vehicle-pedestrian crashes in the United States (Schneider, 2013). Accordingly, one of the American Association of State Highway and Transportation (AASHTO) Strategic Highway Safety Plan (SHSP) objectives is to increase pedestrian safety (SHSP, 2005).

Montella, Aria, Dambrosio, & Mauriello (2012) defined traffic crash as a rare, random, multifactor event always preceded by a state in which one or more roadway users fail to cope with the current environment. Any individual crash is the outcome of a series of directly associated or indirectly associated events. In most of the cases, each individual crash is unique, but there exist common occurrences of a few factors in several individual crashes. The data mining method can provide valuable insight into developing pedestrian crash countermeasure by identifying significant patterns from a large traffic crash dataset. This study uses an a priori (relating to knowledge that proceeds from theoretical deduction rather than from observation) algorithm of association rules mining to determine the key association factors in vehicle-pedestrian crashes pulled from a database.

The vehicle-pedestrian crash record in Louisiana requires immediate attention and provides an opportunity for innovative solutions to improve pedestrian safety. The objective of this research was the application of association rules mining on vehicle-pedestrian crashes in Louisiana to (1) extract significant patterns from the crash dataset, (2) find trivial and non-trivial or unsuspected relationship between the key factors, and (3) provide a more profound learning and insight into pedestrian safety.

2. Literature review

Crash frequency analysis and injury severity analysis are considered as the main focus areas of transportation safety studies. Interested readers can explore four extensive review papers that provide a comprehensive list of existing literature on transportation safety research (Lord & Mannering, 2010; Mannering, Sarkar, & Bhat, 2016; Mannering & Bhat, 2014; Savolainen, Mannering, Lord, & Quddus, 2011). Lord and Mannering (2010) provided a comprehensive list of crash count data studies and provided brief overviews of the existing methods and their limitations. Savolainen et al. (2011) conducted a similar review literature study on injury severity analysis. Recently, Mannering and Bhat (2014) extended and bridged both of these studies with newer studies and methods. Mannering et al. (2016) study presented a detailed discussion of unobserved heterogeneity in crash data analysis along with their strengths and weaknesses.

Past research on the practice of data mining in roadway traffic safety is limited but diversified. In transportation safety studies, some of the popular data mining approaches used by researchers include decision trees (Chung, 2013; Lee & Li, 2015; Sohn & Shin, 2001; Zheng, Lu, & Tolliver, 2016), association rules (Das and Sun, 2014b; Das, Minjares-Kyle, Avelar, Dixon, & Bommanayakanahalli, 2017; Das, Mudgal, Dutta, & Geedipally, 2018; Geurts, Wets, Brijs, & Vanhoof, 2003; Mirabadi & Sharifian, 2010; Montella, 2011; Montella et al., 2012; Pande & Abdel-Aty, 2009; Weng, Zhu, Yan, & Liu, 2016), neural networks (Abdelwahab & Abdel-Aty, 2001; Dia & Rose, 1997; Mussone, Ferrari, & Oneta, 1999; Sohn & Shin, 2001), factor analysis (Chen, Zhang, & Xu, 2016), probabilistic graphical dependency models (Song, Ghosh, Miaou, & Mallick, 2006), self-organizing map (Bibeka, Das, Martin, Jalayer, & Munira, 2018),

correspondence analysis (Das, Avelar, Dixon, & Sun, 2018; Das & Sun, 2014a, 2015, 2016; Jalayer, Pour-Raouholamin, & Zhou, 2018; Jalayer & Zhou, 2016), cluster analysis (Kang & Khattak, 2017; Sasidharan, Wu, & Menendez, 2015), text mining and topic modelling (Das, Brimley, Lindheimer, & Pant, 2017; Das, Dutta, & Zupancich, 2017), and sequence mining (Xi, Gao, Niu, Ding, & Ning, 2013; Xi, Zhao, Li, & Wang, 2016).

Geurts et al. (2003) used association algorithm to identify the patterns of crashes in black spots. Pande and Abdel-Aty (2009) used association mining rules in traffic safety analysis by developing closely associated crash characteristics in the form of rules. Mirabadi and Sharifian (2010) applied association rules data mining to perform knowledge discovery and reveal the association patterns of railway crash data in Iran. Montella (2011) employed association rules mining to explore the association between different types of crashes. Montella et al. (2012) used classification trees and association rules mining to perform an exploratory data analysis on pedestrian crashes in Italy. Notable findings from this research include associations between fatal crashes and rural areas, urban provincial and national roads, pedestrians older than 75, nighttime conditions, pedestrians older than 65 in nighttime crashes, young male drivers in nighttime crashes, and truck involvement. Das and Sun (2014) demonstrated how to apply association rules mining methods to discover hidden patterns in rainy weather crash data with eight years of Louisiana data (2004–2011). Weng et al. (2016) used a priori association rules mining to investigate work zone crash causality patterns. Das et al. (2017) used the second Strategic Highway Research Programme's (SHRP-2) Roadway Inventory Database (RID) crash data for Florida rural roadways to investigate improper passing related crashes by using association rules negative binomial (NB) miner.

Existing literature reveals an extensive variety of contributing factors in vehicle-pedestrian crashes. A recently published report (Turner et al., 2017) provided a comprehensive list of studies that have identified a wider range of risk factors for non-motorized trips. Numerous studies identified several key associated factors for pedestrian crashes, which include higher speed limit (30 mph or over) (Davis, 2001; Sze & Wong, 2007), absence of lighting at night (Moudon, Lin, Jiao, Hurvitz, & Reeves, 2011), poor pedestrian visibility (Sullivan & Flannagan, 2011; Ulfarsson, Kim, & Booth, 2010), certain pedestrian age groups (Eluru, Bhat, & Hensher, 2008; Roudsari et al., 2004), driver gender (Lee & Abdel-Aty, 2005), and impaired pedestrians (Oxley, Lenne, & Corben, 2006; Spainhour, Wootton, Sobanjo, & Brady, 2006).

Association rules mining on the U.S. pedestrian crash dataset has not been performed yet, to the extent of the author's knowledge. This study attempts to execute association rules mining using Louisiana pedestrian crashes. Unlike the parametric statistical modelling, association rules mining doesn't require to predefine assumptions. As the method can deal with the large and complex dataset, the generated rules can identify risk factors associated with unorthodox patterns. The findings from this study can contribute in selecting suitable countermeasures and regulations to improve pedestrian safety.

3. Methodology

Data mining aims to identify valid and understandable patterns underlying in a large data set. It helps to extract refine useable knowledge. It involves machine learning, statistical learning, modelling concepts and database management skillset. It is focused on the search and finding of patterns in data rather than the confirmation of hypotheses. For

this reason, data mining method is not only concerned with algorithmic capabilities but also provides tools to accomplish analyzing work without any prior assumptions.

3.1. Overview of association rules mining

Research on algorithms to resolve the frequent itemset/product problem has been abundant (A set of items is referred to as an itemset). Among these algorithms, *a priori*, developed by Agrawal and Srikant, is a level-wise, breadth-first algorithm which counts transactions. This algorithm can be used to mine frequent itemsets, maximal frequent itemsets and closed frequent itemsets. The implementation of the *a priori* algorithm (principle: if an itemset is frequent, then all of its subsets must also be frequent) can additionally be used to generate association rules.

A frequent itemset generation algorithm digs out frequently occurring itemsets, subsequences, or arrangements from large datasets. It is a process that helps the data miners develop strategies by finding out proper associations between different items. A set of definitions are given here before demonstrating the method with an example. Let $I = \{i_1, i_2, \ldots i_m\}$ be a set of items (e.g. a set of crash categories for a particular crash record) and $C = \{c_1, c_2, \ldots, c_n\}$ be a set of database crash information (transaction) where each crash record c_i contains a subset of items chosen from I. An itemset with k items is called as a k-itemset.

An association rule can be expressed as $X \rightarrow Y$, where X and Y are disjoint itemsets. Here, the antecedent is X and the consequent is Y. The strength of the association rule can be measured in terms of the values of support, confidence and lift. The equations of support are listed in equation 1 to equation 3.

$$S(X) = \frac{\sigma(X)}{N} \tag{1}$$

$$S(Y) = \frac{\sigma(Y)}{N} \tag{2}$$

$$S(X \rightarrow Y) = \frac{\sigma(X \cap Y)}{N} \tag{3}$$

where $\sigma(X)$ = Number of incidents with X antecedent; $\sigma(Y)$ = Number of incidents with Y consequent; $\sigma(X \cap Y)$ = Number of incidents with X antecedent; $\sigma(Y)$ = Number of incidents with Y consequent; $\sigma(X \cap Y)$ = Number of incidents with both X antecedent and Y consequent; N = Total number of incidents; $S(X)$ = Support of antecedent; $S(Y)$ = Support of consequent; $S(X \rightarrow Y)$ = Support of the association rule $(X \rightarrow Y)$.

The equations of confidence and lift are listed in equation 4 and 5. Confidence measures the reliability of the inference of a generated rule. A higher confidence for $X \rightarrow Y$ indicates that presence of Y is highly visible in the transactions having X. The lift of the rule makes an association with the frequency of co-occurrence of the antecedent and the consequent to the expected frequency of co-occurrence.

$$C(X \rightarrow Y) = \frac{S(X \rightarrow Y)}{S(X)} \tag{4}$$

$$L(X \rightarrow Y) = \frac{S(X \rightarrow Y)}{S(X) \cdot S(Y)} \tag{5}$$

where $C(X \rightarrow Y)$ = Confidence of the association rule $(X \rightarrow Y)$; $L(X \rightarrow Y)$ = Lift of the association rule $(X \rightarrow Y)$.

A lift value greater than 1 indicates positive interdependence between the antecedent and the consequent, while a value smaller than 1 indicates negative interdependence, a value of one designates independence (Hahsler, Buchta, Gruen,& Hornik, 2018; Dutta, 2016). A rule with a single antecedent and a single consequent is defined as a 2-product rule; similarly, a rule with two antecedents and single consequent or one antecedent and two consequents is defined as a 3-product rule. A critical inference of the association rules is that the generated rules are not needed to be interpreted as causation rather than association.

3.2. Data analysis

This study used eight years of crash records (2004–2011) obtained from the Louisiana Department of Transportation and Development (LADOTD). The data was stored as an unsorted format in Microsoft Access database tables. Every crash record had many variables (attributes) describing what, when and why a crash had occurred. The detailed information regarding the characteristics of crashes, roadways, vehicles and pedestrians is stored in separate data tables. This study prepared a consolidated dataset by merging all of the tables together with the crash identification number. Only at-fault vehicle crashes were considered to determine the association of at-fault driver's factors. The pedestrian crash numbers (with pedestrian information) were filtered from the consolidated dataset to get a complete profiling of the pedestrian-related crashes. To prepare the final consolidated database, cases with single pedestrian involvements were used so that crash identification numbers and pedestrian crash numbers were representative of each other. Figure 1 illustrates a flowchart of the database preparation.

The total number of observations in the final vehicle-pedestrian dataset was 11,503 (for 2004–2011) including 801 fatal crashes, 8,935 injury crashes (severe, moderate or complaint) and 1,767 property damage only (PDO) crashes. Based on the dataset, seven drivers died in pedestrian-related crashes, 754 drivers were injured (severe, moderate or

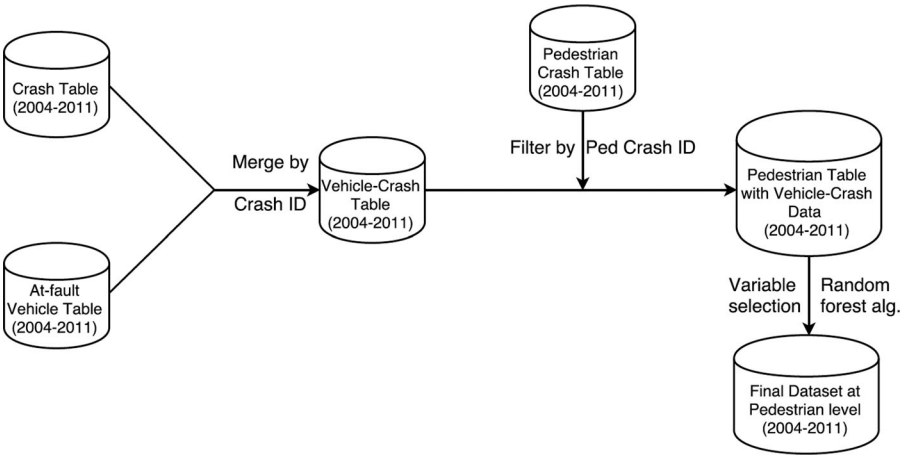

Figure 1. Data preparation flowchart.

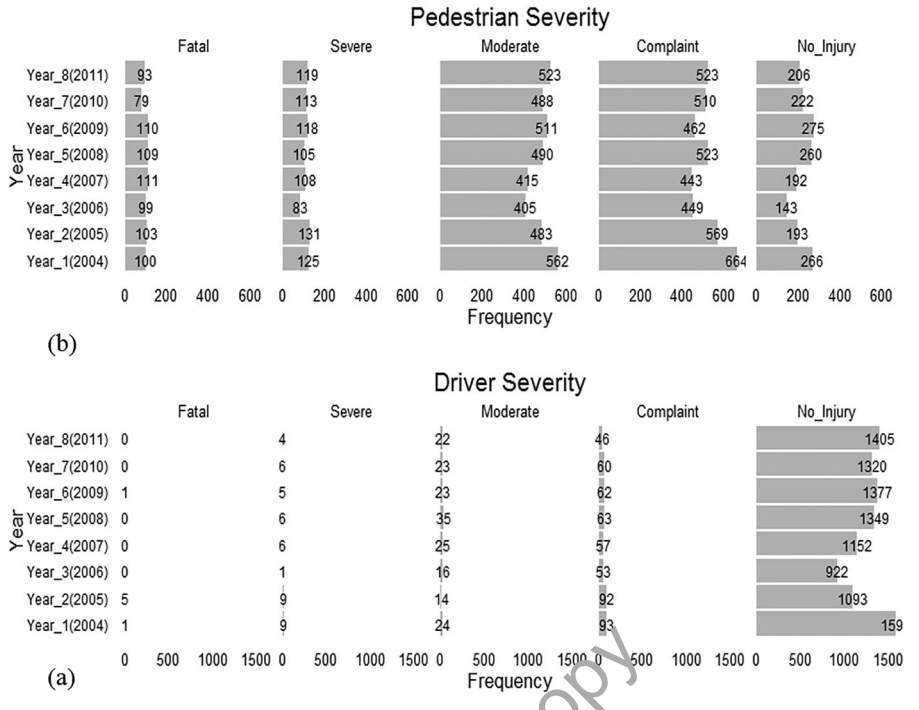

Figure 2. (a) Pedestrian severity in vehicle-pedestrian crashes (b) Driver severity in vehicle-pedestrian crashes.

complaint), and 10,208 of those crashes were PDO crashes. Figure 2 illustrates the outcome of pedestrian severity and driver severity from vehicle-pedestrian crashes. As pedestrians are the exposed group, the severity of pedestrian crashes is much higher than that for the involved driver.

Traffic crash databases contain many variables some of which are redundant in nature for this study (for example, driver's zip code, report modification date, report entered by, driver ejection). In addition, many variables contain missing information (for example, reason for driver's distraction, prior condition before crash, operating speed). To focus on the meaningful analysis, a set of key variables were selected based on prior studies, such as the roadway geometrics (alignment, lighting), collision type, environmental factor (weather), driver related factors (driver gender, age and condition), number of occupants, and pedestrian related factors (pedestrian gender, age, condition and injury severity).

The importance of the variables was ranked by using random forest (RF) algorithms. RF method is based on the bagging principle (Breiman, 2001) and random subspace method (Ho, 1998) that relies on constructing a collection of decision trees with random predictors. Variable selection using random forest algorithm has been widely used by transportation researchers (Das & Sun, 2014b; Das Brimley, Lindheimer, & Pant 2017; Ho, 1998). The overall algorithm can be explained as follows:

1. Produce a cluster sample of size N_C from the overall data with sample size N, to grow the tree by randomly selecting the predictor variables $X = \{x_k,\ i = 1, \ldots, k\}$.

2. Use the predictor x_k at the node of the tree to vote for class label p_B at this node. At each node contains one predictor providing the best split possible.
3. Run the out-of-bag data $N - N_C$ down the tree to gain the misclassification rate, OOB_B
4. Repeat steps 1–3 for large number of trees till the minimum out-of-bag error rate, OOB, is obtained.
5. Assign each observation to a final class k by a majority vote by averaging over the set of trees.

By using this algorithm, a list of final variables was selected based on the higher variable importance values. Table 1 reports the descriptive statistics on the final selected variables.

For the association rules, various settings need to be calibrated in order to arrive at significant findings. Defining adequate minimum support and confidence is essential to generate the meaningful rules; otherwise, the algorithm might produce millions of rules. It is only after a significant number of trials and errors that these minimums are set differently for different cases. One percent of minimum support means that no item or set of items

Table 1. Overview of the vehicle-pedestrian crashes by key variables.

Categories	Frequency	Percentage	Categories	Frequency	Percentage
Alignment (Allign)[a]			Driver Condition (Dr. Cond.)		
Straight-Level	10750	93.45	Normal	6359	55.28
Curve-Level	360	3.13	Impaired	525	4.56
On Grade	174	1.51	Inattentive/Distracted	1914	16.64
Dip, Hump	9	0.08	Sleep, Fatigue	45	0.39
Hillcrest	64	0.56	Other	536	4.66
Unknown (Unk.)	146	1.27	Unknown (Unk.)	2124	18.46
Light			Number Occupant (Num. Occ.)		
Daylight	6272	54.52	One	9021	78.42
Dark – No Street Lights	1442	12.54	Two	1626	14.14
Dark – Street Light	3231	28.09	Three	535	4.65
Dusk, Dawn	358	3.11	Four	164	1.43
Unknown (Unk.)	200	1.74	Five or more	126	1.10
Collision			Unknown (Unk.)	31	0.27
Single Vehicle	4825	41.95	Pedestrian Age (Ped. Age)		
Rear End	466	4.05	15–24	2265	19.69
Right Angle	799	6.95	25–34	1597	13.88
Right Turn	75	0.65	35–44	1450	12.61
Sideswipe	493	4.29	45–54	1484	12.90
Left Turn	209	1.82	55–64	786	6.83
Head-On	185	1.61	65–74	308	2.68
Unknown (Unk.)	4451	38.69	75–Inf	255	2.22
Weather			Unknown (Unk.)	3358	29.19
Clear	8770	76.24	Pedestrian Condition (Ped. Cond.)		
Abnormal	2590	22.52	Normal	4301	37.39
Unknown (Unk.)	143	1.24	Impaired	1232	10.71
Driver Age (Dr. Age)			Inattentive/Distracted	2783	24.19
15–24	2268	19.72	Illness	39	0.34
25–34	2118	18.41	Unknown (Unk.)	3148	27.37
35–44	1638	14.24	Pedestrian Injury (Ped. Inj.)		
45–54	1526	13.27	Fatal	801	6.96
55–64	1036	9.01	Severe	902	7.84
65–74	511	4.44	Moderate	3877	33.70
75–Inf	350	3.04	Complaint	4156	36.13
Unknown (Unk.)	2056	17.87	No Injury	1767	15.36
Driver Gender (Dr. Gender)			Pedestrian Gender (Ped. Gender)		
Female	3759	32.68	Female	3738	32.50
Male	5872	51.05	Male	6958	60.49
Unknown (Unk.)	1872	16.27	Unknown (Unk.)	807	7.02

[a]In the parenthesis, the coded name of the variables is mentioned.

Table 2. Summary chart of the association rules mining.

Case	Consequent	Rules (all)	Rules (Lift ≥ 1)	Support			Confidence			Lift		
				Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
1	Pedestrian Severity = Fatal	43	43	0.020	0.026	0.479	0.100	0.141	0.260	1.445	2.034	3.733
2	Pedestrian Severity = Severe	32	32	0.015	0.020	0.032	0.102	0.115	0.133	1.303	1.472	1.699
3	Pedestrian Severity = Moderate/ Complaint/No Injury	150	122	0.102	0.166	0.361	0.331	0.358	0.443	0.959	1.038	1.225
4	Pedestrian Condition = Impaired	46	46	0.020	0.027	0.048	0.150	0.187	0.245	1.401	1.741	2.290

will be considered frequent for the first analysis if it does not appear in at least 115 traffic crashes (1% of total 11,503 crash records). It can be argued that the choices for the values of these parameters are subjective or case-based. However, the selection of these parameters is not random. One of the reasons behind using supervised rules mining in this study is to get intuitive information from each case by determining the optimized support and confidence values for each case (see Table 2).

A trial and error experiment such as the proposed in this research requires balancing the need for specific answers and the pitfall of slicing the data 'too thin.' On the one hand, setting minimum support too low will result in exponential growth of the number of items in the frequent item sets which will be difficult to interpret. In contrast, by choosing a support parameter too high, the algorithm will be capable of generating a small number of rules, potentially missing important nuances in the data. The purpose of post-processing the association rules set is to identify the subset of interesting rules in a generated set of association rules. A majority of the pedestrian crashes happened in the urban environment; thus, the roadway class is not finally considered for rules generation.

To determine the optimum values of the parameters (support and confidence), different optimization algorithms (for example, genetic algorithm, ant colony optimization) can be applied, which is currently not done in this study. Due to nature of this study (supervised association rules mining), it is also difficult to get rules with high lifts (3 or over). Only a few of the rules will be generated based on the current nature of the database. The authors would consider additional factors and advanced dimensionality reduction method (for example, multiple correspondence analysis) to determine rules with very high lift values (3 or above) in their future studies, which is not currently done in this study.

4. Results and discussion

This study used open source software R and R package 'arules' to conduct the analysis (R Development Core Team, 2006; Hahsler et al., 2018). The primary analysis shows that the dataset has 11,503 rows with 67 items. Figure 3 illustrates the top 20 frequent items. The top five frequent items in the dataset are *Alignment = Straight-level, Number Occupant = One, Weather = Clear, Pedestrian Gender = Male, Driver Condition = Normal*. Performing the association rules mining in an unsupervised way (by keeping minimum support and confidence value at 1% and 30% respectively), results in 156,192 rules, which makes very difficult to extract important patterns. Four different cases were selected based on the nature of the response variables, such as pedestrian severity (Case 1, 2, and 3), and condition of the pedestrian (Case 4).

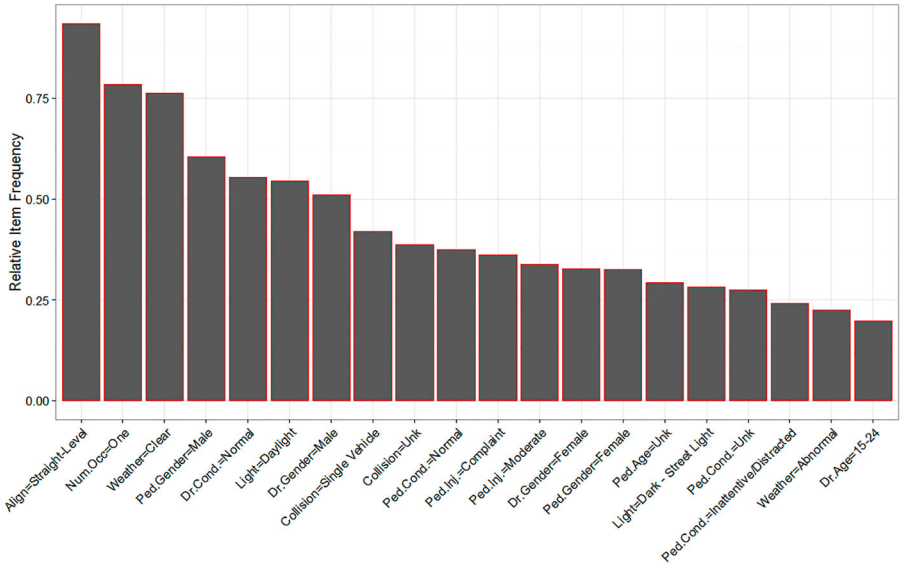

Figure 3. Item frequency plot.

The association rules with the variables of the analyzed cases as consequent were extracted from the generated rules using the *a priori* algorithm. Rules having a shared consequent were sorted according to decreasing value of the lift of the different product rules. Table 2 lists the frequency of the rules generated for different cases and the statistics of support, confidence and lift. Rules are created in two phases. When the minimum threshold of the lift is selected as one, the total counts of the rules are lowered. It is important to note that the minimum values of the supports are determined based on significance and interestingness of the rules.

4.1. Pedestrian severity

In the Louisiana crash database, severity was classified into five groups: fatal, severe, moderate, complaint, and property damage only (PDO). The percentage of fatal and severe crashes for pedestrians is significantly higher than for other road users. This study aims to determine the key factors associated with the different severity types. Three separate cases are considered for further analysis based on pedestrian safety: fatal crashes (Case 1), severe crashes (Case 2) and moderate and complaint injury crashes (Case 3). The rules are shown in Tables 3 –5.

4.1.1. Fatal pedestrian crashes

The association rules with fatal pedestrian severity as consequent were extracted from the generated rules. After performing several iterations, the minimum support was set at 2% and the minimum confidence at 10%. The number of rules generated in this case was 43. All of the rules had a lift higher than 1.

The top 20 rules (based on lift value) generated by keeping pedestrian fatal crashes in consequent is listed in Table 3. Roadways with no street lighting at night (*Light = Dark- No Street Lights*) were highly associated with pedestrian fatal crashes (support = 2.63%,

Table 3. First 20 association rules for pedestrian fatal crashes (Case 1).

Rules	Antecedent	Rule type	Support (%)	Confidence (%)	Lift
1	Light = Dark – No Street Lights	2-product	2.63	20.94	3.008
2	Collision = Single Vehicle	3-product	2.04	26.00	3.733
3	Align = Straight-Level	3-product	2.16	19.38	2.782
4	Collision = Single Vehicle	2-product	4.80	11.44	1.643
5	Dr.Gender = Male	3-product	3.09	14.13	2.029
6	Ped.Gender = Male	3-product	3.49	13.20	1.896
7	Dr.Cond. = Normal	3-product	2.86	11.53	1.655
8	Weather = Clear	3-product	3.58	11.15	1.601
9	Num.Occ = One Align = Straight-Level	3-product	3.49	10.79	1.550
10	Collision = Single Vehicle, Ped.Cond. = Unk	3-product	2.25	22.44	3.223
11	Ped.Cond. = Unk, Dr.Gender = Male	3-product	2.04	15.87	2.279
12	Collision = Single Vehicle	3-product	4.17	10.70	1.537
13	Ped.Cond. = Unk	2-product	3.82	13.95	2.003
14	Collision = Single Vehicle	3-product	2.60	23.77	3.413
15	Ped.Gender = Male	3-product	2.89	18.12	2.602
16	Dr.Gender = Male	3-product	2.35	16.83	2.417
17	Weather = Clear	3-product	2.90	13.82	1.985
18	Num.Occ = One	3-product	2.82	13.20	1.895
19	Align = Straight-Level Dr.Gender = Male	3-product	3.31	13.10	1.882
20	Ped.Gender = Male	3-product	3.20	10.20	1.465

confidence = 20.94%, lift = 3.008). The explanation of the first rule is: 2.63% of vehicle-pedestrian crashes occurred on roadways with no street lights at night and produced fatal severity; Out of all vehicle-pedestrian crashes on roadways with no street lights at night, 20.94% were fatal crashes; the proportion of fatal vehicle-pedestrian crashes on roadways with no street lights at night was 3.008 times the proportion of fatal vehicle-pedestrian crashes in the complete dataset.

Table 4. First 20 association rules for pedestrian severe crashes (Case 2).

Rules	Antecedent	Rule type	Support (%)	Confidence (%)	Lift
1	Light = Dark – No Street Lights	2-product	2.63	20.94	3.008
2	Collision = Single Vehicle	3-product	2.04	26.00	3.733
3	Dr.Gender = Male	3-product	1.79	12.09	1.542
4	Dr.Cond. = Normal	3-product	1.66	11.82	1.507
5	Weather = Clear	3-product	2.49	11.79	1.504
6	Num.Occ = One Weather = Clear	3-product	2.36	10.78	1.375
7	Ped.Cond. = Unk, Ped.Gender = Male Align = Straight-Level	4-product	1.55	12.63	1.611
8	Light = Dark – Street Light	3-product	3.05	11.58	1.476
9	Ped.Cond. = Unk, Ped.Gender = Male	4-product	1.85	12.50	1.594
10	Light = Dark – Street Light, Dr.Gender = Male	4-product	1.70	12.26	1.564
11	Light = Dark – Street Light, Dr.Cond. = Normal	4-product	1.59	11.95	1.524
12	Ped.Cond. = Unk, Dr.Cond. = Normal	4-product	1.55	11.12	1.418
13	Ped.Cond. = Unk, Weather = Clear	4-product	2.00	10.22	1.303
14	Ped.Cond. = Unk	2-product	2.81	10.26	1.308
15	Ped.Gender = Male	3-product	1.99	12.46	1.589
16	Dr.Gender = Male	3-product	1.56	11.22	1.431
17	Dr.Cond. = Normal	3-product	1.65	11.08	1.413
18	Align = Straight-Level	3-product	2.63	10.42	1.329
19	Weather = Clear Collision = Unk	3-product	2.15	10.22	1.304
20	Ped.Gender = Male	3-product	2.34	10.25	1.307

Table 5. First 20 association rules for pedestrian minor injury crashes (Case 3).

Rules	Antecedent	Consequent	Rule type	Support (%)	Confidence (%)	Lift
1	Ped.Cond. = Normal, Num.Occ = One	Ped.Inj. = Complaint	2-product	16.55	44.27	1.225
2	Ped.Gender = Female Align = Straight-Level	Ped.Inj. = Complaint	3-product	10.79	42.66	1.181
3	Ped.Gender = Female	Ped.Inj. = Complaint	3-product	12.96	42.13	1.166
4	Ped.Gender = Female	Ped.Inj. = Moderate	3-product	11.49	37.36	1.108
5	Dr.Gender = Female	Ped.Inj. = Complaint	3-product	11.62	37.63	1.042
6	Collision = Unk	Ped.Inj. = Moderate	3-product	12.60	34.74	1.031
7	Collision = Single Vehicle Weather = Clear	Ped.Inj. = Moderate	3-product	13.50	34.62	1.027
8	Ped.Gender = Female	Ped.Inj. = Complaint	3-product	10.49	42.11	1.166
9	Ped.Gender = Female	Ped.Inj. = Complaint	2-product	13.63	41.95	1.161
10	Ped.Gender = Female	Ped.Inj. = Moderate	2-product	12.05	37.08	1.100
11	Light = Daylight	Ped.Inj. = Complaint	2-product	22.21	40.74	1.128
12	Dr.Cond. = Normal	Ped.Inj. = Moderate	3-product	11.70	36.95	1.096
13	Ped.Gender = Male	Ped.Inj. = Moderate	3-product	11.36	36.00	1.068
14	Dr.Cond. = Normal	Ped.Inj. = Moderate	2-product	19.82	35.85	1.064
15	Dr.Gender = Male Dr.Gender = Male	Ped.Inj. = Moderate	3-product	11.53	35.75	1.061
16	Ped.Gender = Male	Ped.Inj. = Moderate	3-product	10.99	35.03	1.039
17	Ped.Gender = Male Collision = Single Vehicle	Ped.Inj. = Moderate	2-product	21.50	35.54	1.055
18	Num.Occ = One	Ped.Inj. = Moderate	3-product	11.45	35.37	1.049
19	Dr.Gender = Female	Ped.Inj. = Complaint	2-product	12.15	37.19	1.029
20	Collision = Unk	Ped.Inj. = Moderate	2-product	13.40	34.62	1.027

The highest lift value (lift = 3.733) was found for a 3-product rule: *Light = Dark- No Street Lights, Collision = Single Vehicle → Pedestrian Injury = Fatal*, indicating that the proportion of fatal outcomes for vehicle-pedestrian crashes involving single vehicles (only one vehicle involved in pedestrian crash) and no street lights is almost four times the proportion of fatalities for all vehicle-pedestrian crashes. The Moudon et al. study (2011) also found that collision taking place under dark conditions increased the risk of the pedestrian dying or being severely injured. One possible reason is that majority of the pedestrian crashes are single vehicle crashes (about 42% in total crashes, and about 70% in no lighting crashes). The rules also show that the drivers and pedestrians involved in vehicle-pedestrian fatal crashes are overrepresented for male drivers. The study of Lee and Abdel-Aty (2005) also found that male drivers were more likely to be at fault in pedestrian crashes than female drivers. Among different collision types, single vehicle crashes (Rules 4–9) were dominant. Moreover, the roadways with straight-level alignment (Rules 10–12) were significant in the generated rules. Additionally, the proportion of fatal vehicle-pedestrian crashes was higher when the driver was normal in condition. Another interesting finding is that the condition of the pedestrian involved in fatal crashes was mostly unknown which is mainly due to the lack of input on that category in the police report.

4.1.2. Severe pedestrian crashes

The association rules for severe pedestrian crash as consequent were extracted from the generated rules. After performing several iterations, the minimum support was set at 1.5% and minimum confidence at 10%. The number of rules generated in this case was 32. All of the rules had lift value over 1.

Top 20 rules (based on lift value) generated by keeping pedestrian severe crashes in consequent are listed in Table 4. The higher lift value for severe crashes (lift = 3.733) is found for the 3-product rule: *Light = Dark- No Street Lights, Collision = Single Vehicle → Pedestrian Injury = Severe*. This result is consistent with the outcome of fatal crashes and the findings of the study of Moudon et al. (2011). However, lighting shows some mixed effects on pedestrian severities in several studies (Uttley & Fotios, 2017; Welch, Zhang, & Jiao, 2017). The factors like pavement alignments, and straight-level roadways (Rules 8–13) were dominant in severe pedestrian crashes. Moreover, the condition of the pedestrians involved in severe crashes was mostly unknown; which is similar to fatal crashes. While reporting on pedestrian crashes, police officers do not have often had available information about the condition of the pedestrians. In most cases, the condition is described as 'unknown'. The dominant gender in the antecedent part of the severe crashes was male. This finding matches with the findings of the study conducted by Lee and Abdel-Aty (2005).

4.1.3. Minor injury pedestrian crashes

Moderate, and complaint injury crashes for pedestrians were nearly 85% of the pedestrian related crashes. The association rules for moderate/complaint pedestrian crash as consequent were extracted from the generated rules. After performing several iterations, the minimum support was set at 10% and minimum confidence at 33%. The number of rules generated in this case was 150; out of them, 122 had a lift value higher than 1.

Unlike the first two cases, the consequent is not fixed at one attribute in Case 3. The lowest injury group (*Pedestrian Severity = Moderate, or Complaint*) was kept as consequent. The top 20 rules (based on lift value) generated by keeping the severity of this group (minor injury crashes) are listed in Table 5. The rule with the highest lift value (support = 16.55%, confidence = 44.27%, lift = 1.225) is: *Pedestrian Condition = Normal → Pedestrian Injury = Complaint*. This rule can be explained as follows: 16.55% of the total pedestrian crashes involved pedestrians with normal condition and produced complaint severity, 44.27% of the crashes with pedestrian with normal condition were minor injury crashes, and proportion of minor injury outcomes for pedestrians with normal condition was 1.225 times the proportion of crashes of such injury level in the complete dataset.

The findings indicated that females had a higher propensity than males in being involved minor injury crashes. Among differently aligned roadways, straight-level roadways (Rules 3–7) were dominant in moderate, or complaint pedestrian crashes. The rules associated with females produced higher lift values in this case. Lee and Abdel-Aty (2005) showed that male drivers were more likely to be at fault in pedestrian crashes than female drivers. Two other studies attempted to associated pedestrian gender with crashes and driver behaviour (Guéguen, Eyssartier, & Meineria, 2016; Tiwari, Bangdiwala, Saraswat, & Gaurav, 2007). The findings are partly in line with the findings of this study.

Grouped balloon plots are useful in visualizing the larger number of rules with different consequents (see Figure 4). The size of the balloon indicates the support values. The colour intensity indicates the values of the lift. The rules discussed above can be easily identified from this illustration based on the high-lift and high-support combinations.

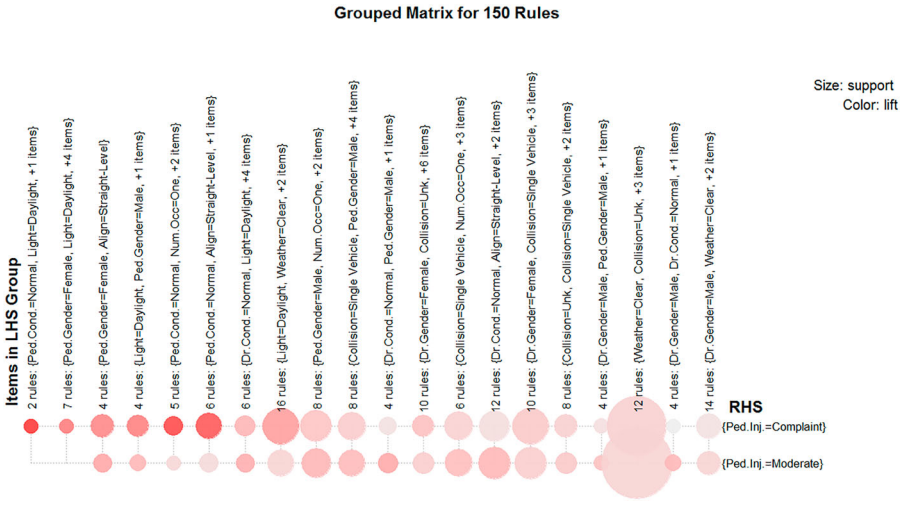

Figure 4. Grouped balloon plot of 150 rules generated for minor and compliant injury crashes.

4.2. Pedestrian condition

An impaired pedestrian is a significant factor in the high toll of pedestrian crashes. Other studies showed that pedestrians under the influence of alcohol have been shown to engage in risky road- crossing behaviours (Oxley et al., 2005; Spainhour et al., 2006). The association rules for impaired pedestrian crash as consequent were extracted from the generated rules. In this case, the minimum support was set at 2% and minimum confidence at 15%. The number of rules generated in this case was 38; all of which had a lift value higher than 1. Table 6 lists the top two rules based on the higher lift values.

Table 6. First 20 association rules for impaired pedestrian crashes (Case 4).

Rules	Antecedent	Rule type	Support (%)	Confidence (%)	Lift
1	Ped.Age = 45–54	2-product	2.59	20.08	1.875
2	Ped.Gender = Male	3-product	2.06	23.65	2.208
3	Ped.Age = 35–44, Align = Straight-Level	2-product	2.35	18.62	1.739
4	Light = Dark – No Street Lights	3-product	2.47	22.19	2.072
5	Ped.Age = 45–54	3-product	2.36	19.83	1.851
6	Collision = Single Vehicle	3-product	2.18	19.78	1.847
7	Ped.Age = 35–44	3-product	2.12	18.11	1.691
8	Light = Dark – Street Light	3-product	4.48	16.99	1.586
9	Light = Dark – No Street Lights	2-product	2.75	21.91	2.046
10	Weather = Clear	3-product	2.06	21.55	2.012
11	Light = Dark – Street Light	2-product	4.88	17.36	1.621
12	Dr.Cond. = Normal	3-product	2.94	20.92	1.953
13	Collision = Single Vehicle	3-product	2.43	20.62	1.925
14	Ped.Gender = Male	3-product	3.55	20.59	1.922
15	Dr.Gender = Male	3-product	2.67	18.02	1.682
16	Weather = Clear	3-product	3.61	17.11	1.598
17	Num.Occ = One Num.Occ = One	3-product	3.69	16.87	1.575
18	Ped.Age = 45–54 Collision = Single Vehicle	2-product	2.01	19.86	1.855
19	Ped.Gender = Male Collision = Single Vehicle	3-product	4.05	15.34	1.433
20	Dr.Cond. = Normal	3-product	3.76	15.14	1.413

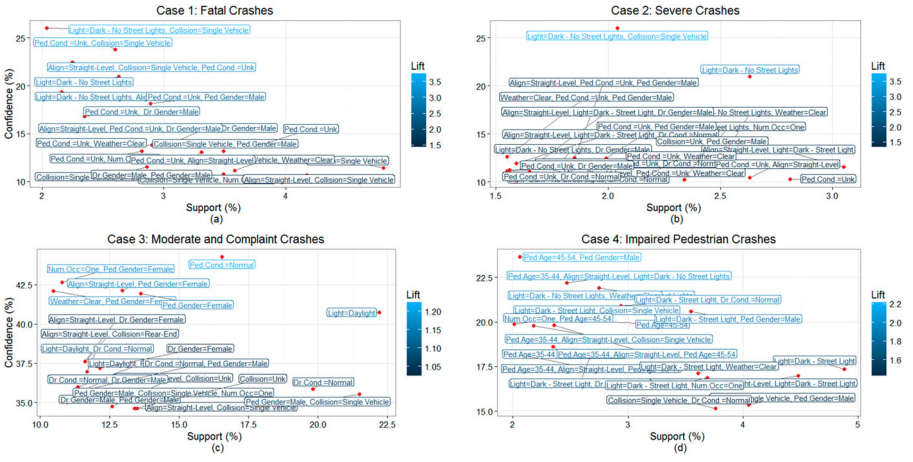

Figure 5. (a) Illustration of the rules for Case 1, (b) Illustration of the rules for Case 2, (c) Illustration of the rules for Case 3, and (d) Illustration of the rules for Case 4.

The top 20 rules (based on lift value) generated by impaired pedestrians as consequent are listed in Table 6. The 2-product rule with the highest lift value (lift = 2.046) is: *Light = Dark- No Street Light→ Pedestrian Condition = Impaired*. The rule with the highest lift value (lift = 2.208) is: *Pedestrian Age = 45–54, Pedestrian Gender = Male → Pedestrian Condition = Impaired*. Two other significant rules are: *Pedestrian Age = 45–54 → Pedestrian Condition = Impaired* and *Pedestrian Age = 35–44 → Pedestrian Condition = Impaired*. The top three rules indicate that middle-aged pedestrians (35–54), male pedestrians, and roadways with no lighting were the dominant factors among the impaired pedestrians. Lighting shows some mixed effects on pedestrian severities in several studies (Moudon et al., 2011; Uttley & Fotios, 2017; Welch et al., 2017). Impaired pedestrians were vulnerable even on straight level roadways (Rules 4–8) and on roadways with lighting at night (Rules 11–17). The dominant gender in the antecedent part of the impaired pedestrian crashes was male. One potential reason might the higher number of male pedestrians. As the exposure rate of gender is not available, this association requires more justification.

Figure 5 illustrates the rules generated from all four cases. As the purpose of this study is to identify significant patterns from the most associated and significant factors, the rules are highlighted in the plots are based on the values of support, confidence, and lift.

5. Conclusions

This study used association rules mining method to analyze the vehicle-pedestrian crashes to extract intuitive knowledge for pedestrian safety improvement. In traffic safety analysis, usually there remains strong correlation between the associated variables. In parametric modelling, there is an essential assumption that the explanatory variables are required to be independent in nature. A non-parametric method like association rules mining has some advantages here as this method does not limit the distribution assumption of variable and their association. The generated trivial and non-trivial association rules

provided valuable insights on the risk factors in pedestrian safety. Moreover, the generated rules can identify significant trends with specific details. Identifying patterns in such level is not possible with conventional statistical modelling.

The findings show that the association rules defined several risk factors and focus groups in pedestrian crashes while some of the rules may be sound trivial findings. The results of this study provide key factors associated with being found at fault. These are:

- Roadways with no lighting at night seemed significantly fatality prone. The recommendation is to provide lighting at high risk locations. It is also important to educate drivers about the difficulty in detecting pedestrians at night, stopping distances and the limitations of headlights. Additionally, educational campaigns are needed to encourage wearing retroreflective materials while walking at night.
- Impaired pedestrians were involved in crashes even on roadways with lighting at night. Educational training programmes should include mandatory components for impaired walking and driving.
- Male drivers showed a great propensity towards fatal and severe crashes. Programmes are needed to enforce existing traffic laws and ordinances for motorists (e.g. keep safe speed limits, yield to pedestrians when turning, traffic signal compliance, obeying drunk-driving laws).
- Middle-aged (35–54) male pedestrians were significant in numbers in the impaired pedestrian group. Educational programmes and campaigns are needed to enforce existing traffic laws and ordinances for pedestrians (e.g. more cautious while crossing roads, crossing the street at legal crossings, obeying traffic and pedestrian signals).
- Younger pedestrians (15–24) were more vulnerable towards traffic crashes. Young pedestrian safety campaigns are necessary to make young pedestrians follow safety rules.
- Female pedestrians showed the greater propensity towards moderate, or complaint crash occurrence. Educational programmes and campaigns are needed to enforce existing traffic laws and ordinances for pedestrians (e.g. more cautious while crossing roads, crossing the street at legal crossings, obeying traffic and pedestrian signals).

The results have consistently indicated that well-lit roadway segments reduce the injury severity of pedestrians. The findings also show that females have lower risk than males in being involved in fatal and severe pedestrian crashes. Other studies also showed that male pedestrians are over represented in vehicle-pedestrian crashes (Clifton & Livi, 2005; Hijar, 2001).

The intoxication of pedestrian had strong effects even on the roadways with lighting at night. Key focus groups for pedestrian crashes are: impaired middle-aged male pedestrians (35–54), younger pedestrians (15–24), and inattentive/distracted male drivers. Among different collision types, single vehicle crashes are dominant in pedestrian involvement. The findings also indicate that in many cases the condition (distracted/impaired) of the pedestrian involved in crashes is unknown which implies that these values are missing in the state maintained police crash reports. The results of this study can be used to increase awareness of pedestrian and driver behaviour that contributes to crash occurrence.

This paper presents a novel approach on how to extract key contributing patterns from a set of pedestrian crash incidents. One of the major contributions of this paper is that it

applies supervised association rules mining algorithm to determine more domain specific patterns associated with pedestrian crashes, which provides better knowledge from conventional rules mining studies. Unlike machine learning approach, this method provides the patterns with higher lift values which can be considered as the parameter for determining the significance of the pattern from the base condition. In summary, the framework developed in this study contributes to a better understanding of pedestrian associated traffic crashes and can provide meaningful information to help city planners and decision makers to improve safety.

Determination of the most suitable support and confidence is an issue that needs further exploration. Future study will focus on the full range of variable values to develop the complete association between the key factors which, hopefully, will lead to new knowledge about vehicle-pedestrian crashes. Another limitation of the paper is that the current study did not explore the issues specific to intersection geometric design and availability of countermeasures. Countermeasures like a hybrid flashing beacon, pedestrian countdown timer shows high safety effectiveness in some recent studies (Fitzpatrick & Park, 2010; Houten, LaPlante, & Gustafson, 2012; Zegeer et al., 2017). A more in-depth look at the nature of crashes at intersections with different pedestrian safety countermeasures using data mining technique would help in identifying patterns and scopes for significant improvements. This issue has potential to be investigated in future studies.

Acknowledgements

The authors are grateful to two reviewers for providing important comments and suggestions which enhanced the original version of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Subasish Das http://orcid.org/0000-0002-1671-2753

References

Abdelwahab, H., & Abdel-Aty, M. (2001). Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record: Journal of the Transportation Research Board*, *1746*, 6–13.

Bibeka, A., Das, S., Martin, M., Jalayer, M., & Munira, S. (2018, January 7–11). *Macro level analysis of pedestrian and bike commuting: Using self-organizing map to associate crime, poverty, and demographics*. The 97th transportation research board annual meeting, Washington, DC.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, T., Zhang, C., & Xu, L. (2016). Factor analysis of fatal road traffic crashes with massive casualties in China. *Advances in Mechanical Engineering*, *8*(4), 1–11.

Chung, Y. (2013). Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. *Accident Analysis & Prevention*, *61*, 107–118.

Clifton, K., & Livi, A. (2005). *Gender differences in walking behavior, attitudes about walking, and perceptions of the environment in three Maryland communities*. The transportation research board annual meeting proceeding papers, Washington, DC.

Das, S., Avelar, R., Dixon, K., & Sun, X. (2018). Investigation on the wrong way driving crash patterns using multiple correspondence analysis. *Accident Analysis and Prevention*, *111*, 43–55.
Das, S., Brimley, B., Lindheimer, T., & Pant, A. (2017). *Safety impacts of reduced visibility in inclement weather* (Report No.: ATLAS-2017-19).
Das, S., Dutta, A., & Zupancich, M. (2017, January 11). *Text mining on 100 years of air crash narratives: Key findings*. The 96th transportation research board annual meeting, Washington, DC.
Das, S., Minjares-Kyle, L., Avelar, R., Dixon, K., & Bommanayakanahalli, B. (2017, January 8–12). *Improper passing-related crashes on rural roadways: Using association rules negative binomial miner*. The 96th transportation research board annual meeting, Washington, DC.
Das, S., Mudgal, A., Dutta, A., & Geedipally, S. (2018, January). *Vehicle consumer complaint reports involving severe incidents: Mining large contingency tables*. The 97th transportation research board annual meeting compendium papers, Washington, DC.
Das, S., & Sun, X. (2014a, January 12–16). *Exploring clusters of contributing factors for single-vehicle fatal crashes through multiple correspondence analysis*. The 93rd transportation research board annual meeting, Washington, DC.
Das, S., & Sun, X. (2014b). *Investigating the pattern of traffic crashes under rainy weather by association rules in data mining*. TRB 93rd annual meeting compendium of papers.
Das, S., & Sun, X. (2015). Factor association with multiple correspondence analysis in vehicle-pedestrian crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2519*, 95–103.
Das, S., & Sun, X. (2016). Association knowledge for fatal run-off-road crashes by multiple correspondence analysis. *IATSS Research*, *39*(2), 146–155.
Davis, G. (2001). Relating severity of pedestrian injury to impact speed in vehicle–pedestrian crashes: Simple threshold model. *Transportation Research Record: Journal of the Transportation Research Board*, *1773*(1), 108–113.
Dia, H., & Rose, G. (1997). Development and evaluation of neural network freeway incident detection models using field data. *Transportation Research Part C: Emerging Technologies*, *5*(5), 313–331.
Dutta, A. (2016). *A smart design framework for a novel reconfigurable multi-processor systems-on-chip (ASREM) architecture* (*Doctoral Dissertation*). Computer Science and Engineering, University of Louisiana at Lafayette.
Eluru, N., Bhat, C., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis and Prevention*, *40*(3), 1033–1054.
Fitzpatrick, K., & Park, E. (2010). *Safety effectiveness of the HAWK pedestrian crossing treatment* (Report No. FHWA-HRT-10-042).
Geurts, K., Wets, G., Brijs, T., & Vanhoof, K. (2003). Profiling of high-frequency accident locations by use of association rules. *Transportation Research Record: Journal of the Transportation Research Board*, *1840*(1), 123–130.
Guéguen, N., Eyssartier, C., & Meineria, S. (2016). A pedestrian's smile and drivers' behavior: When a smile increases careful driving. *Journal of Safety Research*, *56*, 83–88.
Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2018). *Arules: Mining Association Rules and Frequent Itemsets*. Retrieved from http://cran.r-project.org/web/packages/arules/arules.pdf
Hijar, M. (2001). Analysis of fatal pedestrian injuries in Mexico city, 1994–1997. *Injury: International Journal of the Care of the Injured*, *32*, 279–284.
Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844.
Houten, R., LaPlante, J., & Gustafson, T. (2012). *Evaluating pedestrian safety improvements* (Report No. RC-1585). Michigan Department of Transportation.
Jalayer, M., Pour-Raouholamin, M., & Zhou, H. (2018). Wrong-way driving crashes: A multiple correspondence approach to identify contributing factors. *Traffic Injury Prevention*, *19*(1), 35–41.
Jalayer, M., & Zhou, H. (2016). A multiple correspondence analysis of at-fault motorcycle-involved crashes in Alabama. *Journal of Advanced Transportation*, *50*, 2089–2099.

Kang, Y., & Khattak, A. (2017). Cluster-based approach to analyzing crash injury severity at highway–rail grade crossings. *Transportation Research Record: Journal of the Transportation Research Board*, *2608*, 58–69.
Lee, C., & Abdel-Aty, M. (2005). Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. *Accident Analysis and Prevention*, *37*(4), 775–786.
Lee, C., & Li, X. (2015). Predicting driver injury severity in single-vehicle and two-vehicle crashes with boosted regression trees. *Transportation Research Record: Journal of the Transportation Research Board*, *2514*, 138–148.
Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305.
Mannering, F., & Bhat, C. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, *1*, 1–22.
Mannering, F., Sarkar, V., & Bhat, C. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, *11*, 1–16.
Mirabadi, A., & Sharifian, S. (2010). Application of association rules in Iranian railways (RAI) accident data analysis. *Safety Science*, *48*(10), 1427–1435.
Montella, A. (2011). Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. *Accident Analysis and Prevention*, *43*(4), 1451–1463.
Montella, A., Aria, M., Dambrosio, A., & Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis and Prevention*, *49*, 58–72.
Moudon, A., Lin, L., Jiao, J., Hurvitz, P., & Reeves, P. (2011). The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King county, Washington. *Accident Analysis and Prevention*, *43*(1), 11–24.
Mussone, L., Ferrari, A., & Oneta, M. (1999). An analysis of urban collisions using an artificial intelligence model. *Accident Analysis and Prevention*, *31*(6), 705–718.
Oxley, J., Lenne, M., & Corben, B. (2006). The effect of alcohol impairment on road-crossing behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour*, *9*(4), 258–268.
Pande, A., & Abdel-Aty, M. (2009). Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. *Safety Science*, *47*, 145–154.
R Development Core Team. (2006). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org
Roudsari, B., Mock, C., Kaufman, R., Grossman, D., Henary, B., & Crandall, J. (2004). Pedestrian crashes: Higher injury severity and mortality rate for light truck vehicles compared with passenger vehicles. *Injury Prevention*, *10*(3), 154–158.
Sasidharan, L., Wu, K., & Menendez, M. (2015). Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. *Accident Analysis & Prevention*, *85*, 219–228.
Savolainen, P., Mannering, F., Lord, D., & Quddus, M. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, *43*(5), 1666–1676.
Schneider, H. (2013). *Louisiana traffic records data report 2012*. Baton Rouge: Highway Safety Research Group, Louisiana State University.
Sohn, S., & Shin, H. (2001). Pattern recognition for road traffic accident severity in Korea. *Ergonomics*, *44*(1), 107–117.
Song, J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, *97*(1), 246–273.
Spainhour, L., Wootton, I., Sobanjo, J., & Brady, P. (2006). Causative factors and trends in Florida pedestrian crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *1982*, 90–98.
Strategic Highway Safety Plan (SHSP). (2005). *A comprehensive plan to substantially reduce vehicle-related fatalities and injuries on the nation's highways*. Washington, DC: AASHTO.

Sullivan, J., & Flannagan, M. (2011). Differences in geometry of pedestrian crashes in daylight and darkness. *Journal of Safety Research*, *42*(1), 33–37.

Sze, N., & Wong, S. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. *Accident Analysis and Prevention*, *39*(6), 1267–1278.

Tiwari, G., Bangdiwala, S., Saraswat, A., & Gaurav, S. (2007). Survival analysis: Pedestrian risk exposure at signalized intersections. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(2), 77–89.

Turner, S., Sener, I., Martin, M., Das, S., Shipp, E., Hampshire, R., … Robinson, S. (2017). *Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities* (Report No. FHWA-SA-17-041).

Ulfarsson, G., Kim, S., & Booth, K. (2010). Analyzing fault in pedestrian–motor vehicle crashes in North Carolina. *Accident Analysis and Prevention*, *42*(6), 1805–1813.

Uttley, J., & Fotios, S. (2017). The effect of ambient light condition on road traffic collisions involving pedestrians on pedestrian crossings. *Accident Analysis and Prevention*, *108*, 189–200.

Welch, E., Zhang, M., & Jiao, J. (2017). *Identifying factors explaining pedestrian crash severity: A study of Austin, Texas*. The proceedings of transportation research board 96th annual meeting, Washington, DC.

Weng, J., Zhu, J., Yan, X., & Liu, Z. (2016). Investigation of work zone crash casualty patterns using association rules. *Accident Analysis and Preventiom*, *92*, 43–52.

Xi, J., Gao, Z., Niu, S., Ding, T., & Ning, G. (2013). A hybrid algorithm of traffic accident data mining on cause analysis. *Mathematical Problems in Engineering*, *2013*, 8.

Xi, J., Zhao, Z., Li, W., & Wang, Q. (2016). A traffic accident causation analysis method based on AHP-Apriori. *Procedia Engineering*, *137*, 680–687.

Zegeer, C., Srinivasan, R., Lan, B., Carter, D., Smith, S., Sundstrom, C. … Houten, R. (2017). *NCHRP report 841: Development of crash modification factors for uncontrolled pedestrian crossing treatments*. Washington, DC: Transportation Research Board.

Zheng, Z., Lu, P., & Tolliver, D. (2016). Decision tree approach to accident prediction for highway–rail grade crossings. *Transportation Research Record: Journal of the Transportation Research Board*, *2545*, 115–122.