International Information and Engineering Technology Association
Advancing the World of Information and Engineering

International Journal of Safety and Security Engineering
Vol. 12, No. 1, February, 2022, pp. 55-64
Journal homepage: http://iieta.org/journals/ijsse

Analysis of Spatial Concentrations of Large-Truck Crashes Using Data Mining Methods

Syed As-Sadeq Tahfim*, Chen Yan

Department of Maritime Economics and Management, Dalian Maritime University, Dalian 116026, China

Corresponding Author Email: tahfim1963@dlmu.edu.cn

https://doi.org/10.18280/ijsse.120107

Received: 25 September 2021
Accepted: 12 January 2022

Keywords:
large-truck crashes, spatial concentrations, DBSCAN clustering, association rule mining

ABSTRACT

In recent years, the number of studies on crashes involving large-trucks has increased due to its importance to the economy and the higher chance of fatalities. However, none of the previous studies has given attention to the spatial concentrations of large-truck crashes. Moreover, the literature lacks exploration of granular level land use and urban design factors. The current study used the DBSCAN (Density-Based Spatial Clustering of Application with Noise) method to identify the spatial concentrations of crashes involving large-trucks. Additionally, the study explored housing, population, employment, and road network density attributes along with the crash characteristics, roadway attributes, location type, traffic conditions, driver's action and behavior, and environmental factors. The association rule analysis was employed to discover the contributory factors that lead to no injury, non-severe and severe injuries at the spatial concentrations of crashes involving large-trucks. The findings indicated that the rear-end collisions involving drunk drivers often lead to severe injuries in large-truck crashes. Non-interstate roads, speed limit from 40 to 80 kilometers per hour, high road network density, medium and high population density are frequent conditions of non-severe injuries. Lastly, collisions between large-trucks and fixed objects, sideswipe same direction collisions, snowy roads, clear weather, medium road network and employment density are likely to facilitate no injury crashes involving large-trucks. Road traffic authorities can use these insights to reduce the frequency and severity of crashes involving large-trucks at their spatial concentrations.

1. INTRODUCTION

The value of shipments moved by the trucks is the highest in the United States. From 2012 to 2015, the value of shipments moved by the trucks increased by 6.38% approximately [1]. The quantity of shipments is more likely to increase in the coming years due to the higher demand for goods and services. As the quantity of shipments grows, the risk of crashes involving large-trucks also increases. A crash event involving large-trucks can interrupt traffic flow, and lead to severe injuries for drivers, occupants and other vulnerable road users (i.e. pedestrians & cyclists). According to the National Safety Council, the number of fatal crashes that involved large-trucks was 5005 in 2019, which is a 2% increase from 2018. A large-truck is typically defined as "any medium or heavy truck with gross vehicle weight more than 10,000 pounds", excluding the buses and motor homes. The unique characteristics such as the size and weight of the large-trucks, and the fatigue caused by long driving hours in the truck drivers increase the chances of severe or fatal injuries in crashes. The importance of the large-trucks to the economy, and the threats to road safety by the large-trucks validate the special attention that has been given to the analysis of large-truck crashes.

The number of studies on large-truck-related crashes has increased over the years. Most of the studies tried to identify the contributory factors that influence the frequency of large-truck-related crashes [2, 3], the severity of injury in crashes involving large-trucks [3-5], and the risk of occurrence of large-truck involved crashes [6]. Some studies analyzed the crashes involving large-trucks for specific area types (e.g. urban, rural, mountainous freeway) [7, 8] or roadway types (e.g. cross or t-intersection, freeways, highways) [9-12]. A good number of studies have explored and analyzed the spatial clusters of traffic crashes [13-16]. However, very few studies have analyzed the spatial concentration of crashes involving large-trucks. The spatial concentrations of crashes involving large-trucks are different than that of other vehicles due to their unique characteristics and origin-destination routes of large-trucks. Analysis of the spatial concentrations can reveal important insights about the crash contributory factors that frequently occur together in large-truck crash-prone locations. Moreover, the previous studies have rarely explored the land use and urban design attributes in their analysis of crashes involving large-trucks. The knowledge discovered from the analysis of the spatial concentrations of large-truck crashes can improve the road safety policies and make the operation of the trucking industry safer. The objective of the current study is to discover the conditions that influence the different levels of injury severity at the spatial concentrations of crashes involving large-trucks using data mining methods.

2. LITERATURE REVIEW

2.1 The contributory factors of crashes involving large-trucks

The previous studies have identified a wide variety of factors that influence the frequency, the severity of injury and the occurrence of crashes involving large-trucks. The categories of the contributory factors include the crash characteristics (e.g., collision type), vehicle characteristics (e.g., size and weight), driver demographics and actions (e.g., age, moving straight), roadway attributes (e.g., alignment, grade), land use (e.g., rural & urban), traffic conditions (e.g., traffic control & speed limit), and environmental factors (e.g., weather, lighting). Dong et al. [3] reported that driver age, speed limit, and location type have significant influence only on the frequency of crashes involving large trucks. Another study developed a driver-focused truck crash prediction model and found that age, weight, height, gender, employment stability, and previous driving and vehicle history are significantly related to the likelihood of truck crash occurrence.

Different collision type affects the injury outcome of crashes involving large-trucks differently under disparate conditions. Uddin & Huynh [4], and Zhu & Srinivasan [5] both reported that the rear-end collisions decrease the probability of major injury. Zhu & Srinivasan [5] also indicated that head-on collisions involving large-truck lead to more severe injuries. Several studies suggested that the number of vehicles and the characteristics of large-truck also significantly affect the injury severity of crashes. Zheng et al. [17] and Chen F. & Chen S. [18] reported that large-trucks with weights over 20,000 lbs and hazardous material cargo increase the likelihood of severe injury, respectively. Moreover, Chen F. & Chen S. [18] and Islam et al. [8] discovered that there are significant differences between the single-vehicle and multiple-vehicle crashes involving large-trucks. Roads with a right curve also increase the odds of injury in large-truck crashes [6]. With regards to traffic conditions, Uddin & Huynh [4] reported that speed limits between 45 to 60 mph increase the odds of no injury under normal weather conditions, and speed limits over 65 mph increases the odds of major injury under rainy weather.

The effects of driver's age on the injury severity of crashes involving large-trucks have been inconsistent in the past studies. For example, Pahukula et al. [19] reported that young drivers are more likely to decrease the chances of no injury large truck crashes. On the other hand, Chen et al. [20] found that young drivers are more likely to be associated with incapable injuries and fatalities. A few studies have explored the impacts of different environmental factors on the injury severity of large-truck crashes. Naik et al. [21] have explored the weather impacts on the injury severity of large-truck-only crashes. The results of the study indicated that wind speed, rain, humidity, and air temperature have an association with the injury of single-vehicle large-truck crashes. Another study analyzed the injury severity of large-truck crashes under different lighting conditions on the rural and urban roadways of Ohio, USA [9]. The study concluded that the age and gender of occupants, types of trucks, speed, annual average daily traffic, curve roadways, and adverse weather have different effects on the injury severity under different lighting conditions. Concerning the effects of land use, Islam & Hernandez [11] indicated that crashes involving large-trucks in rural areas are likely to result in fatal injuries. In contrast, crashes in urban areas reduced the chances of incapacitating injuries. However, the categorization of crash locations into rural or urban put them in a broad definition of land use. The current study explored the more granular land use (e.g., population, housing, & employment density) and urban design (e.g., road network density) attributes, which rarely has been explored in previous studies.

2.2 The methodologies used for analysis of large-truck crashes

Until now, the majority of the studies have used statistical models to analyze crashes involving large-trucks. Dong et al. [3] used the negative binomial model to identify the factors that influence the frequency of large-truck crashes. Most of the studies that analyzed the injury severity of large-truck crashes have used the discrete-outcome models because typically the injury severity levels in crash data are reported as discrete values [22]. The commonly used discrete-outcome models for injury severity analysis are multinomial logit, ordered logit/probit, random parameters logit/probit model, and Bayesian binary logit model [4, 10, 19, 23]. However, the non-linear relationship between the contributory factors and injury severity makes the application of statistical models questionable. Moreover, several studies have reported that there are substantial correlations among the contributory factors. And, the effects of the contributory factors are difficult to estimate due to the correlations among the contributory factors [24]. Considering these limitations, a few studies have used the machine learning models such as classification and regression tree (CART) [25, 26], and gradient boosting decision tree [17]. However, a crash event occurs due to simultaneous or subsequent interactions of several contributory factors. The dependent and independent model structure only captures the effects of an individual contributory factor on the frequency or injury severity of crashes. The ARM technique not only can discover the set of crash contributory factors that often occur together in certain types of crashes but also identify the direction of associations among them. However, the application of the ARM technique for the analysis of road traffic crashes has been limited. Hong et al. [27] have used the ARM technique to analyze the risk factor of truck-involved crashes that occurred on the expressways of South Korea. Another study used the association rules to analyze the characteristics and contributory factors of work-zone crash casualties [28]. The current study used the ARM technique to discover the contributory factors that are likely to facilitate different levels of injury severity at the spatial concentration of crashes involving large-trucks. In addition to the commonly explored crash contributory factors, this study explored the granular level land use and urban design attributes. The following part of the study includes a description of the crash contributory factors, the methodologies, the results, the discussion, and the conclusion.

3. DATA DESCRIPTION

The final data set was aggregated using two different data sources. One is the crash data set, and another is the smart location data set. The crash data set include only road crashes involving the large-trucks that occurred between 2014 to 2019 in the state of Pennsylvania of the United States (US) [29] Each crash record included a wide variety of variables, which were distributed in multiple tables separately. The data tables are link-able using the crash record number (CRN). For this study, the crash, vehicle, roadway, and flag data tables were retrieved, and the commonly used contributory factors for the analysis of crashes involving large-trucks were selected.

First, the crash records that include only large-trucks were selected from the vehicle data table. The vehicle, crash, flag, and roadway data tables were merged using the CRN. There were multiple crash records with the same CRN because some crash events involved more than one injured person. We kept the crash record with the highest level of injury severity. The injury severity variable in the collected crash data has five categories. They are no injury (0), possible injury (4), suspected minor injury (3), suspected serious injury (2), and killed or fatal injury (1). The numbers in the brackets indicate the index used in the raw data set to label the categories of injury severity. The readers can find the definitions of different levels of injury severity in the following link [30]. According to the data dictionary for the crash data set of Pennsylvania, when there is no visible injury, and only complaints of pain, it is categorized as possible injury. To make to levels of injury severity more discrete and reduce the complexity of the association rules, we have put possible injury and suspected minor injury into one category as non-severe injuries. Moreover, since the proportion of killed or fatal injuries crashes is significantly low, we have put suspected serious injury and fatal injury into the category severe injuries. Such a strategy was used in some previous studies as well [4, 31]. The no injury category was left as it was in the raw data set. The crash records without the geographic coordinates, and with values such as "reported as unknown" or "unknown" were replaced with null values. All the records with null values were removed from the data set to avoid discrepancy. Then, the numerical variables were binned into intervals, and each interval was labeled based on the characteristics of the variables. The geographic coordinates, latitude, and longitude were left in their original format.

The smart location database is a publicly available data product and service provided by the United States Environmental Protection Agency. The population, household, employment, and road network density attributes were selected from the smart location database. These attributes were calculated for every census block group (CBG) of the USA. For details about the database, readers are requested to follow [32]. These attributes were binned into low, medium, and high-density intervals. The geographic coordinates, latitude, and longitude of the crash records, and the CBG's geographic area coordinates were used to identify which crash event belonged to which CBG. Then, each crash record was attributed with its corresponding CBG population, housing, employment, and road network density. The final data set included 39,464 crash records and 23 contributory factors. Table 1 describes the contributory factors.

Table 1. Distribution of the contributory factors

Contributory factors	Nominal Values	Frequency (%)
1. Collision type	Rear-end	10782 (27.32%)
	Angle	9515 (24.11%)
	Hit fixed object	6874 (17.42%)
	Sideswipe same direction	5491 (13.91%)
	No collision	3061 (7.76%)
	Others	1284 (3.25%)
	Head-on	1256 (3.18%)
	Sideswipe opposite direction	1201 (3.04%)
2. Vehicle movement	Going straight	21402 (54.23%)
	Negotiating curve	5382 (13.64%)
	Turning	3736 (9.47%)
	Stopped in traffic lane	2699 (6.84%)
	Changing lanes merging	2570 (6.51%)
	Others	2214 (5.61%)
	Slowing/stopping in lane	1461 (3.7%)
3. Drinking driver	No	38360 (97.2%)
	Yes	1104 (2.8%)
4. Aggressive driving	No	36924 (93.56%)
	Yes	2540 (6.44%)
5. Roadway alignment	Straight	32836 (83.2%)
	Curved	6628 (16.8%)
6. Roadway grade	Level	28747 (72.84%)
	Downhill	5669 (14.36%)
	Uphill	4040 (10.24%)
	Others	1008 (2.55%)
7. Intersect type	Mid-block	28122 (71.26%)
	Four-way intersect	5554 (14.07%)
	T-intersection	3656 (9.26%)
	Ramp	1185 (3%)
	Others (y-intersection, multi-leg intersection, roundabouts etc.)	947 (2.40%)
8. Interstate highway	No	28668 (72.64%)
	Yes	10796 (27.36%)
9. Signalized intersection	No	34655 (87.81%)
	Yes	4809 (12.19%)
10. Housing density	Low (<85th percentile)	38207 (96.81%)
	Medium (between 85th - 95th percentile)	934 (2.36%)
	High (>95th percentile)	323 (0.83%)
11. Population density	Low (<75th percentile)	37001 (93.76%)
	Medium (between 75th -95th percentile)	2159 (5.47%)
	High (>95th percentile)	304 (0.78%)

12. Employment density	Low (<85^{th} percentile)	33575 (85.08%)
	Medium (between 85^{th} -95^{th} percentile)	3819 (9.68%)
	High (95^{th} percentile)	2070 (5.24%)
13. Road network density	Medium (between 25^{th} - 75^{th} percentile)	18356 (46.51%)
	Low (<25^{th} percentile)	17561 (44.46%)
	High (>75^{th} percentile)	3547 (8.99%)
14. Location type	Rural	22244 (56.36%)
	Urban	17220 (43.63%)
15. Traffic control	No	29346 (74.36%)
	Yes	10118 (25.64%)
16. Speed limit	80 – 130 kmh	19923 (50.48%)
	40 – 80 kmh	18712 (47.42%)
	0 – 40 kmh	829 (2.1%)
17. Road surface condition	Dry	28557 (72.36%)
	Wet	6369 (16.14%)
	Snow	2184 (5.53%)
	Ice/frost	1412 (3.58%)
	Others	942 (2.39%)
18. Hour of day	11 am–4 pm	12953 (32.82%)
	6 -11 am	12836 (32.53%)
	0	7388 (18.72%)
	0	6287 (15.93%)
19. Weather	Clear	30417 (77.08%)
	Rain	4759 (12.06%)
	Snow	3379 (8.56%)
	Others	909 (2.30%)
20. Lighting	Daylight	28414 (72%)
	Dark-unlighted	6348 (16.09%)
	Dark-lighted	3180 (8.06%)
	Dawn	926 (2.35%)
	Dusk	423 (1.07%)
	Others	173 (0.44%)
21. Injury Severity	No injury	21938 (55.59%)
	Non-severe Injuries	15441 (39.12%)
	Severe Injuries	2085 (5.28%)
22. Latitude	Numerical	
23. Longitude	Numerical	

4. METHODOLOGIES

4.1 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Ester et al. [33] introduced the DBSCAN method to discover clusters in large spatial databases with noise. Most of the crash data include the geographic coordinates of the crash location. The DBSCAN clustering method can use the geographic coordinates to discover the regions with a high concentration of crashes. Also, roads are not always straight. There are intersections and curves in the road networks. Since DBSCAN can identify clusters with different shapes, it is suitable for application on road crash data. Multiple studies have applied DBSCAN clustering on road crash data [34, 35].

The DBSCAN algorithm has two major input parameters, which determine the density of a region. One is the Eps or ε neighborhood, which specifies the radius from a point to form the dense region. And, another is the MinPts, which specifies the number of points required to label an area around a point as a dense region. The DBSCAN clustering method uses these two input parameters to estimate the density around a point. It should be noted that for this study, a point indicates the geographic latitude and longitude coordinates of a crash event involving large-truck. Below, the cluster formation process of the DBSCAN algorithm is described in the context of road crash data.

Step 1: The algorithm selects an arbitrary point that has not been visited.

Step 2: Counts the number of crash events within the ε neighborhood of the point that was selected at step 1.

Step 3: If the number of crash events is equal to or more than the MinPts then the initially selected point is considered a core point. From the ε neighborhood of this core point, the cluster formation starts. On the other hand, if the number of crash events within the ε neighborhood of the point selected at step 1 is less than the MinPts then the algorithm moves to the closest point and repeats step 2. The algorithm may label the point selected at step 1 as noise or will make it a part of a cluster later, whose core point is within its ε neighborhood.

Step 4: The ε neighborhood of crash events or points within the initially identified cluster also become part of the cluster. If two core points are within each other's ε neighborhood, then the two clusters are joined.

Step 5: The process above continues for other point in the crash data, and ends when all the points in the crash data are visited.

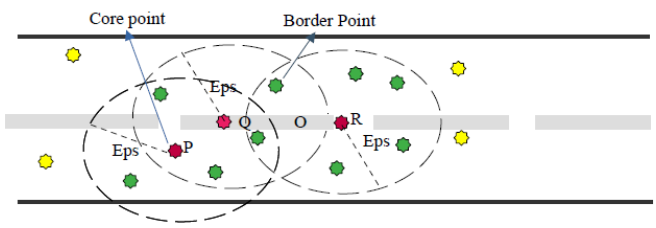

Figure 1. DBSCAN clustering

An illustration of the process of cluster formation by DBSCAN clustering on a road section would more helpful. Figure 1 above illustrates a segment of road, which is indicated by the space between the parallel black lines. The rectangular light gray shapes are markers like in the roads. For the figure above, let's assume the MinPts = 6. In the figure above, the red stars are core points (i.e. P,Q & R), since there are six crash events within the ε neighborhood of those red stars. The dashed line from the core points to the border of circle indicates the Eps. The green points are called border point because they are within the ε neighborhood of the core points. One of the features of DBSCAN clustering is that a point is also part of a cluster if it close to many points from that cluster. Therefore, the three circles in the figure together indicate a cluster. The yellow stars are noise because they are not within the ε neighborhood of core points (P,Q & R).

4.2 Association Rule Mining (ARM)

The ARM technique is based on the market basket data analysis. It was first proposed by Agrawal et al. [36] to identify the frequent sets of items bought together in a transaction data of a supermarket. The technique is simple, and the association rules discovered by the technique are easy to explain. Let, D = {T1, T2, T3, ……., Tn} be a set of transactions, and I = {I1, I2, I3, ……..., In} be a set of items. Assume A and B are two items. A rule discovered by the association rule analysis for the aforementioned items is shown as "A⟶B", where A is the antecedent, and B is the consequent. A rule can have multiple items as antecedent and consequent. In the context of market basket data analysis, the rule means if a customer buys A in a transaction, then the customer is more likely to buy B as well in the same transaction. For this study, a crash record was treated as a transaction in the market basket data. The nominal values of the crash contributory factors were treated as items of a transaction like in the market basket data. In this study, we have employed the commonly used Apriori algorithm because it is simple, and easy to explain, unlike other parametric and non-parametric methods. Generally, the apriori algorithm requires two steps. First, it iteratively scans the whole database for the frequent itemsets. In the second step, it generates association rules from the frequent itemsets. Support, confidence, and lift are the three important indicators of strong association for the rules generated by the apriori algorithm.

4.2.1 Support

The support indicates the proportion of the considered item in the data set. Support for a rule like A→B means the proportion of data records where A and B occur together. Generally, it is determined based on the characteristics of the data, and domain knowledge. The support is calculated using the following equation. Here, N is the total number of crash records. The support for a rule such as A⟶B is the same as B⟶A. Therefore, we need another measure, which considers the direction of the association.

$$S = P\,(A \cap B)/N \tag{1}$$

4.2.2 Confidence

Confidence is a measure that takes into account the direction of a rule and helps to differentiate between the rules such as A⟶B and B⟶A. The confidence of rule A⟶B is defined by the ratio of occurrence of A and B together to the occurrence of A only. The confidence value of rule A⟶B indicates that the chance of occurrence of B increases with the occurrence of A. The confidence value is calculated using Eq. (2). However, the support and confidence alone cannot explain the significance of a rule.

$$C = P\,(A \cap B)/P\,(A) \tag{2}$$

4.2.2 Lift

The lift (L) measures the strength of an association rule. The lift is the ratio between the confidence and the expected confidence of an association rule. The occurrence of A and B together with the occurrence of B is considered the expected confidence. The lift value ranges from 0 to ∞. The chances of A and B occurring together is more than expected when the rule A⟶B has a lift greater than 1. A lift lower than 1 indicates the opposite. There is no association between the items when the lift value is equal to 1. The lift is calculated using Eq. (3).

$$L = P\,(A \cap B)/P\,(A) \times P(B) \tag{3}$$

Unfortunately, there is no standard rule to determine these parameters. The minimum support value ranged from 10 to 40 percent in some studies that used the ARM technique for road crash analysis [37-39]. However, using such high minimum support will leave out interesting rules which may include rare but important crash items. In some studies, the minimum support ranged from 0 to 5 percent [40-42]. It is fair to say that the criteria to set the minimum support value is subjective. It should be noted that the nominal values of the crash contributory factors are mentioned as the crash items in this study.

5. RESULTS

The current work was set in motion to discover the set of crash items that influence no injury, non-severe, and severe injuries at spatial concentrations of crashes involving large-trucks. Figure 2 illustrates the whole process of analysis of this study.

5.1 Spatial concentrations of crashes involving large-trucks

There is no standard rule or criteria to determine the Eps and the MinPts of the DBSCAN clustering method. The value of these parameters depends on the domain and expertise of the user. Selecting a MinPts value of 2 or 3 will lead to the formation of spatial concentrations of very small sizes. On the other hand, the total length of public roads was 120,590 miles (194,149,900 meters) in Pennsylvania until January 2020 [43]. And, the total number of crashes involving large-trucks was 7038 in 2019. Therefore, the approximate rate of crashes involving large-truck per 100 meters of the public road is 0.000036. Though this is a simple way to estimate the distribution of crashes involving large-trucks across the whole state of Pennsylvania, it can aid in estimating the MinPts for the current analysis. Selecting a MinPts that is significantly higher than the rate of crashes involving large-trucks per 100 meters is likely to identify very few numbers of spatial concentrations because in some places the crashes are distributed across much wider areas. Also, we observed that selecting a MinPts value of 7 or 8, identified fewer spatial concentrations where the proportion of severe injuries was over the threshold. Moreover, if the value of Eps is high such as 200 or 300 meters, then the corresponding road and environment characteristics of the crashes in the spatial concentrations may vary too much. Considering the aforementioned facts and issues, the MinPts and Eps were set to 6 and 100 meters, respectively. The DBSCAN method identified 362 spatial concentrations. It should be noted that the distance between the crash locations was calculated using the widely used Euclidean distance.

For practical purposes, considering the temporal characteristics of spatial concentrations is important. In a period of six years, a few crashes may occur at certain locations for random reasons. To account for the temporal stability of the spatial concentrations, the six years were converted into three periods. The spatial concentrations that included crashes from the three periods were considered. After removing the spatial concentrations with low temporal stability, the number of spatial concentrations was 289.

Moreover, the proportion of severe injuries crashes is comparatively low in the crash data of large-trucks, which obscures the overall analysis process. To obtain a more balanced data set, the spatial concentrations were categorized into three groups based on the proportion of severe injuries crashes. The group-1 included the spatial concentrations, where the proportion of severe injuries crashes was less than 5 percent. The group-2 included spatial concentrations with more than or equal to 5 percent but less than 15 percent of severe injuries crashes. The spatial concentrations that involved more than 15 percent of severe injury crashes were put in group-3. The following Table 2 describes the distribution of different levels of injury severity in the different groups of spatial concentrations.

Table 2. Description of the spatial concentrations

Category of Spatial Concentrations	Proportion of Severe Injuries Crashes	Number of Crashes
Group-1	0.05%	1968
Group-2	10.06%	527
Group-3	20.09%	219

5.2 Data preparation for ARM

The data sets need to be transformed into a set of transactions that are similar to the market basket data before applying the Apriori algorithm. The current study has used the python library called “mlxtend”, which offers the Apriori algorithm along with the tools to prepare the data set for the application of the ARM technique. After transforming the data sets, the dimensions of the group-1, group-2, and group-3 data set changed to (1968, 75), (527, 73), and (219, 74), respectively. Table 3 shows a portion of the group-3 data set after transformation for the readers.

Table 3. An example of data set prepared for ARM

Collision type-Sideswipe Same Direction	Hour of day – 10 PM -6 AM	Roadway alignment - straight	Vehicle movement – Going Straight	Maximum Severity - Severe
False	True	True	False	False
False	False	False	False	False
False	False	True	True	True
False	False	True	False	True
True	True	True	False	False

5.3 Rule mining

The apriori algorithm requires three input parameters, which are the minimum support, the maximum length for the set of items, and a metric to rank the rules. After reviewing the data distribution and past studies, the minimum support (S) was set to 1.50%. Also, setting the S value too high will leave out interesting rules including crash items such as the severe injury and drinking driver. Ordonez et al. [44] suggested some constraints to discover important rules and leave out redundant rules. The proposed constraints include maximum association size, an attribute grouping constraint, and antecedent/consequent rule filtering constraint. In this study, we have limited the length of association size to 4. Since injury severity is the consequence of a crash, we have considered only the rules where the consequent is any level of the injury severity (e.g. no injury, non-severe injuries, severe injuries). The rules were ranked according to their lift value, which was set to 1.2 in this study. Under the aforementioned constraints, the number of rules generated for the group-1, group-2, and group-3 spatial concentrations of crashes involving large-trucks are 1733, 4051, and 4343, respectively. The following sections discuss the top five rules with consequent as severe injuries, non-severe injuries, and no injury for each group of spatial concentrations.

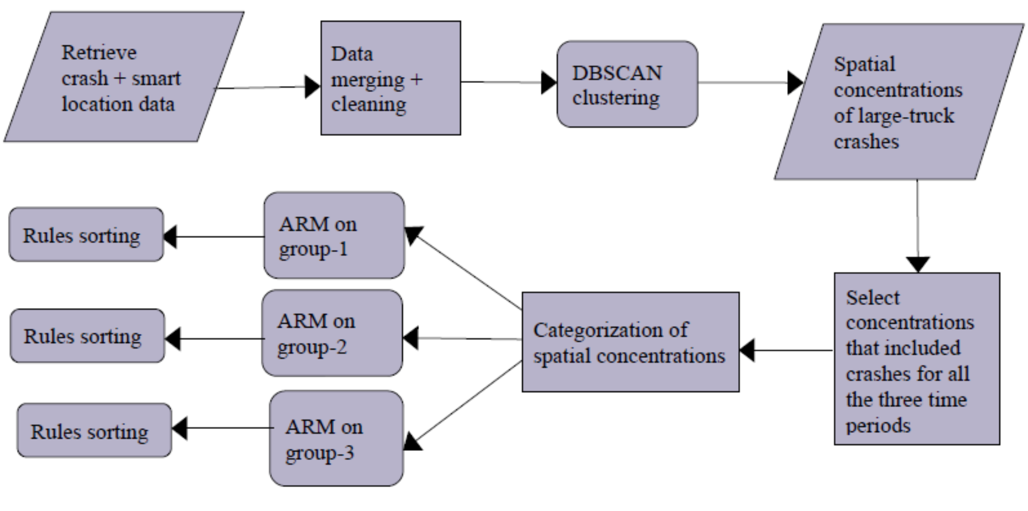

Figure 2. Process flowchart

Table 4. Rules with consequent as the severe injuries

Spatial cluster	Index	Antecedents of Severe Injuries	Support	Confidence	Lift
Group-2	1	{lighting – dark-lighted, roadway alignment – straight, speed limit – 40 to 80 km/h}	1.52%	36.36%	3.62
	2	{lighting - dark-lighted, road condition -dry, speed limit – 40 to 80 km/h,}	1.52%	36.36%	3.62
	3	{lighting – dark-lighted, road condition – dry, interstate - no}	1.71%	34.62%	3.44
	4	{lighting – dark-lighted, employment density – low, speed limit – 40 to 80 km/h}	1.71%	32.14%	3.2
	5	{lighting – dark-lighted, roadway grade – level, speed limit – 40 to 80 km/h}	1.52%	32.00%	3.18
Group-3	6	{drinking driver – yes, road network density – medium, collision type - rear-end}	1.83%	100.00%	4.98
	7	{intersect type – others, roadway alignment – straight, roadway grade - level}	2.28%	100.00%	4.98
	8	{intersect type – others, signalized intersection – no, roadway alignment - straight}	1.83%	100.00%	4.98
	9	{intersect type – others, roadway grade – level, speed limit – 40 to 80 km/h}	1.83%	100.00%	4.98
	10	{intersect type – others, roadway alignment – straight}	2.28%	83.33%	4.15

Table 5. Rules with consequent as the non-severe injuries

Spatial cluster	No	Antecedents of Non-Severe Injuries Crashes	Support	Confidence	Lift
Group-1	11	{collision type – rear-end, lighting – dark-lighted, traffic control - absent}	1.63%	68.09%	1.66
	12	{collision type – rear-end, road network density – high, vehicle movement - straight}	2.44%	66.67%	1.63
	13	{collision type – rear-end, vehicle movement – straight, hour of day – 16 to 22 Pm}	1.52%	66.67%	1.63
	14	{collision type – rear-end, employment density – high, vehicle movement - straight}	1.88%	66.07%	1.62
	15	{collision type – rear-end, lighting – dark-lighted, weather - clear}	1.63%	65.31%	1.6
Group-2	16	{vehicle movement – straight, population density – high, interstate - no}	1.90%	100.00%	2.74
	17	{vehicle movement – straight, population density – medium, speed limit – 40 to 80 km/h}	1.71%	100.00%	2.74
	18	{vehicle movement – straight, road network density – high, interstate - no}	1.90%	90.91%	2.5
	19	{vehicle movement – straight, road network density – high, speed limit – 40 to 80 km/h}	1.71%	90.00%	2.47
	20	{intersect type – four-way, road network density - high}	1.52%	88.89%	2.44
Group-3	21	{employment density – medium, road condition - wet}	1.83%	100.00%	2.61
	22	{traffic control – absent, intersect type - four-way}	1.83%	100.00%	2.61
	23	{road network density – low, vehicle movement – slowing/stopping in lane}	1.83%	100.00%	2.61
	24	{employment density – medium, aggressive driving – no, road condition - wet}	1.83%	100.00%	2.61
	25	{vehicle movement – slowing/stopping in lane, aggressive driving – no, hour of day – 6 to 11 AM}	2.28%	100.00%	2.61

Table 6. Rules with consequent as the no injury

Spatial cluster	Index	Antecedents of No injury Crashes	Support	Confidence	Lift
Group-1	26	{vehicle movement – straight, speed limit – 40 to 80 km/h, collision type – hitting fixed object}	3.25%	94.12%	1.59
	27	{speed limit – 40 to 80 km/h, collision type – hitting fixed object, roadway grade - level}	2.85%	93.33%	1.58
	28	{vehicle movement – turning, road network density – low, signalized intersection - yes}	1.98%	92.86%	1.57
	29	{roadway alignment – straight, speed limit – 40 to 80 km/h, collision type – hitting fixed object}	3.76%	92.50%	1.57
	30	{speed limit – 40 to 80 km/h, employment density – low, collision type – hitting fixed object}	2.49%	92.45%	1.57
Group-2	31	{collision type – sideswipe same direction, intersect type -ramp}	1.52%	100.00%	1.87
	32	{road network density – medium, road condition – snow}	1.90%	100.00%	1.87
	33	{collision type – sideswipe same direction, aggressive driving – no, intersect type - ramp}	1.52%	100.00%	1.87
	34	{road network density – medium, aggressive driving – no, road condition - snow}	1.52%	100.00%	1.87
	35	{collision type – rear-end, weather – clear, road condition - wet}	1.52%	100.00%	1.87
Group-3	36	{roadway alignment – straight, employment density – medium, collision type - angle}	1.83%	100.00%	2.41
	37	{employment density – medium, roadway grade – level, collision type - angle}	1.83%	100.00%	2.41
	38	{employment density – medium, weather – clear, collision type – angle}	1.83%	100.00%	2.41
	39	{collision type -sideswipe same direction, employment density – low, hour of day – 6 to 11 AM}	1.83%	100.00%	2.41
	40	{collision type – sideswipe same direction, interstate – yes, employment density - low}	4.11%	100.00%	2.41

5.3.1 Severe injuries

In the Table 4, rules #1-5 and #6-10 show the top five rules with consequent as the severe injuries for group-2 and group-3 spatial concentrations, respectively. Since the proportion of severe injuries is less than the minimum support in group-1, there were no rules with consequent as severe injuries in crashes of group-1. The confidence and lift values for rules#1-5 were comparatively lower than that of rules#6-10. Rule#6 indicated that the chances of severe injuries are 100% for the rear-end collisions involving drunk drivers in the medium road network density areas. These assumptions reinforced the findings of previous studies where rear-end collisions [8], and drinking & driving [25] led to severe injuries in most of the cases. The set of antecedents in rules #7-9 also have a 100% probability of leading to severe injuries. In the set of those antecedents, intersect type – others (i.e. y-intersection, multi-leg intersections, roundabout, etc.) is a frequent crash item. In addition, rule#1, #2, #4, and #5 indicated that large-truck crashes under dark-lighted conditions and on roads with a speed limit between 40 to 80 km/h have moderate chances of leading to severe injuries. Along with the aforementioned crash items, dry roads, non-interstate roads, roadways with straight alignment and level grade lead to severe injuries in some cases. Some previous studies also reported that large-truck crashes that occur under dark-lighted conditions [5] and on dry road surfaces [17] have some probability of leading to severe injuries. Areas with low employment density and medium road network density may facilitate severe injuries in some cases. In low employment density areas, due to less traffic congestion, drivers are likely to drive at a higher speed, which has the possibility of leading to severe injuries.

5.3.2 Non-severe injuries

A variety of crash items and their combinations have the probability of leading to non-severe injuries. In Table 5, rules#11-15, #16-20, and #21-25 shows the top five rules with non-severe injuries as the consequent of the crashes in the group-1, group-2, and group-3 spatial concentrations, respectively. The confidence and lift values for the rules#11-15 are significantly lower than that of rules#16-25. According to rules#11 and #15, rear-end collisions involving large-trucks under dark-lighted conditions in absence of traffic control or clear weather have significant chances of leading to non-severe injuries. Rules#12 and #14 indicated that large-trucks going straight before rear-end collision in areas with high road network or employment density have more than 66% chance of leading to non-severe injuries. Rear-end collisions involving large-trucks that were going straight during 16 to 22 PM also have significant chances of resulting in non-severe injuries. The previous studies have reported inconsistent effects of rear-end collisions on the injury severity of crashes involving large-trucks. Uddin & Huynh [4] indicated that rear-end collision decreases the likelihood of major injuries (fatal and disabling) under normal weather conditions. On the other hand, Chen & Zhang [45] found that rear-end collisions involving large-trucks are associated with severe injuries.

Table 5 indicates that rules#16 and #15 have the highest confidence and lift values. These rules indicated that large-trucks going straight prior to the collision in areas with medium or high population density on non-interstate roads or on roads with a speed limit between 40 to 80 km/h lead to non-severe injuries in almost all cases. Other crash items that are hugely frequent in non-severe injuries involving large-trucks are high road network density, and four-way intersects. Behnood & Mannering [22] and Islam & Hernandez [23] also reported that large-trucks going straight prior to the collision are inclined to non-severe injuries.

The confidence and lift values of rules#21-25 are equal. The set of crash items in those rules has a 100% probability of leading to non-severe injuries. The frequent crash items in those rules are medium employment density, wet roads, non-aggressive driving, and vehicles slowing/stopping in lanes. Other less frequent crash items are the absence of traffic control, four-way intersect, low road network density, and hours from 6 to 11 AM.

5.3.3 No injury

In Table 5, the rules#26-30, #31-35, and #35-40 show the set of antecedents in rules with consequent as no injury in crashes involving large-trucks of group-1, group-2, and group-3 spatial clusters, respectively. The confidence values of rules#26-30 ranged from 92 to 95%. Rules#26, 27, 29, and 30 indicated that collisions between large-trucks and fixed objects on roads with a speed limit between 40 to 80 km/h have significant chances of leading to no injury crashes. Also, large-trucks that were turning before the collision at signalized intersections in areas with low road network density increase the likelihood of no injury crashes. Behnood & Mannering [22] also reported that collisions with fixed objects by large-trucks increase the likelihood of no injury.

The confidence values of rules#31-40 were equal. But the lift values of rules#36-40 were significantly higher than the other rules in Table 6. According to rules#31 and #33, the probability of no injury is 100% for sideswipe same direction collision involving large-trucks on ramp intersect. On the other hand, rules#32, and #34 indicated that large-truck crashes on snowy roads in areas with medium road network density also have a 100% probability of leading to no injury. During snowy weather condition, drivers become more aware and drive carefully, which probably reduce the chances of injury in crashes. Also, non-aggressive driving is more likely to lead to no injury crashes. Rule#35 indicated that rear-end collisions involving large-trucks on wet roads in clear weather rarely lead to injuries.

Rules#36, #37, and #38 indicated that angle collisions involving large-trucks in areas with medium employment density on roadways with straight alignment or level grade or in clear weather have almost zero percent chance of leading to any type of injury. According to rules#39 and 40, the chances of no injury increase for sideswipe same direction collisions involving large-trucks in areas with low employment density during 6 to 11 AM or on interstate roads.

6. DISCUSSION AND CONCLUSIONS

The current study employed data mining methods to discover the associations between the contributory factors and the different levels of injury severity at the spatial concentrations of crashes involving large-trucks. This study explored the crash characteristics, driver's actions, road geometries, traffic conditions, and environmental factors. Additionally, the newly explored attributes included land use (e.g. housing, population, and employment density), and urban design (e.g. road network density). The application of the DBSCAN clustering method identified the spatial concentrations of large-truck crashes. The spatial concentrations that did not include crashes from the three periods were removed. Then, the remaining spatial concentrations were divided into three groups based on the proportion of severe injury crashes in each spatial concentration to obtain data sets that are less biased towards the no injury category. The apriori algorithm, which is a popular association rule mining technique was employed on each group of spatial concentrations to discover the conditions that lead to the no injury, non-severe injuries, and severe injuries in crashes involving large-trucks.

The proposed framework has significant practical implications. The crash data are generally recorded by police officers, who are not specialized in road crash analysis. Moreover, it is not possible to inspect all the locations of crashes due to limited resources. Investigation of the spatial concentrations of crashes involving large-trucks by road safety experts may reveal more relevant crash contributory factors.

Further, the application of the ARM technique discovered different sets of crash items that are responsible for the different levels of injury severity in crashes involving large trucks. For example, the combination of crash items that significantly increases the chances of severe injuries is the rear-end collision and drinking & driving. Checking the alcohol intoxication level of the drivers passing through those locations can reduce the number and severity of the aforementioned crashes. Additionally, the findings suggested that severe injuries in crashes involving large-trucks are more likely to occur on roads with a speed limit between 40 to 80 km/h, in areas with low employment density and medium road network density. Road safety authorities can improve road conditions, raise traffic control, and put up warnings such as "crash-prone locations" to curve out severe injuries crashes at those locations.

The rear-end collisions and medium or high population density both are frequent conditions of non-severe injuries. In medium or high population areas, traffic congestion is frequent. Such conditions may facilitate rear-end collisions since a lot of vehicles travel in close proximity during congestion. Traffic authorities can put up signs and warning for drivers to avoid driving too close to other vehicles.

Several sets of antecedents can lead to no injury every time they appear together. According to the results, collisions between large-trucks and fixed objects, sideswipe same direction collision, medium road network or employment density, snowy roads, and clear weather have a significant chance of leading to no injury crashes involving large-trucks. Reinforcing the current traffic controls and signals vigorously may reduce the number of no injury crashes involving large trucks.

In summary, the current work was an interesting effort to obtain important and hidden insights about the crashes involving large-trucks. Future studies may apply another clustering algorithm to segment the spatial concentrations into more homogeneous groups. Also, future studies may increase the maximum number of crash items in the rule, which can reveal more insights about the conditions that lead to different levels of injury in crashes involving large-trucks.

ACKNOWLEDGMENT

This work is grateful to the Department of Transportation of Pennsylvania and the US Environmental Protection Agency for making the crash data and smart location data available for research purposes, respectively.

REFERENCES

[1] Freight Facts and Figures. (2017). Bureau of Transportation Statistics. U.S. Department of Transportation.

[2] Dong, C., Nambisan, S.S., Richards, S.H., Ma, Z. (2015). Assessment of the effects of highway geometric design features on the frequency of truck involved crashes using bivariate regression. Transportation Research Part A: Policy and Practice, 75: 30-41. https://doi.org/10.1016/j.tra.2015.03.007

[3] Dong, C., Dong, Q., Huang, B., Hu, W., Nambisan, S.S. (2017). Estimating factors contributing to frequency and severity of large truck–involved crashes. Journal of Transportation Engineering, Part A: Systems, 143(8). https://doi.org/10.1061/JTEPBS.0000060

[4] Uddin, M., Huynh, N. (2020). Injury severity analysis of truck-involved crashes under different weather conditions. Accident Analysis and Prevention, 141: 105529. https://doi.org/10.1016/j.aap.2020.105529

[5] Zhu, X., Srinivasan, S. (2011). A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accident Analysis and Prevention, 43(1): 49-57. https://doi.org/10.1016/j.aap.2010.07.007

[6] Azimi, G., Rahimi, A., Asgari, H., Jin, X. (2020). Severity analysis for large truck rollover crashes using a random parameter ordered logit model. Accident Analysis and Prevention, 135: 105355. https://doi.org/10.1016/j.aap.2019.105355

[7] Cantor, D.E., Corsi, T.M., Grimm, C.M., Özpolat, K. (2010). A driver focused truck crash prediction model. Transportation Research Part E: Logistics and Transportation Review, 46(5): 683-692. https://doi.org/10.1016/j.tre.2009.08.011

[8] Islam, S., Jones, S.L., Dye, D. (2014). Comprehensive analysis of single- and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. Accident Analysis and Prevention, 67: 148-158. https://doi.org/10.1016/j.aap.2014.02.014

[9] Uddin, M., Huynh, N. (2017). Truck-involved crashes injury severity analysis for different lighting conditions on rural and urban roadways. Accident Analysis and Prevention, 108): 44-55. https://doi.org/10.1016/j.aap.2017.08.009

[10] Ahmed, M.M., Franke, R., Ksaibati, K., Shinstine, D.S. (2018). Effects of truck traffic on crash injury severity on rural highways in Wyoming using Bayesian binary logit models. Accident Analysis and Prevention, 117: 106-113. https://doi.org/10.1016/j.aap.2018.04.011

[11] Islam, M.B., Hernandez, S. (2013). Modeling injury outcomes of crashes involving heavy vehicles on Texas highways. Transportation Research Record: Journal of the Transportation Research Board, 2388(1): 28-36. https://doi.org/10.3141%2F2388-05

[12] Wang, Y., Luo, Y., Chen, F. (2019). Interpreting risk factors for truck crash severity on mountainous freeways in Jiangxi and Shaanxi, China. European Transport Research Review, 11: 3-5. https://doi.org/10.1186/s12544-019-0366-4

[13] Zhang, Z., Ming, Y., Song, G. (2019). Identify road clusters with high-frequency crashes using spatial data mining approach. Applied Sciences, 9(24): 5282. https://doi.org/10.3390/app9245282

[14] Jia, R., Khadka, A., Kim, I. (2018). Traffic crash analysis with point-of-interest spatial clustering. Accident Analysis and Prevention, 121: 223-230. https://doi.org/10.1016/j.aap.2018.09.018

[15] Mohaymany, A.S., Shahri, M., Mirbagheri, B. (2013). GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. Geo-Spatial Information Science, 16(2): 113-119. https://doi.org/10.1080/10095020.2013.766396

[16] Xie, Z., Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: An integrated approach. Journal of Transport Geography, 31: 64-71.

https://doi.org/10.1016/j.jtrangeo.2013.05.009
[17] Zheng, Z., Lu, P., Lantz, B. (2018). Commercial truck crash injury severity analysis using gradient boosting data mining model. Journal of Safety Research, 65: 115-124. https://doi.org/10.1016/j.jsr.2018.03.002
[18] Chen, F., Chen, S. (2011). Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis and Prevention, 43(5): 1677-1688. https://doi.org/10.1016/j.aap.2011.03.026
[19] Pahukula, J., Hernandez, S., Unnikrishnan, A. (2015). A time of day analysis of crashes involving large trucks in urban areas. Accident Analysis & Prevention, 75: 155-163. https://doi.org/10.1016/j.aap.2014.11.021
[20] Chen, C., Zhang, G., Tian, Z., Bogus, S.M., Yang, Y. (2015). Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations. Accident Analysis and Prevention, 85: 186-198. https://doi.org/10.1016/j.aap.2015.09.005
[21] Naik, B., Tung, L.W., Zhao, S., Khattak, A.J. (2016). Weather impacts on single-vehicle truck crash injury severity. Journal of Safety Research, 58: 57-65. https://doi.org/10.1016/j.jsr.2016.06.005
[22] Behnood, A., Mannering, F. (2019). Time-of-day variations and temporal instability of factors affecting injury severities in large-truck crashes. Analytic Methods in Accident Research, 23: 100-102. https://doi.org/10.1016/j.amar.2019.100102
[23] Islam, M., Hernandez, S. (2013). Large truck-involved crashes: Exploratory injury severity analysis. Journal of Transportation Engineering, 139(6): 596-604. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000539
[24] Greibe, P. (2003). Accident prediction models for urban roads. Accident Analysis & Prevention, 35(2): 273-285. https://doi.org/10.1016/S0001-4575(02)00005-2
[25] Chang, L.Y., Chien, J.T. (2013). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Safety Science, 51(1): 17-22. https://doi.org/10.1016/j.ssci.2012.06.017
[26] Eustace, D., Alqahtani, T., Hovey, P.W. (2018). Classification tree modeling of factors impacting severity of truck-related crashes in Ohio. Transportation Research Board 97th Annual Meeting, Washington, United States.
[27] Hong, J., Tamakloe, R., Park, D. (2020). Discovering Insightful Rules among Truck Crash Characteristics using Apriori Algorithm. Journal of Advanced Transportation, 2020. https://doi.org/10.1155/2020/4323816
[28] Weng, J., Zhu, J.Z., Yan, X., Liu, Z. (2016). Investigation of work zone crash casualty patterns using association rules. Accident Analysis and Prevention, 92: 43-52. https://doi.org/10.1016/j.aap.2016.03.017
[29] PennDOT Crash Download Map. https://pennshare.maps.arcgis.com/apps/webappviewer/index.html?id=8fdbf046e36e41649bbfd9d7dd7c7e7e, accessed on July 21, 2021.
[30] PennDOT Open Data Portal Cash Data Dictionary and Field Constraints Tables. (2019). https://data-pennshare.opendata.arcgis.com/, accessed on July 21, 2021.
[31] Sun, M., Sun, X., Shan, D. (2019). Pedestrian crash analysis with latent class clustering method. Accident Analysis and Prevention, 124: 50-57. https://doi.org/10.1016/j.aap.2018.12.016
[32] Smart Location Database. (2021). Environmental Protection Agency.
[33] Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. KDD-96 Proceedings, pp. 226-231.
[34] Chantamit-o-pass, P., Pongpum, W., Kongsaksri, K. (2020). Road traffic injury prevention using DBSCAN algorithm. Neural Information Processing. ICONIP 2020. Communications in Computer and Information Science, vol 1333. Springer, Cham. https://doi.org/10.1007/978-3-030-63823-8_22
[35] Qiu, C., Xu, H., Bao, Y. (2016). Modified-DBSCAN clustering for identifying traffic accident prone locations. International Conference on Intelligent Data Engineering and Automated Learning, pp. 99-105. https://doi.org/10.1007/978-3-319-46257-8_11
[36] Agrawal, R., Imieliński, T., Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2): 207-16. https://doi.org/10.1145/170035.170072
[37] Feng, M., Zheng, J., Ren, J., Xi, Y. (2020). Association rule mining for road traffic accident analysis: A case study from UK. In: Ren J. et al. (eds) Advances in Brain Inspired Cognitive Systems. BICS 2019. Lecture Notes in Computer Science, vol 11691. Springer, Cham. https://doi.org/10.1007/978-3-030-39431-8_50
[38] Gupta, M., Kumar Solanki, V., Kumar Singh, V. (2017). A novel framework to use association rule mining for classification of traffic accident severity. Ingeniería Solidaria, 13(21): 37-44. http://dx.doi.org/10.16925/in.v13i21.1726
[39] Kumar, S., Toshniwal, D. (2015). A data mining framework to analyze road accident data. Journal of Big Data, 2(1): 26. https://doi.org/10.1186/s40537-015-0035-y
[40] Geurts, K., Thomas, I., Wets, G. (2005). Understanding spatial concentrations of road accidents using frequent item sets. Accident Analysis and Prevention, 37(4): 787-99. https://doi.org/10.1016/j.aap.2005.03.023
[41] Montella, A. (2011). Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accident Analysis & Prevention, 43(4): 1451-63. https://doi.org/10.1016/j.aap.2011.02.023
[42] Montella, A., Aria, M., D'Ambrosio, 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. https://doi.org/10.1016/j.aap.2011.04.025
[43] Pennsylvania Transportation by the Numbers. (2020). Bureau of Transportation Statistics, U.S. Department of Transportation.
[44] Ordonez, C., Ezquerra, N., Santana, C.A. (2006). Constraining and summarizing association rules in medical data. Knowledge and Information Systems, 9(3): 1-2. https://doi.org/10.1007/s10115-005-0226-5
[45] Chen, C., Zhang, J. (2016). Exploring background risk factors for fatigue crashes involving truck drivers on regional roadway networks: A case control study in Jiangxi and Shaanxi, China. SpringerPlus, 5: 582. https://doi.org/10.1186/s40064-016-2261-y