

## Hit and run crash analysis using association rules mining

Subasish Das, Xiaoqiang Kong & Ioannis Tsapakis

To cite this article: Subasish Das, Xiaoqiang Kong & Ioannis Tsapakis (2019): Hit and run crash analysis using association rules mining, Journal of Transportation Safety & Security, DOI: [10.1080/19439962.2019.1611682](https://doi.org/10.1080/19439962.2019.1611682)

To link to this article: <https://doi.org/10.1080/19439962.2019.1611682>



Published online: 23 May 2019.




Submit your article to this journal [↗](#)



View Crossmark data [↗](#)



## Hit and run crash analysis using association rules mining

Subasish Das , Xiaoqiang Kong, and Ioannis Tsapakis

Texas A&M Transportation Institute, Texas A&M University System, College Station, Texas, USA

### ABSTRACT



Hit-and-run crashes have drawn growing attention because of the severe consequence of the delaying emergency assistance to victims. However, the number of related studies is still limited due to the lack of relevant adequate data. The objectives of this research are to, (1) identify the crash and geometric features, which contribute to hit-and-run crashes, (2) discover how those measures change in the case of the segment and intersection-related crashes. The study applied market basket analysis to mine associations between the crash and geometric features of hit-and-run crashes. Based on the generated rules, the results show single-vehicle crashes are the first common factor of hit-and-run crashes and dark lighting is the second factor. The combination of these two factors was found to clearly associate with more severe crashes. The study also found hit-and-run crashes mostly occurred in urban areas. The rules also show segment-related crashes have higher fatality rates than intersection-related crashes. These findings suggest that improvements such as roadway markings, lighting, and installation of cameras at intersections could help to reduce hit-and-run crashes or detect the hit and run offenders.

### KEYWORDS

hit-and-run crashes; data mining; market basket analysis; rules mining

## 1. Introduction

*Hit-and-run* (H&R) crashes refer to crashes where the at-fault (responsible for crash occurrence) drivers leave the crash location without helping victims or reporting the occurrence of crashes to relevant authorities. These crashes could significantly increase the probability of severe injuries or fatalities, particularly for vulnerable roadway users like pedestrians and bicyclists, due to delays in emergency assistance. Although H&R is a severe crime according to law enforcement agencies, and serious criminal charges befall at-fault drivers if caught later, the frequency of H&R crash occurrences is still high. According to the National Highway Traffic Safety Administration (NHTSA 2016), an estimated 737,100 H&R crashes

**CONTACT** Subasish Das  [s-das@tti.tamu.edu](mailto:s-das@tti.tamu.edu)  Texas A&M Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX 77843.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/utss](http://www.tandfonline.com/utss).

© 2019 Taylor & Francis Group, LLC and The University of Tennessee

happened in the U.S. in 2015, which indicates that a H&R crash happens somewhere in the U.S. in every 43 seconds. The fatalities that resulted from H&R crashes in 2016 added up to 2,049, the highest number ever recorded in a year (Benson, Arnold, Tefft, & Horrey, 2018). In Louisiana, around 109,000 H&R crashes (12% of all crashes) happened during 2010–2015. To put it in context, there are approximately 50 H&R crashes occur on Louisiana roadways every day. The number of fatal crashes represents 5% of all fatal crashes in Louisiana. In 2015, the number of H&R crashes in this state increased by approximately 10% from 2010. This sharp rise calls for a rigorous analysis of H&R crashes and identification of appropriate strategies to address this issue.

There is a limited number of studies on H&R crashes. Most of the studies attempted to identify the driver's traits in leaving a H&R crash scene. In many cases, there is missing information regarding the offender and associated traits. There is a need to study crash and geometric characteristics of H&R crashes. Moreover, Louisiana Crash Data Reports (2019) shows that the count of H&R fatal crashes on segments is approximately three times the H&R crashes at intersections. This study aims to answer two research questions: (1) what crash and geometric measures contribute to H&R crashes, and (2) how these measures change in segment and intersection-related crashes. The answers are important in determining appropriate countermeasures to reduce the number of H&R crashes. This study applied market basket analysis, also known as association rules mining, to a 6-year (2010–2015) H&R crash data set in Louisiana to identify key rules regarding the crash, geometric and environmental characteristics of H&R crashes. The reason for using association rules is related to complex interactions among contributing factors associated with H&R crashes. The interactions in the form of rules can provide valuable insights for the occurrences of H&R crashes. The objectives of this research are to (1) identify the crash and geometric features, which contribute to H&R crashes and (2) discover how those measures change in the case of the segment and intersection-related crashes. The outcomes from the association rules showed that this data mining tool is well suited to describe and improve decision making for addressing H&R crashes.

## 2. Literature review

Crash frequency analysis and crash severity analysis are two major transportation safety research areas, which have been extensively studied. Lord and Mannering (2010) conducted a comprehensive review of state-of-the-art crash frequency studies and their limitations. Savolainen, Mannering, Lord, and Quddus (2011) conducted a similar study on crash-injury

severities in 2011. Mannering and Bhat (2014) summarized analytical methods used in these two transportation research areas and provided future directions. The essential methodology behind transportation safety analysis is to identify the relationship between a large variety of variables and crash occurrence or crash severity. To achieve this goal, a variety of methods have been applied and largely accepted by transportation researchers. The research team developed a weblink that provides a comprehensive bibliographic list of traffic safety studies. Interested readers can consult this webpage (<http://subasish.github.io/pages/TRB2016/crash.html>) for further investigation. The state-of-the-art methods include logistic regression (Al-Ghamdi, 2002; Dissanayake & Lu, 2002; Richard, Kim, & Ulfarsson, 2017), decision trees and neural networks (Chung, 2013; da Cruz Figueira, Pitombo, de Oliveira, & Larocca, 2017; Khan, Bill, & Noyce, 2015; Prato, Gitelman, & Bekhor, 2011; Saha, Alluri, & Gan, 2015), support vector machines (Ahmadi, Jahangiri, Berandi, & Machini, 2018; Chen, Zhang, Qian, Tarefder, & Tian, 2016; Li, Lord, Zhang, & Xie, 2008; Sun, Das, & Broussard, 2016), rough sets (Kim, Pant, & Yamashita, 2008), text mining (Brooks, 2008; Brown, 2016; Gao & Wu, 2013; Rakotonirainy, Chen, Scott-Parker, Loke, & Krishnaswamy, 2015; Zhang, Green, Chen, & Souleyrette, 2019), Twitter mining (Panagiotopoulos, Barnett, Bigdeli, & Sams, 2016), multiple correspondence analysis (Das, Avelar, Dixon, & Sun, 2018; Das, Brimley, Lindheimer, & Pant, 2017; Das & Sun, 2015, 2016; Jalayer, Pour-Rouholamin, & Zhou, 2018), association rules mining (Ait-Mlouk, Gharnati, & Agouti, 2017; Das, Dutta, Avelar, et al., 2018; Das, Dutta, Jalayer, Bibeka, & Wu, 2018; Das, Dutta, & Sun, 2019; Geurts, Thomas, & Wets, 2005; Weng & Li, 2017; Weng, Zhu, Yan, & Liu, 2016), association rules negative binomial miner (Das, Minjares-Kyle, Avelar, Dixon, & Bommanayakanahalli, 2017), and deep learning (Das, Dutta, Dixon, Minjares-Kyle, & Gillette, 2018; Gibert, Patel, & Chellappa, 2017).

The studies on H&R crashes are sparse, with only a few major works written over the last 30 years. The key focus areas are characteristics of the driver and victims and roadway environment and crash characteristics. The first research on H&R crashes by Solnick and Hemenway (1994), which stated a possible association between alcohol and H&R behavior. Since then, H&R crashes have been slowly attracting increasing attention of researchers as H&R data become more accessible, such as Fatality Analysis Reporting System (FARS) databases. Those published results identified a wide range of factors that contributed to H&R crashes (Aidoo, Amoh-Gyimah, & Ackaah, 2013; Bahrololoom, Moridpour, Tay, & Young, 2017; Jiang, Lu, Chen, & Lu, 2016; Kim et al., 2008; Lopez, Glickman, Soumerai, & Hemenway, 2017; MacLeod, Griswold, Arnold, & Ragland, 2012; Roshandeh, Zhou, & Behnood, 2016; Solnick & Hemenway, 1994, 1995;

Tay, Barua, & Kattan, 2009; Tay, Kattan, & Sun, 2010; Tay, Rifaat, & Chin, 2008; Zhang et al., 2014; Zhou, Roshandeh, Zhang, & Ma, 2016). However, only two studies focused on bicycle involved H&R crashes (Bahrololoom et al., 2017; Lopez et al., 2017). These studies found that the most critical contributing variables are geometrics, environmental factors, vehicle and driver features.

The literature identified vehicle features, such as type and age, and human factors are closely correlated with H&R crashes (Aidoo et al., 2013; Bahrololoom et al., 2017; Kim et al., 2008; Lopez et al., 2017; MacLeod et al., 2012; Roshandeh et al., 2016; Solnick & Hemenway, 1995; Tay et al., 2008, 2009, 2010; Zhou et al., 2016). Several researchers pointed out the likelihood of being caught, level of severity of the crash, and risk-taking tendency of the at-fault driver are the main determinants of fleeing. Even with the increasing accessibility of H&R data sets, researchers face the problem of lacking complete data sets. In many cases, researchers conduct studies based on the data set with available variables. For most of the H&R crashes reports, information about the victim is readily available. However, the driver's information is often absent. This could lead to a small size or an incomplete data set. Using such limited data, the associations between potential factors and fleeing from the crash site may produce questionable results. A comprehensive literature review of H&R crashes can be found in Das, Dutta, Kong, & Sun (2018) study.

The main objective of this study is to identify the crash and geometric factors that influence the decisions of drivers to flee after crashes in Louisiana. This research focuses on studying the relationship of level of severity of the crash, geometric, crash types, and environmental variables. Previous studies on this topic extensively used methods like logistic regression. To determine the most relevant patterns, this study applied 'market basket analysis,' which is also known as 'association rules mining.'

### 3. Methodology

With the rapidly growing technology, the increasing size and complexity of data gradually become obstacles of employing conventional research methods. Especially for those data sets with a large number of features, conventional statistical models cannot determine hidden associations. Without any prior assumption, many algorithms in data mining can determine hidden and nontrivial patterns. These algorithms can help researchers to find patterns rather than confirm prior hypotheses. For this reason, data mining methods are not only concerned with algorithmic capabilities but also provide tools to analyze work without any prior assumptions.

### 3.1. Overview of market-based analysis

Market basket analysis is a popular data mining approach. As a nonparametric method, it avoids making any parametric assumptions as most of the parametric methods do. It also has great flexibility while dealing with data sets with a significant amount of variables, which is called frequent itemset (Weng et al., 2016). To tackle the frequent itemset/product problem, researchers invented a large number of algorithms. Among these algorithms, *APRIORI*, developed by Agrawal and Srikant (1994), 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.

This algorithm helps researchers mine out frequently occurring itemsets, consequences, arrangements, and proper associations between various items. A set of definitions are given here before demonstrating the method with an example. Let  $I = i_1, i_2, \dots, i_m$  be a set of items (e.g., a set of crash categories for a particular crash record) and  $C = c_1, c_2, \dots, c_n$  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.

The definition of association rule can be demonstrated as  $A \rightarrow B$ , where  $A$  and  $B$  are disjoint itemsets. Here,  $A$  is known as the antecedent and  $B$  is the consequent. Generally, support is defined as the percentage of casualties in the data set that contains the itemset. Confidence is a ratio of the number of all crashes in  $C$  to the number of crashes that include all items in  $I$ . Lift is a ratio of confidence over expected confidence. The equations of support are listed in Equation 1 to Equation 3.

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

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

$$S(A \rightarrow B) = \frac{\sigma(A \cap B)}{N} \tag{3}$$

Where,

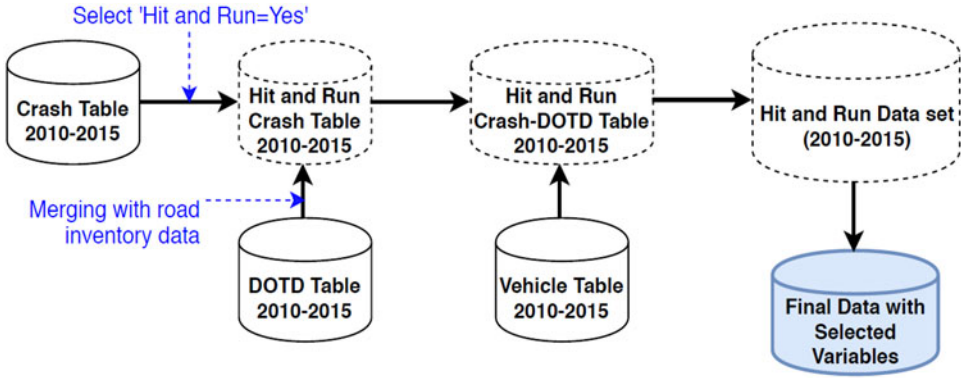
$\sigma(A)$  = Number of incidents with  $A$  antecedent

$\sigma(B)$  = Number of incidents with  $B$  consequent

$\sigma(A \cap B)$  = Number of incidents with both  $A$  antecedent and  $B$  consequent

$N$  = Total number of incidents

$S(A)$  = Support of antecedent



**Figure 1.** Flowchart of data preparation.

$S(B)$  = Support of consequent

$S(A \rightarrow B)$  = Support of the association rule ( $A \rightarrow B$ )

Confidence measures the reliability of the inference of a generated rule. Higher confidence for a  $A \rightarrow B$  indicates that presence of  $B$  is highly visible in the transactions having  $A$ . 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.

$$C(A \rightarrow B) = \frac{S(A \rightarrow B)}{S(A)} \quad (4)$$

$$L(A \rightarrow B) = \frac{S(A \rightarrow B)}{S(A) \cdot S(B)} \quad (5)$$

Where,

$C(A \rightarrow B)$  = Confidence of the association rule ( $A \rightarrow B$ )

$L(A \rightarrow B)$  = Lift of the association rule ( $A \rightarrow B$ )

The lift measure is used to determine the correlation between antecedent and consequent. A lift value above 1 indicates significant interdependence between the antecedent and the consequent, while a value smaller than 1 indicates low interdependence, and a value of 1 designates independence. A rule with a single antecedent and a single consequent is defined as a two-product rule; similarly, a rule with two antecedents and single consequent or one antecedent and two consequents is defined as a three-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.

#### 4. Data preparation

The data set of the current study includes police-reported crashes in Louisiana from 2010 to 2015. Among several variables, one of them shows

**Table 1.** Intersection and segment-related crash severity distribution.

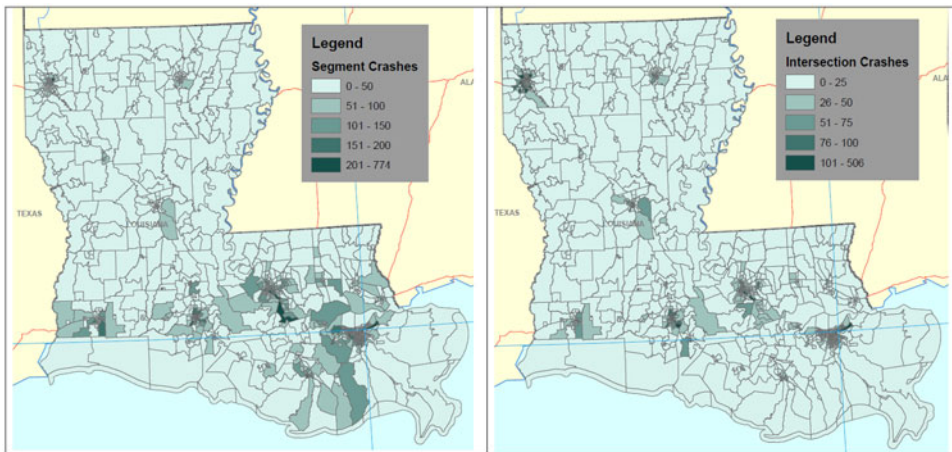
Year	K	A	B	C	O	Yearly Total
<b>Intersection Crashes</b>						
2010	7	44	214	928	4,526	5,719
2011	10	38	238	929	4,412	5,627
2012	8	36	277	1,065	4,998	6,384
2013	5	63	290	1,133	5,331	6,822
2014	13	37	288	1,278	5,587	7,203
2015	12	29	312	1,230	5,285	6,868
Grand Total	55	247	1,619	6,563	30,139	38,623
<b>Segment Crashes</b>						
2010	22	44	337	1,380	9,248	11,031
2011	20	70	329	1,452	9,620	11,491
2012	33	48	312	1,461	9,591	11,445
2013	31	44	367	1,535	10,008	11,985
2014	28	55	364	1,662	10,325	12,434
2015	27	59	350	1,584	9,774	11,794
Grand Total	161	320	2,059	9,074	58,566	70,180

whether the crash was H&R. Louisiana crash database contains four major data tables: (1) crash table, (2) roadway inventory table known as Department of Transportation and Development (DOTD) table, (3) vehicle table, and (4) occupant table. The crash table contains crash and environment-related information. The DOTD table contains roadway geometry information for each crash. The vehicle table provides information on all involved vehicles. All of these tables contain a unique identifier (CRASH\_NUM) for each crash, which is generally used for data merging. At first, this study identified H&R crashes by using the H&R indicator column in the crash table. Later, this table was merged with DOTD and vehicle table. Figure 1 illustrates the data preparation task in a flowchart.

## 5. Descriptive statistics

The data integration process is based on two key variables: crash id and an identifier of H&R crashes. The identifier is based on the structured crash database, which was developed from the individual police report. Louisiana crash data provides crash narratives in electronic format for the majority of the crash events. A preliminary matching shows around 98% accuracies of the reported 'hit-and-run' class identifier. Two separate data sets were prepared based on the location of the H&R crashes: intersection-related and segment-related. Intersection-related crashes were identified by using the values of two columns. If the intersection column shows 'yes' and the distance to intersection is 250 ft, the crash is assigned as intersection related crashes. The geometric information of main roadway (as identified in the police report) is considered as the geometric data associated with the specific crash. The variable selection was conducted based on the literature search. Variables with over-represented category or attribute are removed.





**Figure 2.** Hit-and-run crash locations in Louisiana.

The inclusion of such a variable creates noise in the generated rules. For example, normal weather and dry pavement condition represent above 95%. Some variables are removed due to their high correlation with other variables. For example, crash hour shows high correlation with the lighting condition. [Table 1](#) provides a summary of H&R crashes since 2010. It shows the prevalence of the increase in H&R crashes over the past few years. The values show that segment-related crashes are nearly double compared to intersection-related crashes. The count of intersection-related crashes is almost half of the segment-related crashes. For fatal crashes, segments showed substantially greater crash occurrences than intersections. These differences are not large for injury crashes. For example, nonincapacitating crashes are 1,619 and 2,059 for intersection and segment H&R crashes respectively. H&R property damage only (PDO) crashes are two times more likely to be segment related than intersection related.

The locations reveal that the majority of these crashes happened in urban areas and city streets. This trend is in line with findings of other studies (MacLeod et al., 2012; Solnick & Hemenway, 1994). Additionally, it is also visible that interstates (posted speed limit of 65 mph and above) represent a higher number of H&R crashes. From the data, it is found that around 45% of H&R crashes happened on interstate roadways. This finding is not in line with MacLeod et al. (2012) study that claimed that H&R crashes are less likely to occur in locations with higher posted speed limits. [Figure 2](#) illustrates the distribution of H&R crashes based on the U.S. Census tracts. Census tracts usually have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people. The spatial area of census tracts varies widely depending on the density of settlement. The H&R crashes occur more at populated urban areas. The spatial patterns of segment crashes are more diverse compared to intersection crashes. Another

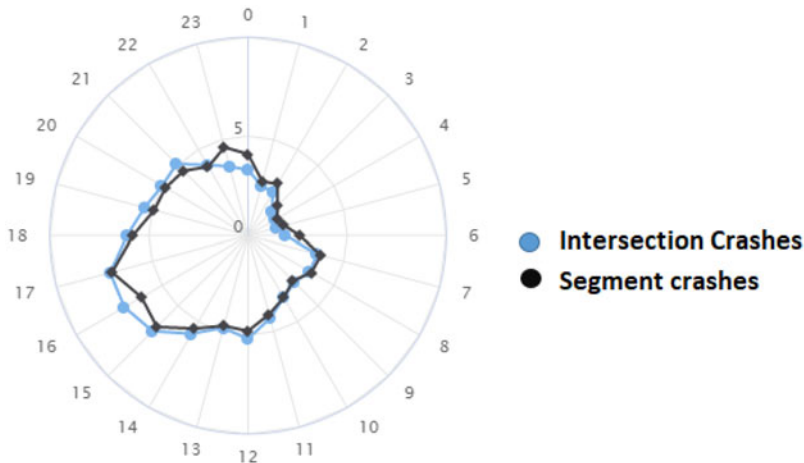


Figure 3. Percentage of crashes by hours.

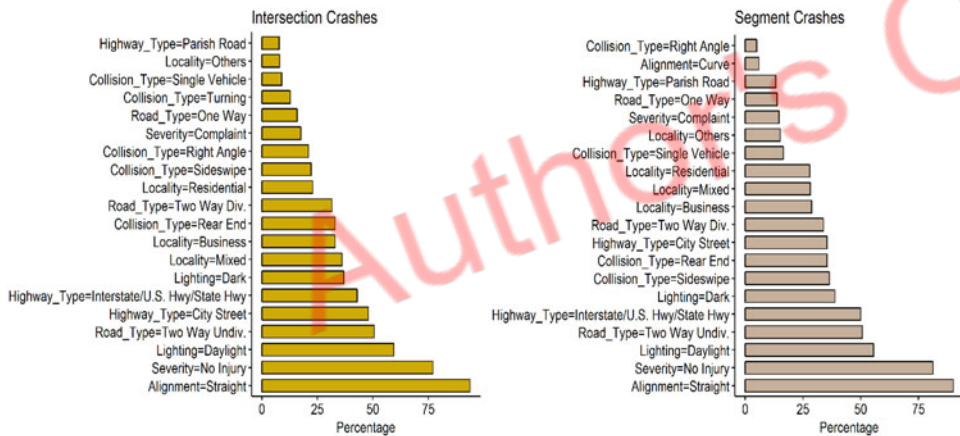


Figure 4. Most frequent items in hit-and-run crashes.

important pattern of the H&R crashes is the temporal pattern. Figure 3 illustrates the temporal percentage distribution of H&R intersection and segment crashes by the hours of the day. Segment crashes are higher in percentages between 11 pm to 6 am. Intersection crashes are higher in percentages between 12 pm to 10 pm.

## 6. Results and findings

The primary data set contains 108,803 H&R crash data with 20 variables. After performing a correlation analysis and missing value entry removal, the final data set contains 87,459 crashes with nine variables. Table 2 lists the chi-squared test values (a simplistic test to determine the difference between the data set attributes) and descriptive statistics of the key variables. The  $p$  values from the chi-squared tests indicate that the variable

**Table 2.** Chi-squared tests and descriptive statistics for key variables by segment and intersection-related hit and run crashes.

Attributes	Segment	Intersection	<i>p</i> -value	Attributes	Segment	Intersection	<i>p</i> value
Alignment (%)			<0.001	Road Type (%)			<0.001
Straight	49,813 (90.0)	30,143 (93.8)		Two-way undivided	28,008 (50.6)	16,281 (50.7)	
Curve	3,323 (6.0)	1,170 (3.6)		Two-way div.	18,744 (33.9)	10,136 (31.5)	
Others	2,194 (4.0)	816 (2.5)		One way	7,631 (13.8)	5,091 (15.8)	
Highway type (%)			<0.001	Others	947 (1.7)	621 (1.9)	
Interstate/U.S. Hwy/State Hwy	27,631 (49.9)	13,796 (42.9)		Severity (%)			<0.001
City street	19,587 (35.4)	15,411 (48.0)		Fatal	156 (0.3)	52 (0.2)	
Parish road	7,312 (13.2)	2,499 (7.8)		Severe	281 (0.5)	217 (0.7)	
Others	800 (1.4)	423 (1.3)		Moderate	1,869 (3.4)	1455 (4.5)	
Lighting (%)			<0.001	Complaint	8,077 (14.6)	5,636 (17.5)	
Daylight	30,736 (55.6)	19,085 (59.4)		No injury	44,947 (81.2)	24,769 (77.1)	
Dark	21,480 (38.8)	11,855 (36.9)		Day of week = Weekend (%)	16,952 (30.6)	9,324 (29.0)	<0.001
Dawn/dusk	1,340 (2.4)	657 (2.0)		Collision type (%)			<0.001
Others	1,774 (3.2)	532 (1.7)		Sideswipe	20,159 (36.4)	7,121 (22.2)	
Locality (%)			<0.001	Rear end	19,589 (35.4)	10,587 (33.0)	
Business	15,902 (28.7)	10,600 (33.0)		Single vehicle or noncollision with vehicle	9,118 (16.5)	2,851 (8.9)	
Residential	15,464 (27.9)	7,380 (23.0)		Right angle	2,739 (5.0)	6,735 (21.0)	
Mixed	15,606 (28.2)	11,586 (36.1)		Turning	2,196 (4.0)	4,092 (12.7)	
Others	8,358 (15.1)	2,563 (8.0)		Head on	1,413 (2.6)	599 (1.9)	
				Others	116 (0.2)	144 (0.4)	

categories are significantly different for the segment and intersection-related H&R crashes. Most of the H&R crashes (around 90% to 94%) happened on straight-aligned roadways. Interstate obviously shows a higher proportion of segment-related crashes and city streets show a higher proportion of intersection-related crashes. Crashes in dark conditions are slightly higher in segment-related crashes. Residential localities show a higher proportion of segment-related H&R crashes. On the other hand, business and mixed localities show higher proportions in intersection-related crashes. One-way roadways show higher intersection-related crashes. Injury crashes are higher in intersection-related crashes. Sideswipe and single-vehicle crashes are higher in proportion in segment-related crashes. Intersection-related crashes show a higher percentage towards angle and turning crashes.

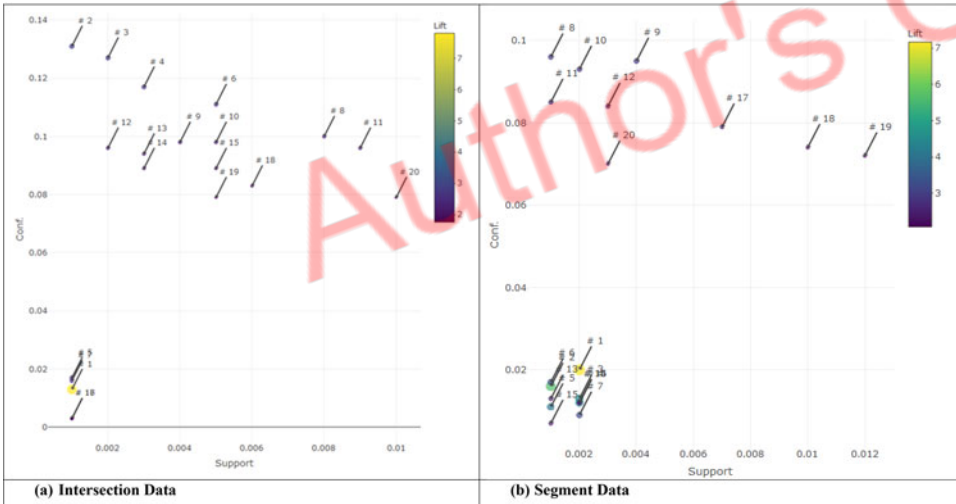
Market basket analysis can handle complex data sets with a large number of variables. Unlike the parametric models, there is no need to predetermine the assumptions and functional forms in association rules mining. Additionally, the generated rules have potentials in exhibiting latent patterns in data. The current study used 'APRIORI algorithm to perform the analysis. The rules are generated by using 'arules' package of the open source R Software (Hahsler, Buchta, Gruen, & Hornik, 2010). Figure 4 shows the top twenty most frequent items in each of the data sets. Straight alignment, no injury, daylight, and two-way undivided roadways are the top four frequent item sets in both of the data sets. These four items are strongly associated with the occurrence of intersection and segment-related H&R crashes. From the fifth ranking, the data sets show the difference between the rankings. For example, the higher functional class is the fifth most frequent itemset in segment data. City streets position in the fifth most frequent itemset in the intersection data set. Localities (business and mixed) are in the top ten most frequent itemsets in the intersection database. These two categories are not in the top ten most frequent itemsets in the segment database.

### 6.1. Key contributing patterns for intersection crashes

This study used APRIORI algorithm that includes two separate steps: (1) minimum support is used to find all of the frequent item sets in the database and (2) these frequent item sets and the minimum confidence constraint are used to form rules. Determining the optimum support and confidence values is important in performing association rules mining. As fatal crashes overall represent a small percentage of total crash occurrences, minimum support and confidence values for intersection data is considered 0.1% and 1%, respectively. These thresholds are determined after conducting several trials and errors. For real-world implications, researchers can perform more tests to determine the suitable values of support and confidence based on different decision criteria. Moreover, there is a need to investigate the rules with more items based on specified conditions. The higher the lift, the higher the associations between the variables. Table 3 lists the top 20 rules that contain highly associated characteristics in H&R intersection crashes. The rule with the highest lift value is single vehicle → fatal. The explanation of this rule is: 0.1% of the single vehicle H&R crashes are associated with fatalities; Out of all single vehicle H&R crashes, 1.3% were fatal crashes. The proportion of fatal single vehicle H&R crashes was 7.802 times the proportion of fatal H&R crashes in the complete data set. Out of the top 20 rules, single vehicle H&R crashes are present in 14 rules. Other frequent items are *straight alignment, right angle collision, one-way/two-way roads, dark, and city street*. Figure 5a shows the scatter plots

**Table 3.** Top 20 rules sorted by the lift values (intersection data).

Rule ID	Rules	Support	Confidence	Lift
#1	Single vehicle → Fatal	0.001	0.013	7.802
#2	One way and single vehicle → Moderate	0.001	0.131	2.891
#3	Business locality and single vehicle → Moderate	0.002	0.127	2.806
#4	Mixed locality and single vehicle → Moderate	0.003	0.117	2.584
#5	Single vehicle → Severe	0.001	0.017	2.493
#6	City street and single vehicle → Moderate	0.005	0.111	2.441
#7	Straight alignment and single vehicle → Severe	0.001	0.016	2.395
#8	Straight alignment and single vehicle → Moderate	0.008	0.100	2.214
#9	Daylight and single vehicle → Moderate	0.004	0.098	2.161
#10	Dark and single vehicle → Moderate	0.005	0.098	2.159
#11	Single vehicle → Moderate	0.009	0.096	2.130
#12	Two-way divided and single vehicle → Moderate	0.002	0.096	2.110
#13	Interstate/U.S. Hwy/State Hwy. and Single vehicle → Moderate	0.003	0.094	2.067
#14	One-way road and right angle → Moderate	0.003	0.089	1.959
#15	Two-way undivided and single vehicle → Moderate	0.005	0.089	1.958
#16	Dark → Fatal	0.001	0.003	1.876
#17	Straight alignment and dark → Fatal	0.001	0.003	1.842
#18	Dark and right angle → Moderate	0.006	0.083	1.839
#19	Residential locality and right angle → Moderate	0.005	0.079	1.755
#20	City street and right angle → Moderate	0.010	0.079	1.753

**Figure 5.** Scatter plots based on support, confidence, and lift values.

of the top 20 rules based on the parameter values. An interactive version of the plot is located at a weblink (<https://rpubs.com/subasish/481019>). It is found that the majority of the rules have high confidence values (0.08 and above). Five rules (#1, #5, #7, #16, and #17) have lower support and confidence values. These rules are associated with either fatal or severe crashes. The key findings are the following:

- The top rules for fatal crashes are associated with single vehicles and dark condition. The spider plots generated for intersection/segment crashes by hour shows a high percentage of nighttime crashes at intersections than segments. Other studies showed that the greater the visibility of a potential

**Table 4.** Top 20 rules sorted by the lift values (segment data).

Rule ID	Rules	Support	Confidence	Lift
#1	Dark and Single vehicle → Fatal	0.002	0.020	7.170
#2	Interstate/U.S. Hwy/State Hwy and single vehicle → Fatal	0.001	0.016	5.682
#3	Straight alignment and single vehicle → Fatal	0.002	0.013	4.562
#4	Single vehicle → Fatal	0.002	0.012	4.279
#5	Two-way undivided and single vehicle → Fatal	0.001	0.011	4.036
#6	Dark and single vehicle → Severe	0.001	0.017	3.330
#7	Interstate/U.S. Hwy/State Hwy and dark → Fatal	0.002	0.009	3.207
#8	One-way road and single vehicle → Moderate	0.001	0.096	2.850
#9	City street and single vehicle → Moderate	0.004	0.095	2.804
#10	Business locality and single vehicle → Moderate	0.002	0.093	2.743
#11	Other alignments and single vehicle → Moderate	0.001	0.085	2.522
#12	Mixed locality and single vehicle → Moderate	0.003	0.084	2.483
#13	Two-way undivided and single vehicle → Severe	0.001	0.013	2.475
#14	Straight alignment and single vehicle → Severe	0.002	0.012	2.450
#15	Dark and two-way undivided → Fatal	0.001	0.007	2.399
#16	Single vehicle → Severe	0.002	0.012	2.375
#17	Dark and single vehicle → Moderate	0.007	0.079	2.351
#18	Straight alignment and single vehicle → Moderate	0.010	0.074	2.186
#19	Single vehicle → Moderate	0.012	0.072	2.123
#20	Two-way divided and single vehicle → Moderate	0.003	0.070	2.068

crash, either through more potential witnesses on heavily trafficked roads or better lighting conditions, the less likely a H&R will occur (Benson et al., 2018; MacLeod et al., 2012).

- The top rules for severe crashes are associated with single vehicle, and straight alignment. The top rules for moderate crashes are associated with single vehicle, one-way roads, business locality, city street, and straight alignment.

## 6.2. Key contributing patterns for segment crashes

Table 4 lists the top 20 rules that contain highly associated characteristics in H&R segment crashes. The minimum support and confidence values for segment data is considered 0.1% and 0.5%, respectively. The rule with the highest lift value is dark and single vehicle → fatal. The explanation of this rule is 0.1% of the single vehicle H&R crashes at dark are associated with fatalities; Out of all single vehicle H&R crashes at dark, 1.3% were fatal crashes. The proportion of fatal single vehicle H&R crashes at dark was 7.170 times the proportion of fatal H&R crashes in the complete data set. Out of the top 20 rules, single vehicle H&R crashes are present in 17 rules. The next frequent item is dark. Other frequent items are Interstate/ U.S. Hwy/State Hwy, and two-way undivided roadways. It is also found that fatal and severe crashes are over-represented in the top rules for segment crashes than intersection crashes. Figure 5b shows the scatter plots of the top 20 rules based on support, confidence and lift. An interactive version of the plot is located at a weblink (<https://rpubs.com/subasish/481018>). The figure shows that around 50% of the rules have low support and confidence scores. These rules are mostly associated with either fatal and severe crashes. The key findings are the following:

- The top rules for fatal crashes are associated with single vehicle, dark condition, Interstate/U.S. Hwy/State Hwy, and two-way undivided roadways. For segment crashes, top five rules are associated with fatal crashes. However, Tay et al. (2009) found that interstate highways and county and municipal roadways are more common H&R locations than the U.S. Hwy/State Hwy.
- The top rules for severe crashes are associated with single vehicle, dark condition, two-way undivided roadways, and straight alignment. Undivided roadways are associated with the high likelihood of a H&R crash in the U.S. (Tay et al., 2009).
- The top rules for moderate crashes are associated with single vehicle, one-way roads, business locality, city street, and straight alignment. This finding is similar to the finding of moderate crashes at intersections.

In many cases, the top 20 or 30 rules cannot shed enough insights from a complex data set. A balloon plot is a good data visualization technique in displaying the association between the representatives of grouped antecedent and consequents from the total set of generated rules. The light gray color of the balloons indicates small lift values while the dark gray indicates high lift value. Size is used as the representation of the support values. Balloon with large radius indicates high support value. On the right hand side (RHS), the antecedent values are presented. To show the spread of the rules, one balloon plot has been developed for intersection crash data. One clear visible pattern is that rules with 'complaint' or 'no injury' as consequents have high support values. It is obvious due to the higher number of 'complaint' or 'no injury' in the database. The left hand side (LHS) of Figure 6 indicates that there are five rules containing both or either of the items single vehicle and curve in the antecedents while the consequent is a fatal crash.

For segment and intersection crashes, single-vehicle crashes are associated with high crash severities. Another frequent item is *dark*. To reduce single-vehicle crashes, the critical measure is aiding so that drivers can keep the vehicles on the traveling path. Locations with inadequate roadway markings and associated countermeasures (rumble strips, raised pavement markers) require additional investigations regarding H&R crashes. Such roadways with high H&R crashes require further attention. Lighting at night is considered as an effective countermeasure to reduce crashes. Roadways with high H&R crash occurrences can be considered as the potential sites for lighting installation. Majority of H&R crashes happen in urban settings. To reduce intersection-related H&R crashes, installation of traffic cameras for road surveillance can improve safety.

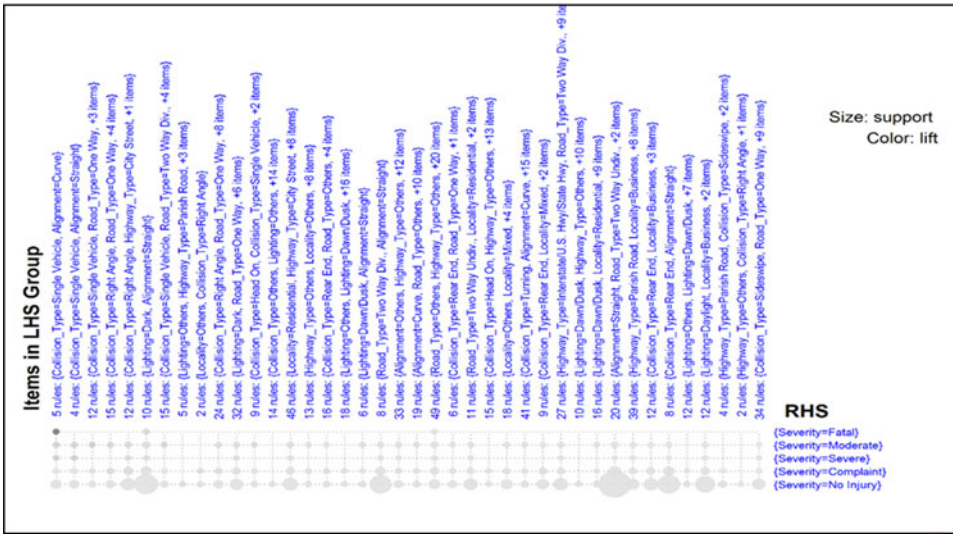


Figure 6. Balloon plot of the rules generated for intersection data.

Majority of the H&R studies are focused on driver characteristics. As driver characteristics are missing in many cases, it is important to examine crash characteristics to identify areas, locations, and patterns of characteristics where potential countermeasures could be focused. Benson et al. (2018) recommended that low-speed roadways, urban areas with high nonmotorized mobilities should all be considered when developing H&R related countermeasures. The maximum sentences (range from 6 months to 30 years) for a H&R crash vary from state to state in the U.S. (Grembek & Griswold, 2012). This study also showed that higher sentencing does not have an immediate effect on the rates of H&R fatalities. Another potential practice is to increase the probability of capturing the offender. A program named 'Yellow Alert Program' was implemented in the city of Los Angeles, California, in 2016. The effectiveness of this program has not been determined yet. There is a need to evaluate the newer countermeasures that focus on mitigating potential scenarios of H&R crashes and identifying offenders. The geometric feature patterns identified in this study can be used in assessing similar locations as 'hit and run crash prone'. There is a need for safety education on H&R crashes. Drivers aware of the punishment of being caught will be reluctant in fleeing the scene. So, there is a need to evaluate the role of public education in public perception of H&R laws.

**7. Conclusions**

There is a small number of studies on H&R crashes, especially when someone considers the high percentage of these incidents. There is a need to



know the reason behind H&R crashes to develop relevant policies and implement appropriate countermeasures. This study attempts to understand the effect of the various crash and geometric variables on H&R crashes, unlike other previous driver-trait-focused studies that used limited data. The goal of the study is to help reducing H&R crashes by determining the key contributing patterns from a data mining approach. Together with visualizations of the rules, the current method provides interpretable results to the transportation safety practitioners. Based on the high lift values in the generated rules, single-vehicle crash is found as the most highly relevant variable. **Dark** condition is the second most common factor in the generated rules. These two factors are also associated with fatal and severe injury crashes. The rules also show that H&R crashes happen mostly in urban settings. For intersection crashes, right-angle crashes are also present in some rules with high lift values. Additionally, it was found that segment-related crashes are over-represented in fatal and severe crashes compared to intersection crashes. The findings call for improvement of roadway markings and lighting conditions. More traffic cameras at intersections will help in detecting H&R offenders. Strict state laws and policies like Yellow Alert Program can help in detecting offenders which in turn will reduce the tendencies of fleeing a crash scene.

This study has some limitations. Determination of the optimum or most suitable support and confidence thresholds requires more thorough investigation. This study excluded driver characteristics in the analysis. Although the study is mainly focused on determining the impact of the crash and geometric properties, an analysis using the driver characteristics may provide more intuitive results. The results are needed to be carefully interpreted due to the unobserved heterogeneity due to driver related factors. Future studies can incorporate driver characteristics to apply to this technique. However, there are challenges in performing the analysis on a smaller dataset due to the absence of driver characteristics in the H&R crash databases.

## Acknowledgments

We like to thank two anonymous reviewers who provided critical and rigorous comments. The current version of the paper is much improved due to the suggested changes that are made to respond to the queries of the reviewers.

## ORCID

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

## References

- Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In the Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1215, 487-499.
- Ahmadi, A., Jahangiri, A., Berandi, V., & Machini, S. G. (2018). Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. *Journal of Transportation Safety & Security*. doi:10.1080/19439962.2018.1505793
- Aidoo, E. N., Amoh-Gyimah, R., & Ackaah, W. (2013). The effect of road and environmental characteristics on pedestrian hit-and-run accidents in Ghana. *Accident Analysis & Prevention*, 53, 23–27. doi:10.1016/j.aap.2012.12.021
- Ait-Mlouk, A., Gharnati, F., & Agouti, T. (2017). An improved approach for association rule mining using a multi-criteria decision support system: A case study in road safety. *European Transport Research Review*, 9(3), 40.
- Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(6), 729–741.
- Bahrololoom, S., Moridpour, S., Tay, R., & Young, W. (2017). *Factors affecting hit and run bicycle crashes in Victoria, Australia*. Australasian Road Safety Conference, 2017, Perth, Western Australia, Australia.
- Benson, A., Arnold, L., Tefft, B., & Horrey, W. J. (2018). Hit-and-run crashes: Prevalence, contributing factors and countermeasures. AAA Foundation, Washington DC.
- Brooks, B. (2008). Shifting the focus of strategic occupational injury prevention: Mining free-text, workers compensation claims data. *Safety Science*, 46(1), 1–21. doi:10.1016/j.ssci.2006.09.006
- Brown, D. E. (2016). Text mining the contributors to rail accidents. *IEEE Transactions on Intelligent Transportation Systems*, 17(2), 346–355. doi:10.1109/TITS.2015.2472580
- Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, 90, 128–139. doi:10.1016/j.aap.2016.02.011
- Chung, Y.-S. (2013). Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. *Accident Analysis & Prevention*, 61, 107–118. doi:10.1016/j.aap.2012.08.015
- da Cruz Figueira, A. D. C., Pitombo, C. S., de Oliveira, P. T. M. e S., & Larocca, A. P. C. (2017). Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil. *Case Studies on Transport Policy*, 5(2), 200–207. doi:10.1016/j.cstp.2017.02.004
- Das, S., Avelar, R., Dixon, K., & Sun, X. (2018). Investigation on the wrong way driving crash patterns using multiple correspondence analysis. *Accident Analysis & Prevention*, 111, 43–55. doi:10.1016/j.aap.2017.11.016
- Das, S., Brimley, B. K., Lindheimer, T., & Pant, A. (2017). Safety impacts of reduced visibility in inclement weather. *Ann Arbor*, 1001, 48109–42150.
- Das, S., Dutta, A., Avelar, R., Dixon, K., Sun, X., & Jalayer, M. (2018). Supervised association rules mining on pedestrian crashes in urban areas: Identifying patterns for appropriate countermeasures. *International Journal of Urban Sciences*, 23(1), 30–48.
- Das, S., Dutta, A., Dixon, K., Minjares-Kyle, L., & Gillette, G. (2018). *Using deep learning in severity analysis of at-fault motorcycle rider crashes*. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(34), 122–134.
- Das, S., Dutta, A., Jalayer, M., Bibeka, A., & Wu, L. (2018). Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration

- using 'eclat' association rules to promote safety. *International Journal of Transportation Science and Technology*, 7(2), 114–123.
- Das, S., Dutta, A., Kong, X. & Sun X. (2018). Hit and run crashes: Knowledge extraction from bicycle involved crashes using first and frugal tree. *International Journal of Transportation Science and Technology*. doi:10.1016/j.ijtst.2018.11.001
- Das, S., Dutta, A., & Sun, X. (2019). Patterns of rainy weather crashes: Applying rules mining. *Journal of Transportation Safety & Security*. doi:10.1080/19439962.2019.1572681
- Das, S., Minjares-Kyle, L., Avelar, R. E., Dixon, K., & Bommanayakanahalli, B. (2017). Improper passing related crashes on rural roadways: Using association rules negative binomial miner. In *the proceedings of the Transportation Research Board 96th Annual Meeting*, Washington DC.
- 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(1), 95–103. doi:10.3141/2519-11
- Das, S., & Sun, X. (2016). Association knowledge for fatal run-off-road crashes by multiple correspondence analysis. *IATSS Research*, 39(2), 146–155. doi:10.1016/j.iatssr.2015.07.001
- Dissanayake, S., & Lu, J. (2002). Analysis of severity of young driver crashes: Sequential binary logistic regression modeling. *Transportation Research Record: Journal of the Transportation Research Board*, 1784(1), 108–114. doi:10.3141/1784-14
- Gao, L., & Wu, H. (2013). Verb-based text mining of road crash report. In *the Proceedings of Transportation Research Board 92nd Annual Meeting*, Washington DC.
- Geurts, K., Thomas, I., & Wets, G. (2005). Understanding spatial concentrations of road accidents using frequent item sets. *Accident Analysis & Prevention*, 37(4), 787–799. doi: 10.1016/j.aap.2005.03.023
- Gibert, X., Patel, V. M., & Chellappa, R. (2017). Deep multitask learning for railway track inspection. *IEEE Transactions on Intelligent Transportation Systems*, 18(1), 153–164. doi: 10.1109/TITS.2016.2568758
- Grembek, O., & Griswold, J. (2012). *On the legal deterrence of pedestrian hit-and-run collisions* (Working Paper). Berkley, CA: University of California, Safe Transportation Research and Education Center.
- Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2010) Arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, 14(15), 1-25.
- Jalayer, M., Pour-Rouholamin, M., & Zhou, H. (2018). Wrong-way driving crashes: A multiple correspondence approach to identify contributing factors. *Traffic Injury Prevention*, 19(1), 35–41. doi:10.1080/15389588.2017.1347260
- Jiang, C., Lu, L., Chen, S., & Lu, J. J. (2016). Hit-and-run crashes in urban river-crossing road tunnels. *Accident Analysis & Prevention, Traffic Safety in China: Challenges and Countermeasures*, 95, 373–380. (October): doi:10.1016/j.aap.2015.09.003
- Khan, G., Bill, A. R., & Noyce, D. A. (2015). Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity. *Transportation Research Part C: Emerging Technologies*, 50, 86–96. doi:10.1016/j.trc.2014.10.003
- Kim, K., Pant, P., & Yamashita, E. (2008). Hit-and-run crashes: Use of rough set analysis with logistic regression to capture critical attributes and determinants. *Transportation Research Record: Journal of the Transportation Research Board*, 2083(1), 114–121. doi: 10.3141/2083-13
- Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention*, 40(4), 1611–1618. doi: 10.1016/j.aap.2008.04.010

- Lopez, D., Glickman, M. E., Soumerai, S. B., & Hemenway, D. (2017). Identifying factors related to a hit-and-run after a vehicle-bicycle collision. *Journal of Transport & Health*, 8, 299–306.
- 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. doi:10.1016/j.tra.2010.02.001
- Louisiana Crash Data Reports. (2019). <http://datareports.lsu.edu/> (Accessed on May 10, 2019).
- MacLeod, K. E., Griswold, J. B., Arnold, L. S., & Ragland, D. R. (2012). Factors associated with hit-and-run pedestrian fatalities and driver identification. *Accident Analysis & Prevention*, 45, 366–372. doi:10.1016/j.aap.2011.08.001
- Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1, 1–22. doi:10.1016/j.amar.2013.09.001
- NHTSA. (2016). *National Automotive Sampling System (NASS) General Estimates System (GES): Analytical User's Manual 1988–2015*. US Department of Transportation, Washington DC.
- Panagiotopoulos, P., Barnett, J., Bigdeli, A. Z., & Sams, S. (2016). Social media in emergency management: Twitter as a tool for communicating risks to the public. *Technological Forecasting and Social Change*, 111, 86–96. doi:10.1016/j.techfore.2016.06.010
- Prato, C. G., Gitelman, V., & Bekhor, S. (2011). Pattern recognition and classification of fatal traffic accidents in Israel: A neural network approach. *Journal of Transportation Safety & Security*, 3(4), 304–323. doi:10.1080/19439962.2011.624291
- Rakotonirainy, A., Chen, S., Scott-Parker, B., Loke, S. W., & Krishnaswamy, S. (2015). A novel approach to assessing road-curve crash severity. *Journal of Transportation Safety & Security*, 7(4), 358–375. doi:10.1080/19439962.2014.959585
- Richard, K., Kim, S., & Ulfarsson, G. F. (2017). A hierarchical Bayesian logistic regression with a finite mixture for identifying higher-than-expected crash proportions at intersections. *Journal of Transportation Safety & Security*. doi:10.1080/19439962.2017.1337054
- Roshandeh, A. M., Zhou, B., & Behnood, A. (2016). Comparison of contributing factors in hit-and-run crashes with distracted and non-distracted drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 38, 22–28. doi:10.1016/j.trf.2015.12.016
- Saha, D., Alluri, P., & Gan, A. (2015). Prioritizing highway safety manual's crash prediction variables using boosted regression trees. *Accident Analysis & Prevention*, 79, 133–144. doi:10.1016/j.aap.2015.03.011
- Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, 43(5), 1666–1676. doi:10.1016/j.aap.2011.03.025
- Solnick, S. J., & Hemenway, D. (1994). Hit the bottle and run: The role of alcohol in hit-and-run pedestrian fatalities. *Journal of Studies on Alcohol*, 55(6), 679–684.
- Solnick, S. J., & Hemenway, D. (1995). The hit-and-run in fatal pedestrian accidents: Victims, circumstances and drivers. *Accident Analysis & Prevention*, 27(5), 643–649. doi:10.1016/0001-4575(95)00012-O
- Sun X., Das, S., & Broussard, N. (2016). *Developing crash models with supporting vector machine for urban transportation planning*. 17th International Conference Road Safety On Five Continents (RS5C 2016), Rio de Janeiro, Brazil, 17-19 May 2016.
- Tay, R., Barua, U., & Kattan, L. (2009). Factors contributing to hit-and-run in fatal crashes. *Accident Analysis & Prevention*, 41(2), 227–233. doi:10.1016/j.aap.2008.11.002

- Tay, R., Kattan, L., & Sun, H. (2010). Logistic model of hit and run crashes in calgary. *Canadian Journal of Transportation*, 4(1), 1–10.
- Tay, R., Rifaat, S. M., & Chin, H. C. (2008). A logistic model of the effects of roadway, environmental, vehicle, crash and driver characteristics on hit-and-run crashes. *Accident Analysis & Prevention*, 40(4), 1330–1336. doi:[10.1016/j.aap.2008.02.003](https://doi.org/10.1016/j.aap.2008.02.003)
- Weng, J., & Li, G. (2017). Exploring shipping accident contributory factors using association rules. *Journal of Transportation Safety & Security*, 1–22.
- Weng, J., Zhu, J.-Z., Yan, X., & Liu, Z. (2016). Investigation of work zone crash casualty patterns using association rules. *Accident Analysis & Prevention*, 92, 43–52. doi:[10.1016/j.aap.2016.03.017](https://doi.org/10.1016/j.aap.2016.03.017)
- Zhang, G., Li, G., Cai, T., Bishai, D. M., Wu, C. & Chan, Z. (2014). Factors contributing to hit-and-run crashes in China. *Transportation Research Part F: Traffic Psychology and Behaviour*, 23, 113–124.
- Zhang, X., Green, E., Chen, M., & Souleyrette, R. (2019). Identifying secondary crashes using text mining techniques. *Journal of Transportation Safety & Security*. doi:[10.1080/19439962.2019.1597795](https://doi.org/10.1080/19439962.2019.1597795)
- Zhou, B., Roshandeh, A. M., Zhang, S., & Ma, Z. (2016). Analysis of factors contributing to hit-and-run crashes involved with improper driving behaviors. *Procedia Engineering, Green Intelligent Transportation System and Safety*, 137, 554–562. doi:[10.1016/j.proeng.2016.01.292](https://doi.org/10.1016/j.proeng.2016.01.292)

Author's Copy