International Journal of Transportation Science and Technology 8 (2019) 146–160

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

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

Hit and run crashes: Knowledge extraction from bicycle involved crashes using first and frugal tree

Subasish Das [a,*], Anandi Dutta [b], Xiaoqiang Kong [a], Xiaoduan Sun [c]

[a] *Texas A&M Transportation Institute, 3135 Texas A & M University, College Station, TX 77843-3135, United States*
[b] *Texas A&M University, 3127 Texas A & M University, College Station, TX 77843-3127, United States*
[c] *University of Louisiana at Lafayette, Lafayette, LA 70503, United States*

ARTICLE INFO

Article history:
Received 22 May 2018
Received in revised form 6 October 2018
Accepted 9 November 2018
Available online 17 November 2018

Keywords:
Hit and run crashes
Bicycle
Fast and frugal algorithm
Heuristics
Decision tree

ABSTRACT

Hit and run crash is a punishable offense. In many cases, it involves higher severity levels of the associated roadway users due to the delay of emergency help. For vulnerable road users like pedestrians and bicyclists, this issue is more gruesome. We present an analysis of the effect of crash, geometric, and environmental characteristics on the bicycle-involved hit and run crashes by using six years of Louisiana crash data with an application of fast and frugal tree (FFT) heuristics algorithm. Over 1000 bicycle-involved hit and run crashes occurred in Louisiana out of around 108,000 hit and run crashes during 2010–2015. The fatal bicycle crashes represent 10% of the total fatal hit and run crashes. Additionally, hit and run bicycle crashes represent 22% of total bicycle crashes in Louisiana. In the preliminary analysis, we provided statistical significance test of the key contributing factors for two major groups (bicycle-involved hit and run crashes, and not bicycle-involved hit and run crashes). We divided the complete dataset into two separate datasets: training data for model development, and testing data for performance evaluation. FFT identifies five major cues or variable threshold attributes that contribute significantly in predicting bicycle-involved crashes. These cues include fatal and injury crashes, right angle/turning/-head on collisions, city streets/others, intersection, and residential/mixed localities. The balanced accuracy is around 76% for both training and testing data. The current model shows higher sensitivity than other complex and black box machine learning models (e.g., support vector machine, random forest). Findings of our study will provide valuable insights for hit and run bicycle crash reduction in both planning and operation levels.

1. Introduction

Hit and run crashes refer to crashes where the at-fault (responsive 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, specifically for vulnerable roadway users like pedestrians and bicyclists, due to delays in emergency assistance. Although hit and run is a severe crime according to the law enforcement

Peer review under responsibility of Tongji University and Tongji University Press.

[*] Corresponding author.
E-mail address: s-das@tti.tamu.edu (S. Das).

https://doi.org/10.1016/j.ijtst.2018.11.001
2046-0430/© 2018 Tongji University and Tongji University Press. Publishing Services by Elsevier B.V.

agencies and drivers who conduct such crime will face serious criminal charges if caught later. However, the frequency of hit and run crash occurrences are still high. Solnick and Hemenway (1994) defined hit and run as "an interesting crime because it is typically spontaneous and not committed by career criminals. It is also a crime that has received virtually no research attention".

A bicycle as a mode of transport has many environmental and societal benefits as well as health benefits. At the same time, it makes bicyclists as one of the most vulnerable roadway user groups. According to National Highway Traffic Safety Administration (2017), 'there were 818 bicyclists killed in motor vehicle traffic crashes in the United States, an increase from 729 in 2014. An additional estimated 45,000 bicyclists were injured in crashes in 2015'. When a collision happens, a significant probability exists that the bicyclist will sustain an injury, while the likelihood is small that the vehicle occupants will be injured (Haworth and Debnath, 2013; Turner et al., 2017). The vulnerability of the bicyclists can be increased when a hit and run crash victimizes them. Tay et al. (2008) showed that the odds of a driver fleeing the crash scene are 4.67 times more likely in vehicle-bicycle crashes when compared to vehicle-vehicle collisions. During 2010–2015, 108,803 hit and run crashes occurred in Louisiana. Among them, 1.01% was recorded as bicycle-involved hit and run crashes. The fatal bicycle crashes represent 10% of the total fatal hit and run crashes. Moreover, hit and run bicycle crashes represent 22% of total bicycle crashes in Louisiana, which triggers another research scope that is currently not examined in this current study. Such a high percentage in terms of all bicycle crashes require in-depth investigation. This focused on understanding the real roadway environment which eventually triggers a driver flee a hit and run bicycle crash scene. The common assertion is that drivers usually do not flee the crash scene unless there is a possibility of higher punishment or occurrence of a severe injury crash.

Based on a six-year (2010–2015) hit and run crash dataset in Louisiana, we applied first and frugal tree (FFT) heuristics algorithm to select significant factors regarding crash, geometric and environmental characteristics and extract the decision rules for bicycle involvement in the hit and run crashes. Our goal is to make the case that the application lens of fast-and-frugal heuristics is well suited to describing and improving applied decision making for the safety improvement of hit and run vehicle-bicycle collisions. The intent of this paper is to demonstrate an approach that can be used to understand better the factors that influence the occurrences of bicycle-involved hit and run crashes.

2. Literature review

The two dominating area of transportation safety analysis are- crash frequency analysis and crash severity analysis. Lord and Mannering (2010) provided a detailed research synthesis on crash frequency analysis methods and limitations for examining such data. Savolainen et al. (2011) presented a similar assessment on crash severity analysis. Recently, Chen et al. (2016), Li et al. (2008), Sun et al. (2016) updated and bridged both of the previous synthesis studies. The conventional approach to traffic safety analysis has been to establish relationships between a wider variety of variables and crash occurrence. Transportation safety researchers widely accepted machine learning and deep learning tools to investigate critical safety issues. Tools that are widely used by the safety researchers are: decision tree (Chung, 2013; Figueira et al., 2017; Khan et al., 2015; Saha et al., 2015), support vector machine (Chen et al., 2016; Li et al., 2008; Sun et al., 2016), rough sets (Kim et al., 2008), text mining (Brooks, 2008; Brown, 2016; Gao and Wu, 2013), Twitter mining (Panagiotopoulos et al., 2016), multiple correspondence analysis (Das et al., 2017a, 2018a; Das and Sun, 2015, 2016; Jalayer et al., 2018), association rules mining (Ait-Mlouk et al., 2017; Das and Sun, 2014; Das et al., 2018c, 2018d; Geurts et al., 2005; Weng et al., 2016), association rules negative binomial miner (Das et al., 2017b), and deep learning (Das et al., 2018b; Gilbert et al., 2017).

In 1994, Solnick and Hemenway (1994) first published a research article addressing a particular safety issue associated with the hit and run crashes. In the years since the first study, a small number of studies have identified multitude factors that contribute to hit and run occurrences (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 and Hemenway, 1995; Tay et al., 2010, 2009, 2008; Zhang et al., 2014). Some studies used FARS databases (MacLeod et al., 2012; Solnick and Hemenway, 1995, 1994), while others used state or country-specific hit and run crash data (Aidoo et al., 2013; Kim et al., 2008; Tay et al., 2008). In the existing literature, only one study focused on bicycle involvement in hit and run crashes (Bahrololoom et al., 2017). The studies mainly focused on identification of critical contributing factors from a wide net of crash, vehicle, driver, roadway user, geometrics, and environmental variables. Table 1 lists the studies conducted on hit and run crashes. This table provides list of used variables in the studies, analytical procedures, and the major contributing factors derived from the study findings.

The reason behind fleeing depends mainly on the likelihood of being caught and the risk-taking tendency of the at-fault driver. Thus, human factors play a crucial role in hit and run crashes. Many studies showed that vehicle type/age and human factors are associated with hit and run crashes (Bahrololoom et al., 2017; Kim et al., 2008; Lopez et al., 2017; MacLeod et al., 2012; Roshandeh et al., 2016; Solnick and Hemenway, 1995; Tay et al., 2010, 2009, 2008; Zhang et al., 2014). Driver and vehicle information on the hit and run crashes are available only for the cases that have been identified in the police reports. Additionally, solid evidence on driver and vehicle information is not available for many cases. As driver and vehicle information is missing in many cases, the inclusion of these variables would make the data size small and biased to extract significant relationships. Although victim information is readily available in most of the hit and run crashes, the absence of driver information would not help in extracting significant associations. Our research scope was limited to crash, geometric, and environmental variables. However, the roadway environment plays a significant role in hit and run crash occurrences. It

Table 1
Data, methods, and key factors identified in previous studies.

Study	Data/Years	Variables used in the study	Approach	Key factors
Solnick and Hemenway (1994)	FARS 1989–1990 pedestrian fatal crashes	Day of week, time of day, driver (age, gender, licensing status, previous DWI record, previous license suspension, previous crash record, previous speeding suspension, BAC), vehicle model	Logistic regression	• Weekends, nighttime. • Previous intoxication, previous DWI • Male drivers
Solnick and Hemenway (1995)	FARS 1989–1991 pedestrian fatal crashes	Driver (age, gender, previous DWI record, license validity, BAC), pedestrian (age, gender), region, urban/rural, season, day of week, speed limit, lighting	Exploratory analysis	• Urban roadways, summer, weekends, nighttime. • Previous DWI, invalid license • Male and young drivers • Young pedestrians • Intoxication
Kim et al. (2008)	Hawaii (2002–2005)	Gender, age, human factors, resident type, alcohol, stolen, vehicle problem, location, roadway type, surface condition, road defects, alignment, day, time of day, weather.	Rough set analysis; Logistic regression;	• Male, tourist, intoxicated • Stolen vehicle driving • Horizontal curve, weather, lighting
Tay et al. (2008)	Singapore (1992–2002)	Time trend, time of day, occurrence area, vehicle type, foreign vehicle, road type, location, surface, surface condition, age, offending, gender, race, collision type, maneuver, severity	Logistic regression	• Bridges and flyovers, bends, straight roads and shop houses. • Age group (45–69)
Tay et al. (2009)	California FARS (1994–2005)	Day of week, time of day, number of vehicles, pedestrian involvement, collision type, functional class, roadway type, median type, segment/intersection, speed limit, traffic control device, lighting, alignment, road surface, weather, special jurisdiction, construction zone.	Logistic regression	• Day of week, time of day, number of vehicles, pedestrian involvement • Collision type • Route type, median type, types of roadway segment, speed limit, traffic control device, functioning of traffic control device, lighting condition, horizontal alignment and vertical alignment
Tay et al. (2010)	Calgary (2010)	Day of week, time of day, severity, road class, weather, surface condition, lighting, vehicle age, driver age, number of vehicles.	Logistic regression	• Night, weekend, undivided roadways • Single vehicle
MacLeod et al. (2012)	FARS 1998–2007 pedestrian fatal crashes	Day of week, time of day, number of vehicles, region, population density, season, age, gender, lighting, traffic control, speed limit, BAC, prior DWI, prior suspension, valid license, older vehicle.	Logistic regression	• Early morning, poor light conditions, and weekend • Alcohol use and invalid license.
Aidoo et al., 2013	Ghana pedestrian crashes (2004–2010)	Year, region, road environment, day of week, accident severity, weather condition, lighting, road description, surface condition, location, traffic condition,	Logistic regression	• Lighting and weather condition, • Roadway type, median separation, road surface condition and repair • Location and traffic condition
Zhang et al. (2014)	Guangdong Province in China (2006–2010)	77 variables	Logistic regression	• Male, age (25–44), new drivers, no valid license, no insurance • Wet road, merging lane, elevated roadways

Bahrololoom et al. (2017)	Victoria, Australia bicycle crashes (2004–2013)	Crash time, age, helmet use, speed zone, lighting, intent, traffic control	Logistic regression	• Dark AM (0:00 to 6:00 AM) • Bicyclists age (>45 years) • No bike helmet • Speed limit >70 km/hr • No traffic control • Wrong way driving, wrong way biking • Lighting dark
Jiang et al. (2016)	13 urban road tunnels surrounding Huangpu river, Shanghai (2011–2012)	Season, day of week, time of day, weather, speed limit, number of lanes, tunnel length, crash type, vehicle types, number of vehicles, severity	Logistic regression	• Night, exit tunnel • Season, day of week, speed limit
Roshandeh et al. (2016)	Cook county, Illinois (2004–2012)	Day of week, population, crash type, severity, urban/rural, national highway system, surface condition, location, lighting, alcohol, intersection, number of lanes, alignment, median type	Logistic regression	Comparison between distracted and non-distracted crashes are: • Poor lighting condition • Curve level • Non-distracted drivers less flee
Zhou et al. (2016)	Cook county, Illinois (2004–2012)	Time trend, time of day, occurrence area, vehicle type, foreign vehicle, road type, location, surface, surface condition, age, offending, gender, race, collision type, maneuver, severity	Logistic regression	• Variables contributing to hit-and-run crashes varied for different improper driving behaviors.
Lopez et al. (2017)	Boston bicycle crashes (2009–2012)	Time of day, day of week, temperature, precipitation, main street, intersection, taxi, extended door, gender, age, severity, ethnicity	Logistic regression	• Night, weekend • Extended door • Taxi • Male bicyclists

is very important to know at what roadway environment a driver flees a hit and run scene involved with a bicyclist. An in-depth analysis of this issue has not been conducted before. The current study aims to mitigate this research gap.

Historically, studies examining hit and run crashes have focused on developing logistic regressions with a variety of crash, environment, driver, victim, and roadway variables. As bicycle safety is a critical issue, there is a need for developing modern predictive tools to understand the key undertakings associated with the bicycle-involved hit and run crashes. We aimed to mitigate the current research gap by applying a predictive learning tool FFT to determine the key issues associated with bicycle-involved hit and run crashes.

3. Fast and frugal tree (FFT)

Fast and frugal tree (FFT) is a heuristic for binary decisions. It is defined as a decision tree that has $n + 1$ exits, with one exit for each of the first $n - 1$ cues and two exits for the last cue or variable threshold. As process models of decision-making, FFT makes predictions not only about what cues will influence decisions but also how decision makers might use these cues. Initially, an FFT starts by checking the thresholds on the first cue to examine the exit condition and continues to the other cues one after another until final exit criteria is met.

In considering only a few binary variables, FFTs, like all other heuristics, simplify decision problems. FFT is generally composed of three building blocks, similar to take the best:

- Search rule: Look up predictor variables and attributes (variable categories) in the order of their importance.
- Stopping rule: Stop search as soon as one predictor allows it.
- Decision rule: Classify according to this predictor variable.

FFT was first introduced by Laura Martignon in 2003. This binary decision tree has at least one outcome or exit at every node (cue or threshold of variable attribute set) and the decision led by the outcome will not be influenced by rest of unchecked nodes (Martignon et al., 2003). Growing attention has been drawn to this algorithm in recent years in both academia and industries because of its effectiveness and robustness comparing to regression and other classification algorithms (Luan et al., 2011; Phillips et al., 2017).

FFTs have three key advantages over the statistical and machine learning models. These advantages are based on the frugality, simplicity, and prediction accuracy of the FFTs. First, FFTs tend to be both fast and frugal as they typically provide lean decision outcomes. Second, the modeling results are very easy to understand. Finally, FFTs can make good predictions even on the basis of a small amount of noisy data because they are relatively robust against a statistical problem known as overfitting (Phillips et al., 2017).

Three FFT algorithms, max, zigzag, and fan, are commonly used. First, two algorithms have been introduced by Martignon et al. (2003), and the last one was introduced by Phillips et al. (2017). There are four tasks to construct FFT algorithms: select cues; choose a decision threshold for each cue; rank cues; determine exit of each cue (Martignon et al., 2008; Martignon and Hoffrage, 2002; Phillips et al., 2017). The main differences between Martignon's and Phillips' algorithms are: First, fan algorithm takes sensitivity and specificity weighting parameters into consideration while selecting and ordering the cues. The decision maker has the authorization to designate their desired weights of these error trade-offs; Second, fan algorithm sets size restrictions of the decision tree rather than using all available cues as other two algorithms do.

The *fan* algorithm consists of two variants: *ifan* and *dfan*. The first step of constructing FFT is to determine the decision threshold for each cue. Thresholds are single values for numerical cues and sets of factor values for nominal cues. Appling each cue into to training dataset while ignoring others, the single value or factor could maximize the cue's accuracy would be selected. In *ifan* algorithm, the decision maker defines the measurement of the cue's accuracy. Second, the order of cues is ranked based on the value of cue's accuracy. Different with *max* and *zigzag* algorithms, fan algorithm applies a stopping-par to limit the number of cues can be used in the final algorithm. Then, a set of trees will be generated with a fixed order of cues and all possible exit structures. The *fan* algorithm would choose the FFT with the highest accuracy.

An important assumption in the *ifan* algorithm, same as *max* and *zigzag* algorithms, is that all cues do not have any interaction. However, dependencies between cues are possible in some cases. The *dfan* algorithm has been developed to fill this gap. Similar to the *ifan* algorithm, dfan still ranks cues based on the accuracy value. However, the *dfan* could iteratively recalculate the thresholds of each cue and rank cues based on subsets of cases to address the possible interactions rather than only use their accuracy only once on the whole dataset (Phillips et al., 2017).

4. Data preparation

Dataset of the current study are police-reported crashes in Louisiana from 2010 to 2015. Among several variables, one of them shows whether the crash was hit and run. Louisiana crash database contains four major data tables: (1) crash table, (2) roadway inventory table known as 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. The driver information of each vehicle is also included in vehicle table. The occupant table contains occupant information other than the drivers. All of these tables contain a unique identifier (CRASH_-

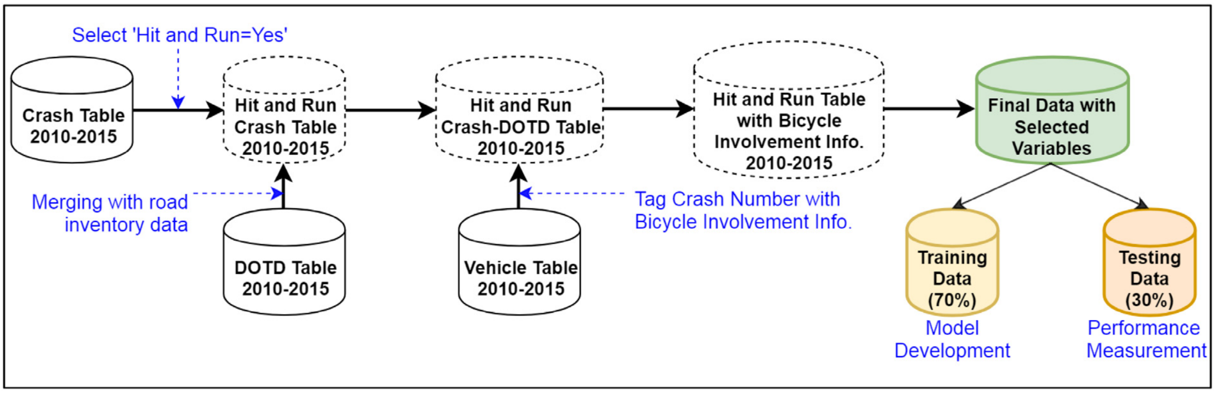

Select 'Hit and Run=Yes'
Crash Table 2010-2015
Hit and Run Crash Table 2010-2015
Hit and Run Crash-DOTD Table 2010-2015
Hit and Run Table with Bicycle Involvement Info. 2010-2015
Final Data with Selected Variables
Merging with road inventory data
DOTD Table 2010-2015
Vehicle Table 2010-2015
Tag Crash Number with Bicycle Involvement Info.
Training Data (70%)
Testing Data (30%)
Model Development
Performance Measurement

Fig. 1. Flowchart of data preparation.

NUM) for each crash, which is generally used for data merging. At first, we identified hit and run crashes by using the hit and run indicator column in the crash table. Later, we merged this table with DOTD and vehicle table. Vehicle table contains supporting columns to identify bicycle involvement in a crash. The following rule was used to identify bicycle-involved in hit and run crashes:

- Crash Table [Hit and Run Crash = Yes] AND
- Vehicle Table: [First Harmful Event = Pedalcycle OR Most Harmful Event = Pedalcycle OR Second Harmful Event = Pedalcycle OR Third Harmful Event = Pedalcycle OR Vehicle Type = Pedalcycle]

From the vehicle table, unique crash numbers are filtered to create the database of bicycle-involved hit and run crashes. The rest hit and run crashes are tagged as not bicycle-involved hit and run crashes. Fig. 1 illustrates the data preparation task in a flowchart.

5. Descriptive statistics

Fig. 2(a) illustrates the locations of the hit and run crashes and Fig. 2(b) shows the locations of the bicycle-involved hit and run crashes in Louisiana. The locations reveal that majority of these crashes happened in urban areas and city streets. As our goal is to identify patterns or significant factors of hit and run bicycle-involved crashes, a comparison between bicycle-

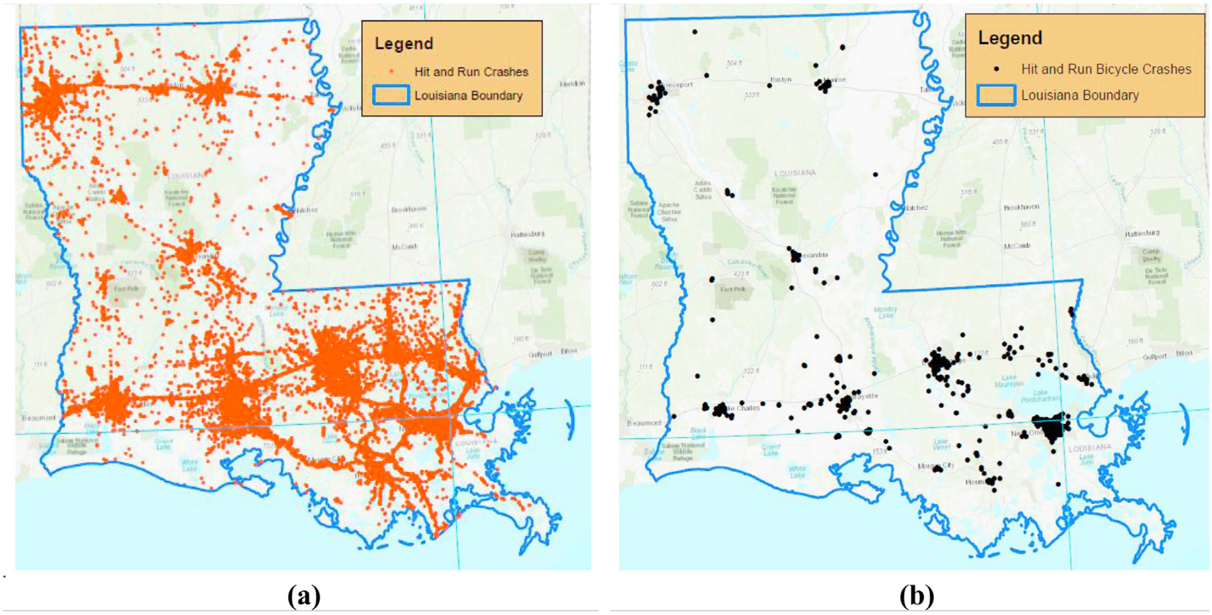


Fig. 2. (a) All Hit and Run crashes in Louisiana, (b) Hit and run bicycle crashes in Louisiana.

involved and not bicycle-involved hit and run crashes was performed. Fig. 3(a) shows the percent distribution of hit and run crashes (in relation to bicycle involvement or not) for the top ten cities with high hit and run frequencies in between 2010 and 2015. The blue dots indicate the percentage of not bicycle-involved crashes and the red dots indicate the percentage of bicycle-involved crashes. For example, New Orleans city represents 56.9% (the percentage is calculated by dividing the frequency of hit and run bicycle-involved crashes by frequency of all hit and run bicycle crashes in top 10 cities) of bicycle-involved hit and run crashes. For not bicycle-involvement, this percentage is 43.4%. These two values indicate that New Orleans data is over-represented for bicycle-involved hit and run crashes. The other cities that show high bicycle-involved hit and run crashes are Opelousas, Lake Charles, Kenner, Houma, and Alexandria. Fig. 3(b) illustrates the percent distribution of hit and run crashes for the DOTD districts. District 2 is over representative for hit and run crashes. For this district, the bicycle-involvement shows high percentage than not bicycle-involvement. The other districts with higher bicycle-involved hit and run crashes are District 3, District 7, and District 8.

Fig. 4 shows the heat chart of hit and run crash frequencies by year for each group with different severity levels. The severity levels are defined with KABCO scale:

- K: Fatal
- A: Incapacitating injury
- B: Non-incapacitation injury
- C: Minor injury
- O: No injury or Property Damage Only (PDO)

The heat chart is the easiest and self-explanatory table to understand the variability for different cases (for example, year and severity in Fig. 4). The table shows the following insights:

- Bicycle-involved crashes represent 1.0% of total hit and run crashes.

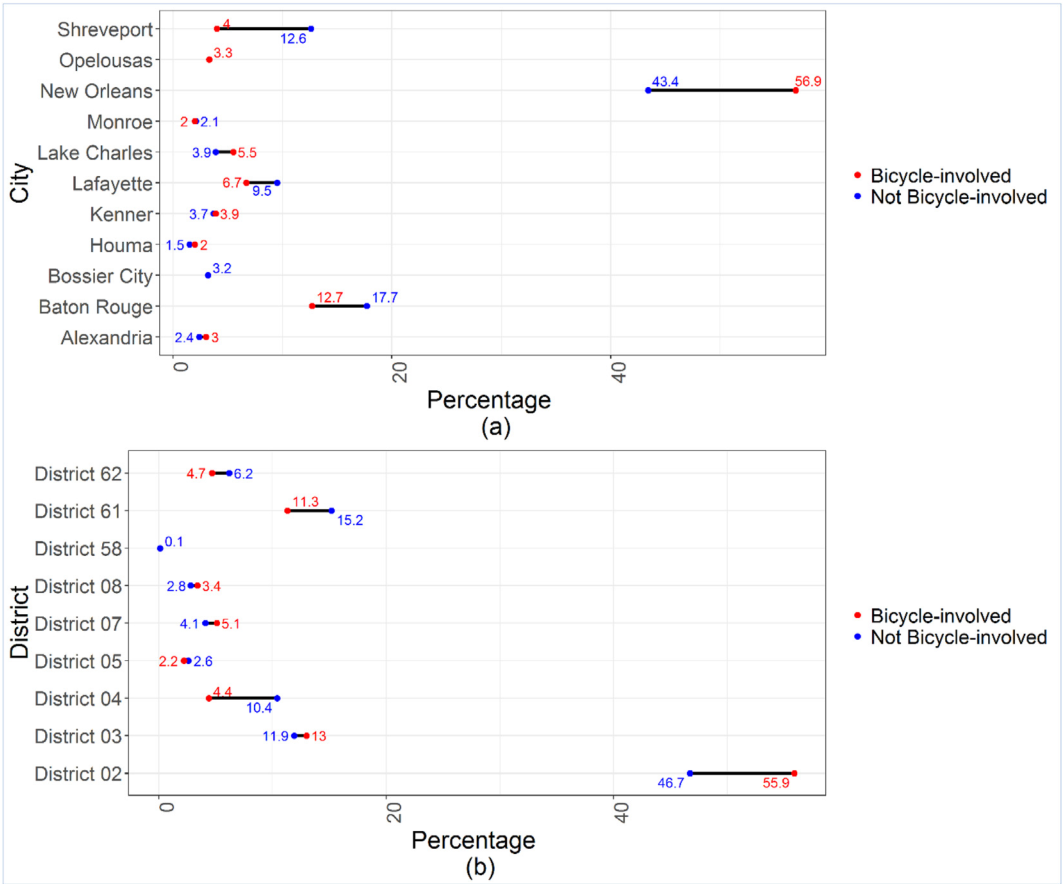
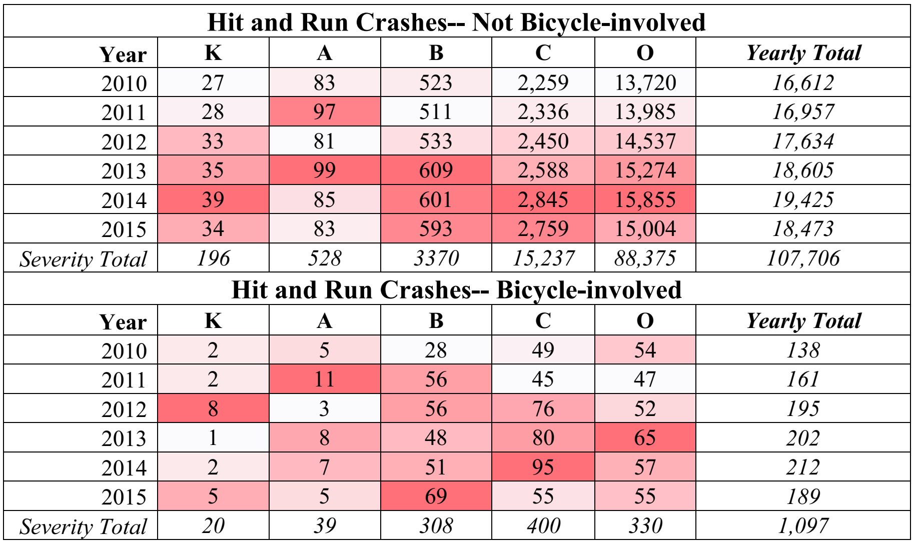


Fig. 3. (a) Percentage distribution of hit and run crashes in top 10 cities, (b) Percentage distribution of hit and run crashes in DOTD districts.



Hit and Run Crashes-- Not Bicycle-involved						
Year	**K**	**A**	**B**	**C**	**O**	***Yearly Total***
2010	27	83	523	2,259	13,720	*16,612*
2011	28	97	511	2,336	13,985	*16,957*
2012	33	81	533	2,450	14,537	*17,634*
2013	35	99	609	2,588	15,274	*18,605*
2014	39	85	601	2,845	15,855	*19,425*
2015	34	83	593	2,759	15,004	*18,473*
Severity Total	*196*	*528*	*3370*	*15,237*	*88,375*	*107,706*

Hit and Run Crashes-- Bicycle-involved						
Year	**K**	**A**	**B**	**C**	**O**	***Yearly Total***
2010	2	5	28	49	54	*138*
2011	2	11	56	45	47	*161*
2012	8	3	56	76	52	*195*
2013	1	8	48	80	65	*202*
2014	2	7	51	95	57	*212*
2015	5	5	69	55	55	*189*
Severity Total	*20*	*39*	*308*	*400*	*330*	*1,097*

Fig. 4. Yearly crash frequencies by severity types.

- Not bicycle-involved hit and run crashes increased by 11.2% in 2015 from 2010. For bicycle-involved crashes, the growth is 37.0%.
- Highest number of fatalities associated with bicycle-involved crashes happened in 2012. For non-bicycle related hit and run crashes, 2014 shows the highest frequency.

Our study focused on a wider net of crash, geometric, and environmental variables for the preliminary analysis. Variable selection is one of the key issues during the statistical modeling. Although fast and frugal decision tree can overcome issues like missing data and outlier problem, we conducted correlation test and Goodman Kruskal test (Pearson, 2018) to identify the correlation between both numerical and categorical factors. Goodman and Krustal's tau measure is the association between variables x and y is generally not the same as that between y and x. For example, two categorical variables, x and y, having M and N distinct values, we like to quantify the extent to which these variables vary or associate. This text uses Cramer's V measure (a normalized measure varies from 0 to 1) that can be defined as:

$$V = \sqrt{\frac{X^2}{N_{min}\{(M-1,N-1)\}}} \tag{1}$$

Several variables (for example, functional class) were dropped by observing the correlation outputs. Thirteen variables were selected for the final analysis. The original database comprised a total of 108,803 hit and run crashes. The final dataset for analysis included only 87,459 hit and run cases with all thirteen variables. Around 20% of the hit and run crashes were not included in the final analysis due to the missing data problem. As this study is more focused on understanding the real roadway environment that triggers a driver flee the scene of a bicycle involved hit and run crashes, missing value imputation was not applied in this study. Pearson's Chi-square test was performed to identify the variability of the attributes for each variable in both groups: (1) hit and run bicycle-involved crashes, and (2) hit and run not bicycle-involved crashes. Table 2 lists the frequency, percent distribution, and p-value (as a measure of significance) for each attribute for hit and run crashes. Based on the p-values, 12 variables were found to be statistically significant ($p \leq 0.05$). The findings are stated below:

- Bicycle-involved hit and run crashes contribute approximately 1% of total hit and run crashes. The bicycle-involved fatal crash percentage is ten times (around 10%) when compared with the percentage of all crashes.
- Right angle crashes are more likely to involve bicyclists. On the other hand, not bicycle-involved crashes are more inclined towards rear-end crashes. These findings are partially consistent with Tay et al. (2009) study.
- Many studies showed that nighttime has a major influence on crash occurrences (Aidoo et al., 2013; Bahrololoom et al., 2017; Lopez et al., 2017; Roshandeh et al., 2016; Solnick and Hemenway, 1995; Tay et al., 2010, 2009, 2008). Tay et al. (2008) mentioned that 'lighting condition is a crucial factor in determining whether the driver leaves the scene without reporting the crash or not. Inadequate lighting may encourage hit-and-run behavior because of the perceived lower probability of being identified.' In low light conditions, the perpetrators are more likely to flee (MacLeod et al., 2012; Solnick and Hemenway, 1995; Tay et al., 2008, 2009, 2010). In contrast, when the crash occurs in daylight, drivers may decide to remain at the scene because they realize that the chance of escaping detection is low (Solnick and Hemenway, 1995). Around 63% of bicycle involved hit and run crashes happened at nighttime (7 PM–6 AM). The lighting condition statistics also show that around 39% of the crashes happened during dark light condition and dawn/dusk.

Table 2
Chi-square test statistics for selected categorical variables.

Description	Not Bicycle-involved	Bicycle-involved	p-value	Description	Not Bicycle-involved	Bicycle-involved	p-value
Count	86,554	905		Alignment (%)			<0.001
Collision_Type (%)			<0.001	Straight	79,082 (91.4)	874 (96.6)	
Right Angle	9159 (10.6)	315 (34.8)		Curve	4479 (5.2)	14 (1.5)	
Rear End	30,047 (34.7)	129 (14.3)		Others	2993 (3.5)	17 (1.9)	
Head On	1980 (2.3)	32 (3.5)		Road_Type (%)			0.001
Sideswipe	27,034 (31.2)	246 (27.2)		One Way	12,558 (14.5)	164 (18.1)	
Turning	6195 (7.2)	93 (10.3)		Two Way Div.	28,621 (33.1)	259 (28.6)	
Single Vehicle	11,881 (13.7)	88 (9.7)		Two Way Undiv.	43,817 (50.6)	472 (52.2)	
Others	258 (0.3)	2 (0.2)		Others	1558 (1.8)	10 (1.1)	
Time_of_Day (%)			<0.001	Highway_Type (%)			<0.001
1 AM–6 AM	12,306 (14.2)	85 (9.4)		City Street	34,411 (39.8)	587 (64.9)	
7 AM–12 PM	17,791 (20.6)	161 (17.8)		Parish Road	9736 (11.2)	75 (8.3)	
1 PM–6 PM	31,501 (36.4)	343 (37.9)		Interstate/U.S. Hwy/State Hwy	41,202 (47.6)	225 (24.9)	
7 PM–12 AM	24,956 (28.8)	316 (34.9)		Others	1,205 (1.4)	18 (2.0)	
Day_of_Week = Weekend (%)	26,006 (30.0)	270 (29.8)	0.919	Locality (%)			<0.001
Season (%)			<0.001	Business	26,290 (30.4)	212 (23.4)	
Fall	21,532 (24.9)	245 (27.1)		Mixed	26,827 (31.0)	365 (40.3)	
Winter	20,835 (24.1)	173 (19.1)		Residential	22,563 (26.1)	281 (31.0)	
Spring	23,258 (26.9)	224 (24.8)		Others	10,874 (12.6)	47 (5.2)	
Summer	20,929 (24.2)	263 (29.1)		Lighting (%)			0.002
Intersection = Yes (%)	31,649 (36.6)	480 (53.0)	<0.001	Daylight	49,282 (56.9)	539 (59.6)	
Access_Control (%)			<0.001	Dark	33,013 (38.1)	322 (35.6)	
Full Control	10,172 (11.8)	21 (2.3)		Dawn/Dusk	1965 (2.3)	32 (3.5)	
Partial Control	6471 (7.5)	62 (6.9)		Others	2294 (2.7)	12 (1.3)	
No Control	68,867 (79.6)	805 (89.0)		Weather (%)			<0.001
Others	1044 (1.2)	17 (1.9)		Clear	63,957 (73.9)	733 (81.0)	
Severity = PDO (%)	69,450 (80.2)	266 (29.4)	<0.001	Rain/Cloudy/Fog	20,707 (23.9)	155 (17.1)	
				Others	1890 (2.2)	17 (1.9)	

- Two way undivided roadways show higher propensity towards bicycle-involved hit and run crashes. Non bicycle-involved crashes are also overrepresented on two way undivided roadways. This finding contradicts with the findings of Tay et al. (2008), which showed that crashes on undivided roads decreased the probability of hit-and-run when compared to crashes on one-way roads. Divided highways without traffic barriers and divided highways with median strips and two-way continuous left-turn lanes are believed to reduce hit-and-run behavior in fatal crashes compared with undivided and one-way streets (Zhang et al., 2014).
- Our study showed that weekend is not statistically significant in differentiating between bicycle-involved and not bicycle-involved hit and run crashes. Our findings are similar to three studies (Tay et al., 2008, 2009; Zhang et al., 2014). However, other studies showed that probability of hit and run behavior in fatal crashes is higher on weekends than weekdays (MacLeod et al., 2012; Solnick and Hemenway, 1995; Tay et al., 2009).
- The results show that – in line with the literature (Solnick and Hemenway, 1995) – fall and summer are dominating attributes in bicycle-involved hit and run crashes.
- Inclement weather does not contribute much in hit and run crashes for both groups. Other studies also showed that weather conditions have no evident effect on hit and run crashes (Tay et al., 2008, 2010; Zhang et al., 2014).
- The present study adds to the literature by finding that residential and mixed (business and residential) locations show higher propensity towards bicycle-involved hit and run crashes.
- Crash severity patterns are significantly different in bicycle-involved and not bicycle-involved hit and run crashes. Around 70% of the bicycle-involved crashes are KABC. On the other hand, this percentage is 20% for non bicycle-involved group. This is an obvious finding. As bicyclists are the vulnerable road user groups, fatalities and injuries are more dominant in bicycle-involved hit and run crashes.
- An increased risk of hit and run incidents is associated with urban locations (MacLeod et al., 2012; Solnick and Hemenway, 1995). Our study showed that 65% of bicycle-involved crashes occurred on city streets.
- Our study showed that roadways with no access control are overrepresented in both groups. This finding is in line with Tay et al. (2009) study, which stated that 'crashes occurring at locations with properly or improperly functioning traffic control devices are less likely to result in hit-and-run than crashes at locations where the traffic control devices are not functioning'.

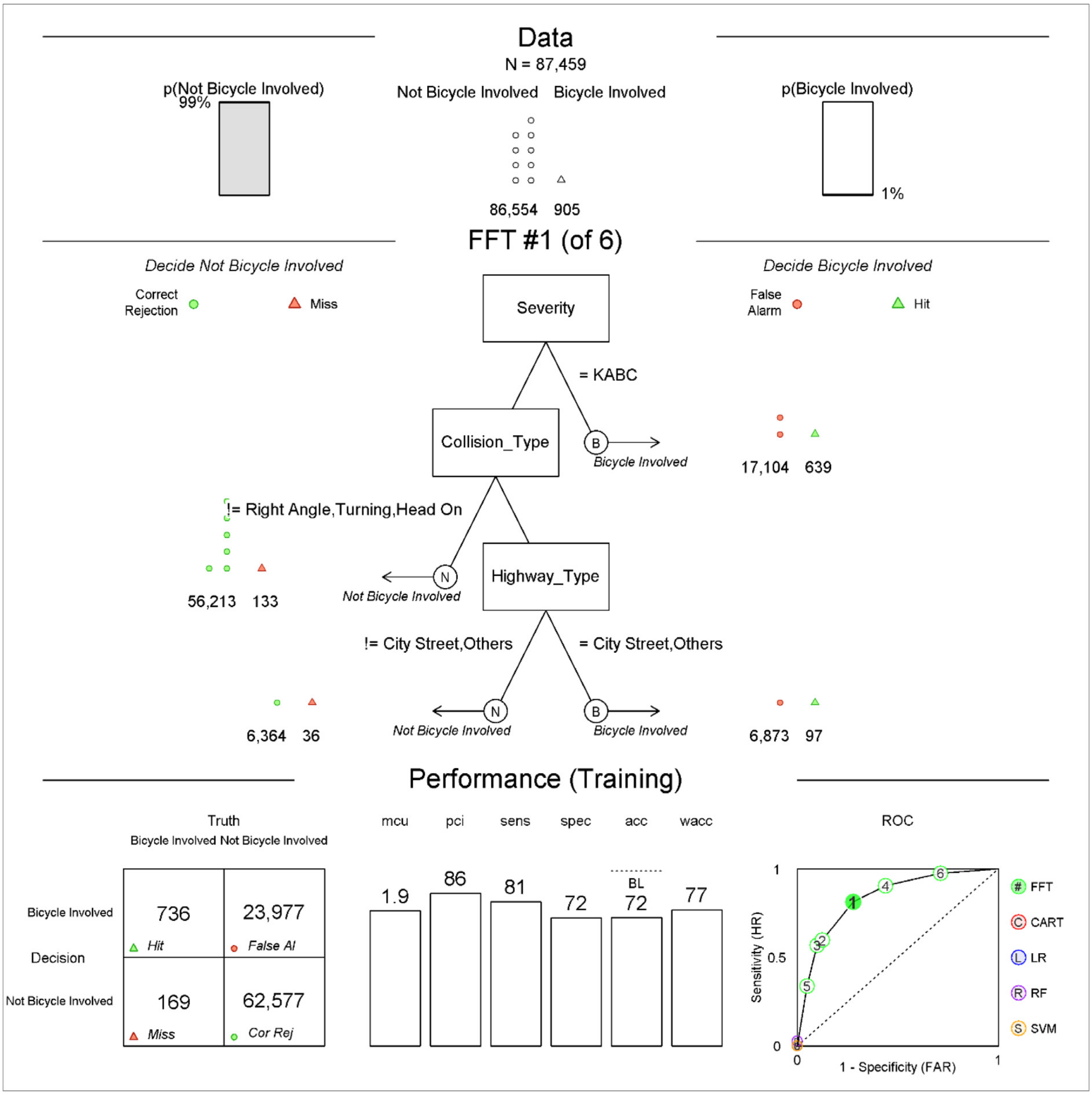


Fig. 5. First and frugal tree outputs by considering all data as a training group.

- Consistent with Tay et al. (2009) study, our study found that curve aligned roadways are not skewed towards hit and run crashes.

6. Model development

The final dataset used for FFT analysis consists of 87,459 hit and run crashes with 13 variables. The preliminary model involves the whole dataset as the training set to develop the model (as shown in Fig. 5). To evaluate the performance of the model, both training (a dataset with 70% from bike-involved crashes, and 70% from not bike-involved crashes), and testing (the rest 30% from each group) was used in the final analysis. The research team extensively used R package ‘FFTrees’ to perform the analysis and data visualization (Phillips et al., 2017). As this study aims to determine the decision factors associated with a hit and run bicycle crash occurrence, severity is considered as the response variable. This is also justified for two other findings: (1) severity is found as the most significant variable in the variable importance assertion, and (2) descriptive statistics show that 70% of the hit and run bicycle crashes involved some sort of injury. Fig. 6(a) and (b) illustrate the outputs of FFTs by using training and testing data respectively. This study applied random data selection procedure for

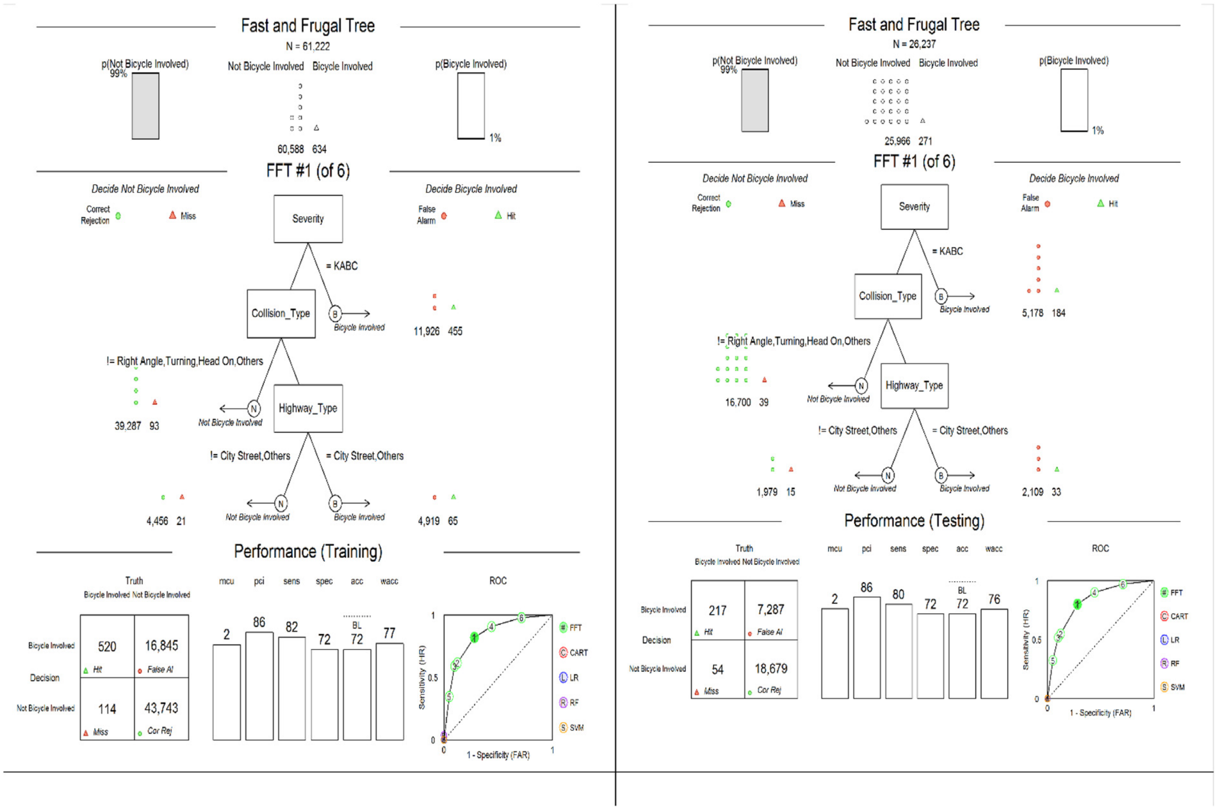


Fig. 6. (a) First and frugal tree outputs for the training data (70% of randomly selected data), (b) First and frugal tree outputs for the testing data (rest 30% data, which are not used for model development.)

the training and test datasets. A total of 20 different training and test sets were selected. This study applied FFT algorithms on all 20 training and 20 test set data. The outputs shown in this study are the mean trend of the similar outputs. The output visualization has three basic parts: top part, middle part, and bottom part.

- The top panel shows information about the dataset, including the frequencies and percentage distributions of bicycle-involved and not bicycle-involved cases. For example, top part of Fig. 5 indicates that the sample size of the dataset is 87,549. The count of bicycle-involved hit and run crashes is 905, which represents 1% of the total hit and run crashes.
- The middle part contains the FFT and icon arrays showing the count and confusion table matrices at each node. It is interesting that FFTs developed for each cases (complete data as training data, randomly selected 70% data as training data, and randomly selected 30% data as test data) are same. It indicates that three major thresholds can identify the key measures for bicycle-involvement in hit and run crashes: KABC severity, right-angle/turning/head on collisions, and city streets/others.
- The bottom part shows the FFT's performance in receiver operating characteristics (ROC) curve, confusion matrix, and levels for a range of statistics.

A confusion matrix table provides a simple representation of the model's accuracy and the types of errors the model makes. The values in cells *tp* and *tn* infer the correct prediction (true positive and true negative), whereas counts in cells *fp* and *fn* refer to errors (false positive and false negative). The FFT algorithm is designed such a way that it targets to maximize frequencies in cells *tp* and *tn* while minimizing those in cells *fp* and *fn*. This study focused on five important measures: sensitivity (sens), specificity (spec), overall accuracy (acc), weighted accuracy, and balanced accuracy (bacc). These terms can be expressed as follows:

$$Sensitivity\ \ (sens) = \frac{tp}{tp + fn} \tag{1}$$

$$Specificity\ \ (spec) = \frac{tn}{tn + fp} \tag{2}$$

$$Accuracy\ \ (acc) = \frac{tp + tn}{tp + fp + fn + tn} \tag{3}$$

$$Weighted\ \ ccuracy\ \ (wacc) = \frac{tp}{tp + fn} \times w + \frac{tn}{tn + fp} \times (1 - w) = sens \times w + spec \times (1 - w) \tag{4}$$

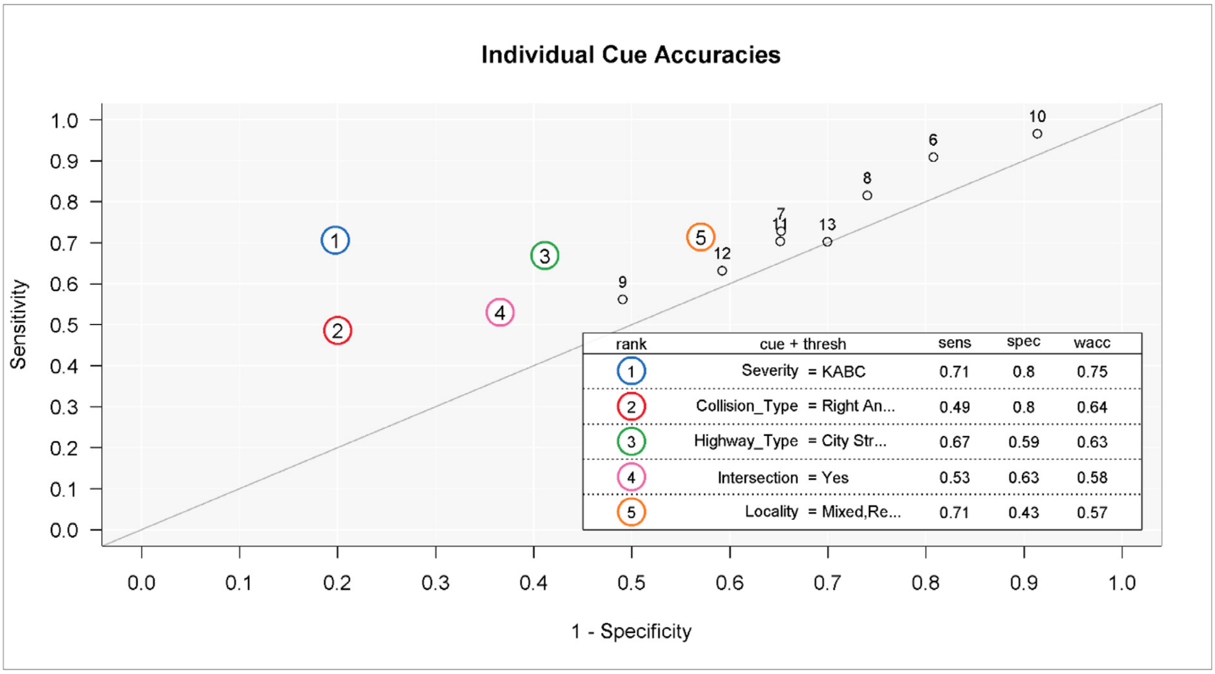
$$Balanced\ \ Accuracy\ \ (bacc) = \left(\frac{tp}{tp + fn} + \frac{tn}{tn + fp}\right) \times 0.50 = \frac{sens + spec}{2} \tag{5}$$

where,
tp = true positive (correct hit)
fn = false negative (miss)
tn = true negative (correct rejection)
fp = false negative (false alarm)
w = weighting factor

Sens represents the percentage of cases with positive criterion values that are correctly predicted by the algorithm. Similarly, *spec* infers the percentage of cases with negative criterion values correctly predicted by the algorithm. The next three measures define accuracy across all cases. *Accuracy* (*acc*) is defined as the overall percentage of correct decisions by ignoring the difference between hits and correct rejections. To make a balance between sensitivity and specificity, one can use the measure named as *weighted accuracy* (*wacc*). This measure depends on a weighting factor, which ranges in between 0 and 1. In cases, when sensitivity is more important than specificity, one can consider the weighting factor above 0.50. For an ideal balance, one can use 0.50, which can be termed as *bacc*.

The *ifan* algorithm explicitly selects and ranks cues (threshold of variable category or category-groups) accuracies. Visualizing marginal cue helps in understanding the ranking of each cues in terms of *balanced accuracy* (*bacc*). The top five cues are colored and described in the legend. All other cues in the data are shown as black points. Fig. 7 shows the resulting plot for the training data (70% of all hit and run crashes). Inspecting the graph reveals that the three cues (severity, collision type, and highway type) used in FFT #1 have the highest individual wacc values. Fig. 7 also shows that the two next best cues are intersection, and locality. This information can be useful in guiding a top-down process of future FFT construction. Table 3 shows *bacc* values for all 13 variables. Day of the week shows lowest *bacc* values, which is in line with the findings of chi-square test.

A useful way to present the characteristics of a diagnostic test is the ROC curve. A ROC curve illustrates the trade-off between *sens*and *spec* in classification algorithms. As *sens* increases, *spec* decreases (i.e., 1 – *spec* increases). Ideal performance (*bacc* = 1.0), is represented by the cross in the upper-left corner. The bottom right plot shows the performance of



rank	cue + thresh	sens	spec	wacc
1	Severity = KABC	0.71	0.8	0.75
2	Collision_Type = Right An...	0.49	0.8	0.64
3	Highway_Type = City Str...	0.67	0.59	0.63
4	Intersection = Yes	0.53	0.63	0.58
5	Locality = Mixed,Re...	0.71	0.43	0.57

Fig. 7. Top five cue (threshold of variable attribute groups) accuracies.

Table 3
Accuracies of the variable thresholds.

Variables	Threshold	Direction	bacc
Severity	KABC	=	0.7542
Collision_Type	Right Angle, Turning, Head On	=	0.6430
Highway_Type	City Street, Others	=	0.6285
Intersection	Yes	=	0.5824
Locality	Mixed, Residential	=	0.5716
Access_Control	No Control, Others	=	0.5503
Time_of_Day	7 PM–12 AM, 1 PM–6 PM	=	0.5380
Weather	Clear, Not Reported	=	0.5377
Season	Summer, Fall	=	0.5354
Alignment	Straight	=	0.5260
Road_Type	One Way, Two Way Undiv.	=	0.5257
Lighting	Daylight, Dawn/Dusk	=	0.5194
Day-of_Week	Weekday	=	0.5041

Table 4
FFT outputs.

Dataset	% of Complete Dataset	Count of Hit and Run Crashes	Count of Bicycle-involved Hit and Run Crashes	Count of Not Bicycle-involved Hit and Run Crashes	tp	fn	tn	fp	sens	spec	acc	bacc
Complete	100%	87,459	905	86,554	736	169	62,577	23,977	81.33	72.30	72.39	76.82
Training	70%	61,222	634	60,588	520	114	43,743	16,845	82.02	72.20	72.30	77.11
Testing	30%	26,237	271	25,966	217	54	18,670	7287	80.07	71.93	72.01	76.00

all FFTs in ROC space (green circles with numbers correspond to FFTs). FFT #1 has the highest weighted accuracy (colored in solid green). Additional points in this plot correspond to the performance of competing classification: standard decision trees (CART), logistic regression (LR), random forests (RF), and support vector machines (SVM). This study examines the potential of using an optimized method to better interpret the results. The current study does not provide discussions on other machine learning models as it is not in the scope of the analysis. These machine learning models are used here to justify the strength of a simple algorithm like FFT. It is important to note that the application of machine learning encounters hurdles to interpret a machine learning model or to explain the outcomes. FFT has an advantage over machine learning models due to its easy interpretability. In our study, FFT #1 has a higher *sens* than other algorithms, but at the cost of a lower *spec*.

Table 4 lists the model comparison results for three datasets: complete, training, and testing. It shows that the measure parameters (*sens*, *spec*, *acc*, and *bacc*) are almost identical for complete and training data. The balanced accuracy is higher for complete, and training dataset. For testing data, the accuracy is promising (around 76%). Our study showed that a simplified FFT can perform better than complex machine learning models (CART, RF, and SVM) in terms of *sens*. As FFT algorithm has exit branch on every node, they typically make data representation faster than standard decision trees while simultaneously being easier to understand and use than other machine learning black box approaches.

7. Conclusions

The issue of cyclist safety is crucial. Hit and run crashes refer to crashes where at-fault drivers flee crash scenes without reporting the incidents to the emergency response services or similar authorities. Due to lack of immediate emergency medical assistance, hit and run crashes could significantly increase the likelihood of serious injury and even fatality for vulnerable roadway users like bicyclists. Fleeing from the crash location is more likely to depend on the situational factors surrounding the crash. To reduce the occurrences of these crashes, it is crucial to identify appropriate associations. We used all hit and run crashes in Louisiana for six years (2010–2015) and particularly identifies patterns associated with bicycle involvements.

Many factors affect the instantaneous decision to leave the scene of a crash. In the current study, various factors which might been associated with hit and run decisions were taken into account, including collision type, day of week, time of day, season, crash severity, roadway type, locality, highway type, alignment, access control, lighting, intersection presence, and weather. It was found out that variables contributing to hit and run crashes varied based on the involvement of bicycles in the crashes. Results from FFT heuristics revealed that crash severity, collision type, and highway type were the most important predictors of bicycle involvement in hit and run crashes. Crash severity (KABC) shows the most influential factor that is associated with the bicycle-involved hit and run crashes. Next most strongly associated with the bicycle-involved hit and run crashes is collision types in relation to right-angle/turning/head on collisions. The third influential cue is city street/others. These findings are in line with findings from earlier studies (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 and Hemenway, 1995; Tay et al., 2010, 2009, 2008; Zhang et al., 2014). The present study adds to the literature by finding that residential and mixed (both business and residential) localities are over-represented in the bicycle-involved hit and run crashes. Surprisingly, weekends, dark lighting condition, and roadway types (e.g., two way undivided roadways) were not determined as key contributing factors in the bicycle-involved hit and run crashes. However, variabilities of these attributes in the two groups (bicycle-involved and not bicycle-involved) were statistically significant in the preliminary analysis. Our model perform better than complex machine learning models regarding $sensitivity$. Findings from our study will provide valuable insights for hit and run bicycle crash reduction in both planning and operation levels.

Our study is subject to several limitations that call for future studies. One problem is that the current study focused largely on reported hit and run cases for which there was complete information on geometric and environmental factors. Future studies can incorporate additional variables like population and demographic characteristics, unbiased driver information, and real-time driver behavior. However, the contribution of built roadway environment during the crash occurrence is highly associated with hit and run crash occurrences. It is important for the safety professionals to know at what roadway environment a driver flees a hit and run scene involved with a bicyclist. The major contribution of this study is to investigate the tipping point of the hit and run bicycle crash occurrence. The findings of this study can help the safety researchers to design the environment suitable for reducing hit and run bicycle crashes. Additionally, more improvements and enhancements in the FFT algorithms can be done by extending the tree structures. Notwithstanding these limitations, our study added a simplified decision mechanism (with high accuracies) to identify bicycle-involved hit and run crashes instead of incorporating more complex and black box machine learning algorithms. Multiple correspondence analysis (MCA) could be a useful tool in exploring the association patterns in the complex dataset of hit and run crashes. MCA with missing data handling algorithm would be effective in reducing the biases associated with the missing driver information. One such study by the authors is currently in progress.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijtst.2018.11.001.

References

Aidoo, E.N., Amoh-Gyimah, R., Ackaah, W., 2013. The effect of road and environmental characteristics on pedestrian hit-and-run accidents in Ghana. Accid. Anal. Prev. 53, 23–27. https://doi.org/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. Eur. Transp. Res. Rev. 9 (3), 40. https://doi.org/10.1007/s12544-017-0257-5.

Bahrololoom, S., Moridpour, S., Tay, R., Young, W., 2017. Factors affecting hit and run bicycle crashes in Victoria, Australia. Presented at the Australasian Road Safety Conference, 2017, Perth, Western Australia, Australia.

Brooks, B., 2008. Shifting the focus of strategic occupational injury prevention: mining free-text, workers compensation claims data. Saf. Sci. 46 (1), 1–21. https://doi.org/10.1016/j.ssci.2006.09.006.

Brown, D.E., 2016. Text mining the contributors to rail accidents. IEEE Trans. Intell. Transp. Syst. 17 (2), 346–355. https://doi.org/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. Accid. Anal. Prev. 90 (Supplement C), 128–139. https://doi.org/10.1016/j.aap.2016.02.011.

Chung, Y.-S., 2013. Factor complexity of crash occurrence: an empirical demonstration using boosted regression trees. Accid. Anal. Prev. Emerg. Res. Methods Appl. Road Saf. 61, 107–118. https://doi.org/10.1016/j.aap.2012.08.015.

Das, S., Sun, X., 2014. Investigating the pattern of traffic crashes under rainy weather by association rules in data mining. In: Transportation Research Board 93rd Annual Meeting. Transportation Research Board, Washington D.C..

Das, S., Avelar, R., Dixon, K., Sun, X., 2018a. Investigation on the wrong way driving crash patterns using multiple correspondence analysis. Accid. Anal. Prev. 111, 43–55. https://doi.org/10.1016/j.aap.2017.11.016.

Das, S., Brimley, B., Lindheimer, T., Pant, A, 2017a. Safety impacts of reduced visibility in inclement weather. Report No. ATLAS-2017-19.

Das, S., Dutta, Anandi, Dixon, K., Minjares-Kyle, L., Gillette, G., 2018b. Using deep learning in severity analysis of at-fault motorcycle rider crashes. Transp. Res. Rec. J. Transp. Res. Board.

Das, S., Dutta, A., Avelar, R., Dixon, K., Sun, X., Jalayer, M., 2018c. Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. Int. J. Urban Sci., 1–19 https://doi.org/10.1080/12265934.2018.1431146.

Das, S., Dutta, A., Jalayer, M., Bibeka, A., Wu, L., 2018d. Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: Exploration using 'Eclat' association rules to promote safety. Int. J. Transp. Sci. Technol. 7 (2), 114–123. https://doi.org/10.1016/j.ijtst.2018.02.001.

Das, S., Minjares-Kyle, L., Avelar, R., Dixon, K., Bommanayakanahalli, B., 2017b. Improper passing-related crashes on rural roadways: using association rules negative binomial miner. In: Transportation Research Board 96th Annual Meeting. Transportation Research Board, Washington D.C.

Das, S., Sun, X., 2016. Association knowledge for fatal run-off-road crashes by multiple correspondence analysis. IATSS Res. 39 (2), 146–155. https://doi.org/10.1016/j.iatssr.2015.07.001.

Das, S., Sun, X., 2015. Factor association with multiple correspondence analysis in vehicle-pedestrian crashes. Transp. Res. Rec. J. Transp. Res. Board 2519, 95–103. https://doi.org/10.3141/2519-11.

Figueira, A. da 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 Stud. Transp. Policy 5 (2), 200–207. https://doi.org/10.1016/j.cstp.2017.02.004.

Gao, L., Wu, H., 2013. Verb-based text mining of road crash report. In: Transportation Research Board 92nd Annual Meeting. Transportation Research Board, Washington D.C..

Geurts, K., Thomas, I., Wets, G., 2005. Understanding spatial concentrations of road accidents using frequent item sets. Accid. Anal. Prev. 37 (4), 787–799. https://doi.org/10.1016/j.aap.2005.03.023.

Gilbert, X., Patel, V.M., Chellappa, R., 2017. Deep multitask learning for railway track inspection. IEEE Trans. Intell. Transp. Syst. 18 (1), 153–164.

Haworth, N., Debnath, A.K., 2013. How similar are two-unit bicycle and motorcycle crashes? Accid. Anal. Prev. 58 (Supplement C), 15–25. https://doi.org/10.1016/j.aap.2013.04.014.

Jalayer, M., Pour-Rouholamin, M., Zhou, H., 2018. Wrong-way driving crashes: a multiple correspondence approach to identify contributing factors. Traffic Inj. Prev. 19 (1), 35–41. https://doi.org/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. Accid. Anal. Prev. Traffic Saf. China: Challenges Countermeasures 95, 373–380. https://doi.org/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. Transp. Res. Part C Emerg. Technol. Special Issue Road Saf. Simulat. 50, 86–96. https://doi.org/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. Transp. Res. Rec. J. Transp. Res. Board 2083, 114–121. https://doi.org/10.3141/2083-13.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using Support Vector Machine models. Accid. Anal. Prev. 40 (4), 1611–1618. https://doi.org/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. J. Transp. Health. https://doi.org/10.1016/j.jth.2017.10.005.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part Policy Pract. 44 (5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001.

Luan, S., Schooler, L.J., Gigerenzer, G., 2011. A signal-detection analysis of fast-and-frugal trees. Psychol. Rev. 118 (2), 316.

MacLeod, K.E., Griswold, J.B., Arnold, L.S., Ragland, D.R., 2012. Factors associated with hit-and-run pedestrian fatalities and driver identification. Accid. Anal. Prev. 45, 366–372. https://doi.org/10.1016/j.aap.2011.08.001.

Martignon, L., Hoffrage, U., 2002. Fast, frugal, and fit: simple heuristics for paired comparison. Theory Decis. 52 (1), 29–71. https://doi.org/10.1023/A:1015516217425.

Martignon, L., Katsikopoulos, K.V., Woike, J.K., 2008. Categorization with limited resources: a family of simple heuristics. J. Math. Psychol. 52 (6), 352–361. https://doi.org/10.1016/j.jmp.2008.04.003.

Martignon, L., Vitouch, O., Takezawa, M., Forster, M.R., 2003. Naive and yet enlightened: from natural frequencies to fast and frugal decision trees. Think. Psychol. Perspect. Reason. Judgm. Decis. Mak., 189–211

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. Technol. Forecast. Soc. Change 111 (Supplement C), 86–96. https://doi.org/10.1016/j.techfore.2016.06.010.

Pearson, R., 2018. GoodmanKruskal: association analysis for categorical variables. R package version 0.0.2. https://CRAN.R-project.org/package=GoodmanKruskal.

Phillips, N.D., Neth, H., Woike, J.K., Gaissmaier, W., 2017. FFTrees: a toolbox to create, visualize, and evaluate fast-and-frugal decision trees. Judgm. Decis. Mak. 12 (4), 344.

Roshandeh, A.M., Zhou, B., Behnood, A., 2016. Comparison of contributing factors in hit-and-run crashes with distracted and non-distracted drivers. Transp. Res. Part F Traffic Psychol. Behav. 38, 22–28. https://doi.org/10.1016/j.trf.2015.12.016.

Safety Administration, N.H.T., 2017. Traffic Safety Facts (2015 Data): Bicyclists and Other Cyclists.

Saha, D., Alluri, P., Gan, A., 2015. Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. Accid. Anal. Prev. 79, 133–144. https://doi.org/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. Accid. Anal. Prev. 43 (5), 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025.

Solnick, S.J., Hemenway, D., 1995. The hit-and-run in fatal pedestrian accidents: victims, circumstances and drivers. Accid. Anal. Prev. 27 (5), 643–649. https://doi.org/10.1016/0001-4575(95)00012-O.

Solnick, S.J., Hemenway, D., 1994. Hit the bottle and run: the role of alcohol in hit-and-run pedestrian fatalities. J. Stud. Alcohol 55 (6), 679–684.

Sun, X., Das, S., Broussard, N., 2016. Developing crash models with supporting vector machine for urban transportation planning. Presented at the 17th International Conference Road Safety On Five Continents (RS5C 2016), Rio de Janeiro, Brazil, 17–19 May 2016. Statens väg- och transportforskningsinstitut.

Tay, R., Barua, U., Kattan, L., 2009. Factors contributing to hit-and-run in fatal crashes. Accid. Anal. Prev. 41 (2), 227–233. https://doi.org/10.1016/j.aap.2008.11.002.

Tay, R., Kattan, L., Sun, H., 2010. Logistic model of hit and run crashes in Calgary. Can. J. Transp. 4, 1.

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. Accid. Anal. Prev. 40 (4), 1330–1336. https://doi.org/10.1016/j.aap.2008.02.003.

Turner, S., Sener, I., Martin, M., Das, S., Shipp, E., Hampshire, R., Fitzpatrick, K., Molnar, L., Wijesundera, R., Colety, M., Robinson, S., 2017. Synthesis of Methods for Estimating Pedestrian and Bicyclist Exposure to Risk at Areawide Levels and on Specific Transportation Facilities (Literature Review No. FHWA-SA-17-041), FHWA Report.

Weng, J., Zhu, J.-Z., Yan, X., Liu, Z., 2016. Investigation of work zone crash casualty patterns using association rules. Accid. Anal. Prev. 92 (Supplement C), 43–52. 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. Transp. Res. Part F Traffic Psychol. Behav. 23, 113–124. https://doi.org/10.1016/j.trf.2013.12.009.

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 Eng Green Intelligent Transp. Syst. Saf. 137, 554–562. https://doi.org/10.1016/j.proeng.2016.01.292.