Accident Analysis and Prevention 144 (2020) 105620

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Understanding speeding behavior from naturalistic driving data: Applying classification based association rule mining

Xiaoqiang Kong (Jack)[a,*], Subasish Das[b], Kartikeya Jha[b], Yunlong Zhang[a]

[a] Texas A&M University, 3136 TAMU, College Station, TX 77843, United States
[b] Texas A&M Transportation Institute, 1111 RELLIS Parkway, Bryan, TX, 77807, United States

ARTICLE INFO

Keywords:
Trip features
Driving characteristics
Geometric features
Speeding duration
Speeding pattern
Association rules

ABSTRACT

Speeding is considered as one of the most significant contributing factors to severe traffic crashes. Understanding the associations between trip/driving/roadways features and speeding behavior is crucial for both researchers and practitioners. This research utilized naturalistic driving data collected by the Safety Pilot Model Deployment (SPMD) program and roadway features from a road inventory dataset - Highway Performance Monitoring System (HPMS), provided by the United States Department of Transportation (USDOT), to investigate the hidden rules that associated trip/driving/roadway features with speeding behavior. A classification-based association (CBA) algorithm was adopted to explore the hidden rules from two perspectives of speeding: speeding duration and speeding pattern. Results indicate that the combinations of longer trips (more than 60 min), driving on the roadways with a relatively higher functional class are highly associated with longer speeding events (speeding longer than 2 min). The moderate speeding events (speeding longer than 2 min and longer than 30 s) are found highly associated with the combination of driving on roadways with lower functional class, absence of a median and relatively short trip time (less than 30 min). The research also found the combinations of driving on roadways with relatively lower functional class, experienced congestion before a speeding event, and the presence of a median is a leading cause that triggers a higher speeding pattern (speeding more than 5mph above the speed limit). Furthermore, the moderate speeding pattern (speeding more than 1mph above the speed limit and less than 5mph of the speed limit) is associated with the combinations of factors like experiencing congestion before a speed event, driving on roadways with higher functional class and a relatively shorter trip (less than 30 min). The findings can help practitioners understand the composite effect of these factors more comprehensively and provide corresponding countermeasures to mitigate the negative consequences of speeding wherever possible. These can also help in calibrating driver behavior parameters for transportation-related simulation tools.

1. Introduction

The National Highway Traffic Safety Administration reported that 52,274 drivers were involved in 34,247 fatal crashes resulting in 37,133 fatalities in the year 2017, and 17 % of the drivers involved at crashes were speeding, and about 26 % of those killed were in a crash involving at least one speeding driver (National Center for Statistics and Analysis, 2019). Speeding could reduce drivers' ability to control the vehicle by increasing the stopping distance after the driver wants to stop the vehicle and could increase the injury severity if the crash occurred (National Highway Traffic Safety Administration, 2020). Elvik published a research in 2008 which suggested that around 1/3 to 1/4 fatal crashes are speeding related (Elvik, 2008). These statistics highlight a pressing need for an in-depth examination of potential contributing factors that lead to speeding events and underline the importance of understanding the driver's speeding behavior. This research utilized GPS based trajectory data and road inventory data to explore the associations between trip/driving/roadway features and speeding behavior from two perspectives: speeding duration (how long the speeding event lasts) and speeding pattern (how much does this speeding event exceed the speed limit). It examined the individual feature or a combination of features that might encourage drivers to speed for a relatively longer duration or a higher speed by employing a classification and association rule algorithm. The findings could provide transportation engineers, law enforcement and transportation agencies more understanding of the speeding behavior and offer supports for corresponding countermeasures.

The naturalistic driving data are extensively analyzed to understand

* Corresponding author.
E-mail address: jackxqkong@gmail.com (X. Kong).

https://doi.org/10.1016/j.aap.2020.105620
Received 27 October 2019; Received in revised form 11 March 2020; Accepted 26 May 2020

the relationship between speeding behavior and factors like roadway environment, driver demographic and personality factors (Avrenli et al., 2013; Chevalier et al., 2017, 2016; Reagan et al., 2013; Richard et al., 2012). This research utilized a naturalistic trajectory dataset provided by the Safety Pilot Model Deployment (SPMD) program, which is a part of the Connected Vehicle Safety Pilot Program, contains detailed information about trips and driving characteristics. Data for each trip consists of points with spatial information at a centiseconds level. The geospatial information (latitude and longitude) of each point is used to conflate its corresponding roadway features. With information on speed limit from road inventory data, collected from the Highway Performance Monitoring System (HPMS), speeding behavior can be identified.

Existing studies have identified a list of factors that might associate with the speeding behavior from various perspectives like demographic, personality, roadway environment and situational, such as speed loss caused by the congestion (Chevalier et al., 2017; National Highway Traffic Safety Administration, 2020; Richard et al., 2013, 2012). This research investigated the associations rules of speeding behavior from three perspectives: a) trip features like total travel time and whether driving during peak hour; b) driving features like tailgating other vehicles and speed loss before a speeding event; c) roadway features like functional class, access control, shoulder width, lane width, median width, speed limit, median type, and Annual Average Daily Traffic (AADT). Instead of identifying a single feature that may associate with speeding behavior, the association rule technique mines a set of features that may associate with speeding behavior as a combination. Presumably, the speeding behavior is more likely triggered by a set of features comparing with one single feature. Moreover, the association rule of the speeding behavior was investigated from two perspectives: speeding duration and speeding pattern. Speeding duration has been classified as a moderate speeding duration and longer speeding duration. Additionally, the speeding pattern is classified as a moderate speeding pattern and a higher speeding pattern.

This research applied the classification-based association (CBA) algorithm (Liu et al., 1998). To authors' knowledge, this is the first of its kind application of the CBA algorithm to this area of research on speeding behavior. One advantage of this algorithm is that this non-parametric method avoids making any parametric assumptions, which are normally subjective and sensitive to the correctness of the underlying hypothesis. Another advantage of applying this algorithm is that it can generate rule sets for two classes in each category. For example, it generates rule sets for two speeding duration classes: moderate speeding duration and longer speeding duration. These classification-based rule sets provide not only ways to understand the pattern of speeding behavior, but also generate in-depth insights of speeding behavior by comparing rule sets between two classes.

The paper will elaborate on the existing researches on the speeding behavior issue, methodology, data description, followed by results discussion and conclusion.

2. Earlier work and research context

The exploration of the associations between roadway geometry features and speeding behavior has brought attention to transportation researchers for years. Many existing researches found significant effects of road and shoulder width, section length and extent of road markings on driver behavior. Road markings along both the centerline and the shoulders gave a sense of added safety to drivers and encouraged higher speed choice. Presence of woodland (vegetation or shrubbery) and intersections also influenced speed choice along with other contributing factors such as gender, time of day and age of the vehicle (Andersen et al., 2016; Ben-Bassat and Shinar, 2011; Liu et al., 2016; Lobo et al., 2018). Bassani et al. (2014) analyzed urban arterial and collector road sections instead of rural roads and found that the presence of both right and left shoulders resulted in higher speed dispersion, as did the presence of a dedicated bus and taxi lane adjacent to the traveled way. On the other hand, the presence of sidewalk and pedestrian crossing significantly reduced speed dispersion. The study concluded that speed reduction is influenced more by transversal geometric characteristics than longitudinal characteristics. Lane position and number of traveled ways (lanes) were found to have the most significant effect on mean speed (Bassani et al., 2014).

Besides the road geometry, the speed limit could also affect speeding behavior. There are several findings from previous studies. First, high posted speed limits are highly associated with moderate speed limit violations compare to minor or major speed limit violations (Afghari et al., 2018). They also pointed out that the reason could be that road segments with higher speed post speed limit normally exhibit higher standards such as high-quality pavements, wide and paved shoulders, etc., which might facilitate moderate speeding. Moreover, the research pointed out that major speed limit violations are highly associated with road segments with a divided median and higher functional class. Second, several studies have also highlighted the tendency in drivers to disregard posted speed limits to varying extents depending on the road environment. While Fildes et al. (1991) found in a study conducted in Australia that a major portion of drivers did not consider driving 30 kph above the posted limit to be risky in both rural and urban settings, Eksler et al. (2009) found in Scandinavia that more than half of the driving population violate the speed limit by more than 10 kph on rural roads 2009). Varhelyi (1997) concluded that in general, drivers tend to underestimate the risks of speeding, and base their speed choice primarily on their own assessment of the road and traffic environment.

There are existing studies on investigating relationships between speeding behavior and geometric features, speed limits, demographic and other factors. However, the trigger of speeding behavior might not be just one individual factor, rather a combination of factors. It is rare to find related research. Moreover, most of the research only considered speeding behavior as a whole, rather than investigating speeding behavior from more specific perspectives: speeding duration and speeding pattern/severity.

3. Methodology

3.1. Speeding event definition

It is critical to select reasonable thresholds for defining a speeding event in the naturalist driving dataset. Two important thresholds are considered while identifying the speeding event: the minimum speed threshold to be considered as speeding and the minimum speeding duration. In this dataset, the speeding event is identified while the average speed of a continuous driving period is higher than the minimum speeding threshold and speeding time is longer than the minimum speeding duration. In existing researches using naturalistic driving data, the minimum speeding threshold of a speeding event is considered as the speeding limit plus 1mph (Reagan et al., 2013; Zhao and Wu, 2013). This would provide a buffer for unintentional speeding or the error from measuring devices to avoid inaccurate classification. Furthermore, in the United States, drivers could receive speeding warnings if speeding 1 to 5mph over the speed limit (Reagan et al., 2013). The minimum speeding duration is also introduced to avoid inaccurate classification due to the unintentional speeding and measuring errors. There are existing researcher consider an event as a valid speeding event only if the continuous speeding time is longer than 30 s (Chevalier et al., 2017, 2016). Therefore, a speeding event is located in this dataset when the driver continuously drives faster than the speeding limit more than 1mph and longer than 30 s, and a speeding event ends when the driving speed drops back to the minimum speeding threshold.

A speeding event is examined from two perspectives: speeding

pattern and speeding duration. Speeding pattern indicates the speeding severity. A speeding event is classified as moderate speeding pattern, while the speed is between speed limit plus 1mph and speed limit plus 5mph, and higher speeding pattern, when the speed is higher than 5mph (Avrenli et al., 2013; Reagan et al., 2013; Richard et al., 2013; Zhao and Wu, 2013). Speeding duration is classified as moderate speeding duration, when a speeding event is lasting between 30 s to 2 min, and longer speeding duration, when a speeding event is longer than 2 min. The purpose of classifying the speeding events into moderate speeding duration and longer speeding duration is to understand the reason behind various speeding durations. The existing studies about investigating speeding duration are rare.

3.2. Classification based association (CBA) mining

Data mining involves machine learning, statistical knowledge, modeling concepts and database management. Association rules mining, a descriptive analytics technique, discovers significant rules showing variable category conditions that occur frequently together in a dataset. It involves the use of machine learning models to analyze data for patterns, or co-occurrence, in a database. It identifies frequent if-then (antecedent-consequent) associations, which are called association rules. As a non-parametric method, it avoids making any parametric assumptions as most parametric methods do. Association rules are created by searching data for frequent if-then patterns and using the criteria support and confidence to identify the most important relationships. Support indicates how frequently the items appear in the data. Confidence is an indicator of the number of times the if-then statements are found true. In other words, support represents how much historical data supports the identified rule and confidence represents how confident we are that the rule holds. A third metric, called lift, which is the ratio of confidence to support, can be used to compare observed confidence with expected confidence (Agrawal and Srikant, 1994)

Association rules are derived from *itemsets* which comprise of two or more items. The Apriori algorithm considers any subset of a frequent itemset to also be a frequent itemset. No superset of any infrequent itemset is generated or tested which allows pruning of many item combinations. *Apriori* 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. This algorithm helps researchers mine out frequently occurring itemsets, subsequences, arrangements and interesting associations between various items.

A set of definitions is provided here before demonstrating the method with an example. Let $I = i_1, i_2, \ldots i_m$ be a set of items (e.g., a set of crash categories for a particular crash record) and $C = c_1, c_2, \ldots, c_n$ be a set of database crash information (transaction) where each crash record c_i contains a subset of items chosen from I. An itemset with k items is called as a k-itemset.

An association rule can be demonstrated as A → B, where A and B are disjoint itemsets. Here, A is known as the antecedent and B is the consequent. The strength of the association rule can be measured using the values of measures like support, confidence and lift. For the purpose of our analysis, support is defined as the percentage of casualties in the dataset that contains the itemset. Confidence is the ratio of the number of all crashes in C to the number of crashes that include all items in I. Lift is the ratio of confidence over expected confidence. The equations of support are listed in Eq. 1 through 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 antecedent A

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

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

N = Total number of incidents

S(A) = Support of antecedent

S(B) = Support of consequent

S(A → B) = Support of the association rule (A → B)

The definitions for confidence and lift are provided in Eqs. 4 and 5 respectively. Confidence measures the reliability of the inference of a generated rule. Higher confidence for A → B indicates that presence of B is highly likely in transactions (or observations) containing A. The lift of the rule A → B associates the frequency of co-occurrence of the antecedent and the consequent to the expected frequency of co-occurrence.

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

$$L(A \rightarrow B) = \frac{S(A \rightarrow B)}{S(A).\ S(B)} \tag{5}$$

where,

C(A → B) = Confidence of the association rule (A → B)

L(A → B) = Lift of the association rule (A → B)

A lift value > 1 indicates significant interdependence between the antecedent and the consequent, while a value < 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 2-product rule; similarly, a rule with two antecedents and single consequent or one antecedent and two consequents is defined as a 3-product rule. A critical inference of the association rules is that the generated rules need not be interpreted as causation, rather as association.

There are several existing transportation related researches employed association rule methods (Das et al., 2019, 2018; Das and Sun, 2014). While association rule mining tries to find all rules in the database that satisfy the desired minimum support and minimum confidence constraints, Classification Based Association (CBA) rule mining aims to discover a small set of rules in the database to form an accurate classifier based on class association rules (Liu et al., 1998). Therefore, there is a pre-set target (called class) for the classification rule mining process. Instead of using an itemset as in association rules mining, a *ruleitem* is used, which consists of a *condset* (a set of items) and a class. A rule pruning technique is adopted to prune off non-predictive and overfitting rules. This is an extension to the approach used in association rule mining.

4. Data description

4.1. Data sources

This study uses two data sources for this analysis: 1) SPMD data, and 2) HPMS data of Michigan for the year 2016. The following subsections provide a brief description of these two datasets and the final data used in this study.

4.1.1. Safety pilot model deployment (SPMD) data

SPMD is a part of the Connected Vehicle Safety Pilot Program (Hamilton, 2015). SPMD study researchers collected several types of driving data, including data on vehicle trajectory, data about surrounding vehicles, and network data. This study focused on the vehicle trajectory data and the surrounding vehicle data collected by Data Acquisition System 1 (DAS1). University of Michigan Transportation Research Institute (UMTRI) was responsible for DAS1. It uses the Mobileye sensor for data collection (Mobileye, 2020). Trips with less than

Table 1
Overview of Selected Variables.

Variable	Description	Levels	Count	Percentage
vstime	speeding time for one speeding event	1: moderate speeding duration (0.5 – 2 min.)	1772	63.02
		2: longer speeding duration (longer than 2 min.)	1040	36.98
vsspeed	average amount of speed over the speed limit during one speeding event	1: moderate speeding pattern (> speed limit + 1 & < speed limit +5)	1887	67.11
		2: higher speeding pattern (> speed limit +5)	925	32.89
pre_spdloss	total speed lose before the speeding event occurred	1: great speed loss (-inf to -0.1)	1001	35.60
		2: moderate_speedloss (-0.1 to 0.01)	746	26.53
		3: minor speed loss (-0.01 to 0)	811	28.84
		4: no speed loss (o to inf)	254	9.03
talgate	if this vehicle is tailgating	1: yes	730	25.96
		2: no	2082	74.04
f_system	functional class system	1: interstate highway	835	29.69
		2: principle arterial	70	2.49
		4: minor arterial	413	14.69
		5: major collector	1494	53.13
access_con	if there is access control	1: yes	761	27.06
		2: no	2051	72.94
shoulder_w	width of the shoulder	no shoulder	798	28.38
		1 – 8 feet	1187	42.21
		8 – 10 feet	606	21.55
		larger than 10 feet	221	7.86
lane_width	width of the lane	lw10 feet	52	1.85
		lw11 feet	305	10.85
		lw12 feet	2340	83.21
		lw13 feet	115	4.09
median_wid	width of the median	0 feet	520	18.49
		less than 25 feet	201	7.15
		25 - 60 feet	1093	38.87
		60 – 70 feet	903	32.11
		> 70 feet	95	3.38
speed_limi	speed limit	less than 40mph	506	17.99
		between 40–60 mph	286	10.17
		higher than 60 mph	2020	71.83
peak	if the event occurred during peak hour	0: non peak hour	1950	69.35
		1: peak hour	860	30.58
median_t	median type	none	232	8.25
		curbed	520	18.49
		rigid barrier	558	19.84
		semirigid barrier	275	9.92
		unprotected	1193	42.43
aadt	aadt of the speeding road segment	1,000 – 25,000	669	23.79
		25,000–50,000	373	13.26
		50,000 – 70,000	838	29.80
		larger than 70,000	932	33.14
traveltime	total travel time for the trip which the speeding event occurred	less than 30 min	1829	65.04
		30 – 60 min	778	27.67
		larger than 60 min	205	7.29

ten minutes trip duration were excluded from the dataset. Data for 92 vehicles was available after filtering for trip duration. This data was collected in between October 2012 and April 2013.

4.1.2. Highway performance monitoring system (HPMS)

HPMS contains roadway inventory data for all the states in the U.S. In this study, the authors used the roadway inventory data for Michigan, 2016. The data included roadway and traffic features such as number of lanes, roadway functional classification, and average annual daily traffic (AADT), etc. Table 1 lists the roadway features included in this study. Roadway features such as type of shoulder, AADT and lane width had missing values for some rows, were thus not included in this study.

4.2. Data integration

The researchers reduced the trajectory data before merging it with the roadway inventory data. SPMD data contains data for every 100 centiseconds. The researchers only used data at the second level for merging. This was done to speed up the merging process without losing accuracy. Open source GIS tool QGIS was used for merging the two datasets. After the conflation of two datasets, every geographical point in the SPMD dataset will have its corresponding information from HPMS, such as speed limit. The authors defined the speeding event as a speeding behavior, driving over the speed limit + 1mph and lasting more than 30 s. Therefore, each trip can be divided into events like "speeding event" and "non speeding section" (Fig. 1). At each "non speeding event," the loss of speed is defined as the amount of speed difference between the average speed along this "non speeding event," and the speed limit times the total travel time at this section.

In other words, this speed loss is an index of the magnitude of speed lost during this "non speeding event" before a speeding event starts, and generally, this value is negative. For each speeding event, three variables, "vstime", "vsspeed" and "pre_spdloss" are generated. "Vstime" refers to the duration of the speeding event. It is classified into two categories: moderate (0.5 – 2 min) and longer (longer than 2 min). "Vsspeed" represents the speeding pattern of this speeding event. It has been classified as a moderate speeding pattern and a higher speeding pattern, based on whether the average speed of the speeding event is larger than the speed limit +5 mph.

The final dataset contains 2812 speeding events. Table 1 presents all variables used for the association rule mining process.

5. Results and findings

In order to extract useful association rules, two key parameters for performing the association rules mining are support and confidence, which were set as 0.001 and 0.7 respectively after some trial and error runs. Tables 2 and 3 contain mined association rules for the speeding duration and speeding pattern. For better illustration purposes, the top 20 rules of each class, if the number of minded rules is more than 20, with the highest lift value are reported in these two tables. The research team used open-source R software 'arulesCBA' package to perform this analysis (Johnson, 2017)

5.1. Rules for speeding duration

There are two categories on the right-hand side of the algorithm: *vstime = Moderate* and *vstime = Longer*. The consequent *vstime = Moderate* indicates that the speeding event lasted for 0.5 – 2 min. The consequent *vstime = Longer* indicates that the speeding event lasted more than 2 min. The first chunk of association rules in Table 2, from rule 1 to rule 15, shows the mined association rules for

Fig. 1. Speeding events and non-speeding events.

Table 2
Association Rules for Speeding Duration.

#	Antecedent	Consequent	support	confidence	lift
Rules for Longer Speeding Duration					
1	{tailgate = no,aadt = 50000 − 70000, traveltime = longer than 60 min}	vstime = Longer	0.007	0.913	2.469
2	{shoulder_w = 1 − 8 ft,median_w = 25 − 60 ft,traveltime = longer than 60 min}	vstime = Longer	0.007	0.792	2.141
3	{pre_spdloss = minor,shoulder_w = 1 − 8 ft,traveltime = longer than 60 min}	vstime = Longer	0.009	0.765	2.068
4	{pre_spdloss = minor,tailgate = no,traveltime = longer than 60 min}	vstime = Longer	0.006	0.750	2.028
5	{pre_spdloss = high,shoulder_w = 1 − 8 ft,traveltime = longer than 60 min}	vstime = Longer	0.006	0.750	2.028
6	{tailgate = no,speed_limi = higher than 60mph,traveltime = longer than 60 min}	vstime = Longer	0.014	0.745	2.015
7	{pre_spdloss = minor,f_class = interstate,shoulder_w = 1 − 8 ft}	vstime = Longer	0.015	0.737	1.992
8	{shoulder_w = 1 − 8 ft,peak = Peak = 0,traveltime = longer than 60 min}	vstime = Longer	0.021	0.732	1.978
9	{pre_spdloss = minor,peak = Peak = 0,traveltime = longer than 60 min}	vstime = Longer	0.013	0.725	1.962
10	{shoulder_w = 1 − 8 ft,aadt = 50000 − 70000,traveltime = longer than 60 min}	vstime = Longer	0.014	0.717	1.939
11	{f_class = interstate,aadt = 50000 − 70000,traveltime = longer than 60 min}	vstime = Longer	0.011	0.714	1.931
12	{f_class = interstate,shoulder_w = lager than 10 ft,median_w = larger than 60 ft}	vstime = Longer	0.007	0.714	1.931
13	{pre_spdloss = minor,f_class = interstate,median_w = larger than 60 ft}	vstime = Longer	0.011	0.711	1.923
14	{pre_spdloss = high,f_class = interstate,shoulder_w = 1 − 8 ft}	vstime = Longer	0.010	0.711	1.921
15	{pre_spdloss = no,aadt = larger than 70,000,traveltime = 30 − 60 min}	vstime = Longer	0.006	0.708	1.915
Rules for Moderate Speeding Duration					
16	{shoulder_w = 1 − 8 ft,median_w = 0,speed_limi = less than 40mph}	vstime = Moderate	0.010	1.000	1.587
17	{pre_spdloss = high,lane_w = LW13}	vstime = Moderate	0.023	0.985	1.563
18	{lane_w = LW13,speed_limi = less than 40mph}	vstime = Moderate	0.030	0.977	1.550
19	{lane_w = LW12,median_w = 0,speed_limi = less than 40mph}	vstime = Moderate	0.038	0.972	1.543
20	{speed_limi = less than 40mph,aadt = 25000 − 50000}	vstime = Moderate	0.012	0.971	1.540
21	{speed_limi = less than 40mph,traveltime = longer than 60 min}	vstime = Moderate	0.011	0.969	1.537
22	{f_class = minor arterial,traveltime = longer than 60 min}	vstime = Moderate	0.011	0.968	1.536
23	{f_class = minor arterial,shoulder_w = 1 − 8 ft,traveltime = less than 30 min}	vstime = Moderate	0.015	0.956	1.516
24	{tailgate = yes,lane_w = LW13}	vstime = Moderate	0.028	0.952	1.510
25	{lane_w = LW11,median_w = 0,peak = Peak = 1}	vstime = Moderate	0.014	0.951	1.509
26	{f_class = major collector,speed_limi = less than 40mph,traveltime = less than 30 min}	vstime = Moderate	0.012	0.946	1.501
27	{median_w = 0,speed_limi = less than 40mph,peak = Peak = 0}	vstime = Moderate	0.074	0.946	1.501
28	{f_class = principal arterial,median_w = 0,peak = Peak = 0}	vstime = Moderate	0.055	0.945	1.500
29	{median_w = 0,speed_limi = less than 40mph,traveltime = less than 30 min}	vstime = Moderate	0.075	0.942	1.495
30	{tailgate = no,median_w = 0}	vstime = Moderate	0.037	0.937	1.487
31	{f_class = principal arterial,access_con = no,median_w = 0}	vstime = Moderate	0.071	0.935	1.483
32	{pre_spdloss = high,median_w = 0,peak = Peak = 1}	vstime = Moderate	0.020	0.932	1.479
33	{median_w = 0,aadt = 25000 − 50000}	vstime = Moderate	0.028	0.930	1.476
34	{f_class = minor arterial,lane_w = LW12,peak = Peak = 0}	vstime = Moderate	0.022	0.925	1.468
35	{lane_w = LW10,peak = Peak = 0}	vstime = Moderate	0.013	0.925	1.468

longer speeding duration events and rule 16 to rule 35 are the mined rules of moderate speeding duration events. The lift value for all association rules is greater than 1.4 and the confidence value is greater than 0.7. All rules are sorted by decreasing values of lift.

5.1.1. Long duration of speeding

Based on the rules in Table 2, it is perceived that the combination of the following features is strongly associated with events with longer speeding events. These features include relatively longer trip time (more than 60 min), the presence of shoulder and median, speed loss before a speeding event and roads with relatively high functional classification (interstate highway).

For instance, rule 2 in Table 2, *{shoulder_w = 1 − 8 ft, median_w = 25 − 60 ft, traveltime = longer than 60 min}*, is the rule with the second-highest value of lift. It asserts that trips with longer than 60 min of travel time and driving on roads with the presence of median and shoulder are highly associated with longer duration of speeding behavior. Corresponding indices are support = 0.7 %, confidence = 79.2 %, and lift = 2.141. Those indices represent the following:

(a) 0.7 % of longer duration speeding events are associated with these three items: driving on a roadway with the presence of shoulder and median, and total trip travel time more than 60 min.
(b) In this dataset, out of all speeding events containing this combination, 79.2 % exhibited speeding behavior for longer than 2 min.
(c) The percentage of all longer duration speeding events containing this combination was 2.141 times the percentage of long duration speeding events in the overall dataset.

Out of all rules associated with longer duration speeding, the trip characteristic of longer travel time (more than 60 min) was highly associated with longer speeding events in 10 (rule1 − 6 and rule 8–12) out of 15 rules. Speed loss before speeding events is contained in 7 (rule 3–5, rule 7, rule 9 and rule 13–14) out of 15 rules. Roads with higher functional classification and presence of median and shoulder repeatedly present in 13 out of 15 rules. One-point worth to be pointed out is that roads with higher functional class can be indicated by many features/items: aadt = 50,000 − 70,000 (rule1, 10, 11), median_w = 25 − 60 feet (rule2), speed_limi = ligher than 60mph(rule6), f_class = interstate highway (rule7, rule 11–14), shoulder_w = larger than 10feet (12),aadt = larger than 70,000 (rule15). It is also worth pointing out that the observation that two items – travel time longer than 60 min and experienced congestion before a speeding event appeared in 4 rules (rule 3–5, rule 9). It indicates that drivers of long trips are more likely to do speeding longer after being congested in traffic to compensate for the lost travel time.

5.1.2. Moderate duration of speeding

It was observed that the speeding event with moderate duration is strongly associated with the combination of features including roadways with relatively lower function classification, absence of the median, relatively short trip time (less than 30 min) and speed loss before the speeding event. In these rules, roadways with relatively lower functional classification is a collectively way to describe many items/characteristics presented in the rules. These items include f_class = minor arterial, f_class = major collector, median_w = 0 (no median exists), speed_limi = less than 40mph, aadt = 25,000–50,000, lane_w = LW10 (lane width is 10 feet) in the 18 out of 20 min. d rules of

Table 3
Association Rules for Speeding Patterns.

#	Antecedent	Consequent	support	confidence	lift
Rules for Higher Speeding Pattern					
1	{lane_w = LW12,median_t = curbed,aadt = 1000 − 25000}	vsspeed = higher	0.009	0.897	2.726
2	{pre_spdloss = high,lane_w = LW12,median_t = curbed}	vsspeed = higher	0.006	0.895	2.720
3	{f_class = principal arterial,shoulder_w = 1 − 8 ft,speed_limi = less than 40mph}	vsspeed = higher	0.008	0.880	2.675
4	{access_con = no,lane_w = LW12,median_w = 25 − 60 ft}	vsspeed = higher	0.011	0.838	2.547
5	{shoulder_w = 1 − 8 ft,median_w = less than 25 ft,aadt = larger than 70,000}	vsspeed = higher	0.006	0.773	2.349
6	{lane_w = LW12,speed_limi = less than 40mph,traveltime = 30 − 60 min}	vsspeed = higher	0.007	0.769	2.338
7	{shoulder_w = 1 − 8 ft,lane_w = LW12,speed_limi = less than 40mph}	vsspeed = higher	0.008	0.767	2.331
8	{shoulder_w = 1 − 8 ft,speed_limi = less than 40mph,peak = Peak = 0}	vsspeed = higher	0.011	0.762	2.316
9	{tailgate = no,f_class = major collector}	vsspeed = higher	0.005	0.750	2.280
10	{tailgate = no,lane_w = LW12,speed_limi = less than 40mph}	vsspeed = higher	0.008	0.742	2.255
11	{tailgate = no,aadt = 50000 − 70000,traveltime = longer than 60 min}	vsspeed = higher	0.006	0.739	2.247
12	{pre_spdloss = moderate,f_class = minor arterial,lane_w = LW12}	vsspeed = higher	0.005	0.737	2.240
Rules for Moderate Speeding Pattern					
13	{pre_spdloss = minor,shoulder_w = lager than 10 ft,traveltime = less than 30 min}	vsspeed = Moderate	0.013	0.902	1.345
14	{shoulder_w = lager than 10 ft,aadt = 50000 − 70000,traveltime = less than 30 min}	vsspeed = Moderate	0.018	0.893	1.331
15	{pre_spdloss = high,speed_limi = between 40–60 mph,aadt = 25000 − 50000}	vsspeed = Moderate	0.014	0.889	1.325
16	{f_class = principal arterial,lane_w = LW11}	vsspeed = Moderate	0.019	0.883	1.316
17	{shoulder_w = lager than 10 ft,median_t = unprotected,traveltime = less than 30 min}	vsspeed = Moderate	0.012	0.872	1.299
18	{pre_spdloss = high,shoulder_w = 8 − 10 ft,aadt = larger than 70,000}	vsspeed = Moderate	0.031	0.861	1.284
19	{pre_spdloss = high,shoulder_w = No shoulder,speed_limi = between 40–60 mph}	vsspeed = Moderate	0.027	0.854	1.273
20	{speed_limi = between 40–60 mph,peak = Peak = 0,aadt = 25000 − 50000}	vsspeed = Moderate	0.012	0.854	1.272
21	{shoulder_w = lager than 10 ft,speed_limi = higher than 60mph,traveltime = less than 30 min}	vsspeed = Moderate	0.037	0.847	1.262
22	{pre_spdloss = minor,shoulder_w = lager than 10 ft,median_w = 25 − 60 ft}	vsspeed = Moderate	0.017	0.845	1.259
23	{pre_spdloss = high,median_w = 0,aadt = 25000 − 50000}	vsspeed = Moderate	0.016	0.830	1.237
24	{pre_spdloss = high,shoulder_w = 1 − 8 ft,aadt = 25000 − 50000}	vsspeed = Moderate	0.010	0.829	1.235
25	{pre_spdloss = moderate,shoulder_w = 8 − 10 ft,median_w = larger than 60 ft}	vsspeed = Moderate	0.010	0.829	1.235
26	{f_class = principal arterial,speed_limi = between 40–60 mph,traveltime = less than 30 min}	vsspeed = Moderate	0.023	0.825	1.229
27	{pre_spdloss = high,f_class = principal arterial,speed_limi = between 40–60 mph}	vsspeed = Moderate	0.023	0.823	1.226
28	{pre_spdloss = moderate,shoulder_w = 8 − 10 ft,median_t = unprotected}	vsspeed = Moderate	0.011	0.821	1.223
29	{pre_spdloss = minor,shoulder_w = lager than 10 ft,speed_limi = higher than 60mph}	vsspeed = Moderate	0.022	0.816	1.216
30	{shoulder_w = No shoulder,speed_limi = between 40–60 mph,peak = Peak = 0}	vsspeed = Moderate	0.028	0.813	1.211
31	{pre_spdloss = high,median_t = unprotected,aadt = 25000 − 50000}	vsspeed = Moderate	0.011	0.811	1.208
32	{shoulder_w = No shoulder,speed_limi = between 40–60 mph,traveltime = less than 30 min}	vsspeed = Moderate	0.029	0.810	1.207

moderate speeding duration events in Table 2.

Additionally, one road feature – the absence of median presented in 10 out of 20 rules. It states the dominant impact of the absence of a median on the speeding duration. The absence of a median inevitably generates a sense of insecurity from the opposing traffic for the speeding driver, which shortens the speeding duration.

5.1.3. Comparison of two sets of rules generated for speeding duration classification

(a) Road segments with a higher functional class have a significant impact on longer-duration speeding events. On higher road functional roadways, the speeding behavior tends to last longer. On the contrary, in Table 2, a majority of rules for speeding events with moderate speeding duration have a relatively lower functional class.
(b) Speed loss is one of the factors that trigger the speeding behavior since this factor presents in many rules for both longer and moderate speeding duration events. Whether it will trigger a longer or moderate speeding event depends on the combination with other factors.
(c) The presence or absence of a median is a significant factor associated with speeding duration. Speeding events tended to not last long on the road segment without a median. On the contrary, with the presence of a wide median, the speeding duration tends to last longer.
(d) The total travel time of a trip poses an impact on the speeding duration of a speeding event. Longer trips (longer than 60 min) are associated with longer speeding events and shorter trips (less than 30 min) are associated with moderate speeding events. There are two possible explanations. One is that longer trips provide more opportunities for drivers to speeding longer. Another one is that longer trips normally have a higher possibility of using roadways with higher functional classification. The roadway with higher functional classification is normally associated with higher construction and operational standards: wide lanes, better pavement conditions, more access control, which ensure fewer interruptions for drivers and encourage longer speeding behavior.

5.2. Rules for speeding patterns

There are two categories in the consequent: *vsspeed = Higher* and *vsspeed = Moderate*. The consequent vsspeed = higher indicates that the drivers' average speed during this speed event was 5 mph more than the speed limit. The consequent *vsspeed = Moderate* represents that the average speed during a speeding event was more than the speed limit + 1mph but less than speed limit + 5mph. The first chunk of Table 3, rule 1–12, contains the rules for the higher speeding pattern and the rest are rules for the moderate speeding pattern.

5.2.1. Higher speeding pattern

Based on the frequency of the features presented in the rules of high speeding pattern, Table 3 shows that several trip/driving features and several road characteristics are highly associated with the higher speeding pattern. Trip/driving features include experiencing speed loss before a speeding event. Road characteristics include the presence of shoulder and curbed median and lower functional classification of the roadways comparing with an interstate highway. As mentioned above, roadways with a lower functional classification could be identified through several features: median = curbed, acces_con = no (no access control), speed_limi = less than 40mph, aadt = 1000 − 25000 and f_class = major collector/minor arterial/principal arterial.

Six out of twelve rules (rule 1–5, 12) generated from the higher speeding pattern dataset contain the presence of a median. With the

barrier or buffer separating the traffic from the opposite direction, this would provide a sense of safety that might encourage speeding drivers to speed more severely. In addition, a lower functional class of the roadway was found in 10 out of 12 rules (rule 1–4, rule 6–10 and rule 12). It is interesting to note that the higher speeding pattern dominantly occurs on the roadways with a relatively lower functional classification, rather than on an interstate highway. Commonly, drive on the roadway with a lower functional classification has a higher probability of experiencing constant interruptions from traffic lights, congestions/speed loss and low-speed limits. The combination of these factors may trigger more aggressive driving behavior.

5.2.2. Moderate speeding pattern

In this set of rules for moderate speeding pattern events, the combination of experiencing speed loss before a speeding event, short trip time (less than 30 min.), roadways with relatively higher functional class (shoulder width larger than 10 feet, unprotected median, speed limit higher than 40mph, AADT larger than 25,000) shows strong association with the moderate speeding pattern.

It is noteworthy that 11 out of 20 rules generated for moderate speeding pattern events contain the feature - speed loss before the speeding event occurred. In other words, the moderate speeding pattern often occurred after experienced congestion. Moreover, the combination of travel time less than 30 min and roadways with relatively higher functional classification appeared in 6 rules (rul2, 13, 14, 17, 21, 26 and 32). That says short trips on higher functional classification roadways are more likely to speed moderately.

5.2.3. Comparison of two sets of rules generated for speeding pattern classification

(a) The speed loss is a preeminent factor that triggers both the higher speeding pattern and the moderate speeding pattern.
(b) Conversely, the higher speeding pattern occurs more on the roadways with lower functional classification with the presence of median and the moderate speeding pattern mostly occurs on the roadways with relatively higher functional classification.
(c) Travel time is not a frequent item for the rules for the higher speeding pattern, however, short travel time – less than 30 min presented in many rules of the moderate speeding pattern. The combination of short travel time and traveling on a roadway with a higher functional class leads to a moderate speeding pattern.

6. Conclusions

This research investigated the potential associations between speeding behavior and trip/driving/roadway features by exploring a naturalistic driving dataset using the classification and association rule mining algorithm. This paper explored the associations from two perspectives: duration of speeding behavior and speeding patterns. Moreover, the investigation identified not only individual features but also the combination of features that might associate with speeding behavior. The purpose of the research is to understand the speeding behavior from speeding duration and speeding pattern perspectives and to develop countermeasures to mitigated speeding behavior, which has been proved as a leading cause of fatal crashes.

The duration of speeding behavior was classified as a longer duration and a moderate duration. The speeding pattern was classified as a higher pattern and a moderate speeding pattern. By comparing the mined rules, this research found the combinations of speed loss caused by congestion, presence of a median, longer trip and driving on a roadway with higher functional class are highly associated with longer speeding duration (more than 2 min). On the contrary, the combination of driving on roadways with lower functional class, absence of a median and trip time shorter than 30 min are highly associated with relatively moderate speeding events (less than 2 min). It is also found that higher speeding patterns (more than speed limit + 5mph) are strongly associated with driving on roadways with relatively lower functional class, experiencing speed loss/congestion and the presence of a median. Furthermore, the moderate speeding pattern is associated with the combinations of factors like experiencing congestion before a speed event, driving on roadways with higher functional class and a relatively shorter trip (less than 30 min).

The corresponding countermeasures should aim to reduce these speeding events with a higher speeding pattern and longer speeding duration. To reduce the speeding events with higher speeding pattern, the findings suggest transportation agencies, law enforcement or roadway designer should invest more attention and develop countermeasures on the road segments with relatively lower function class, regularly congested and with median existed, because these locations are highly associated with the higher speeding pattern, which could lead to a higher probability of incidents. For example, the countermeasure could be the installation of dynamic signals on the busy local corridors could help to relieve the regular occurring congestions, which would greatly reduce the incentives of a higher speeding pattern. Moreover, combinations of features associated with longer speeding events worth of transportation agencies' attention as well, since it often occurs on the longer trips and roadways with higher functional class. The tiredness and distraction caused by longer trips and driving at a higher speed on the roads with higher function class could lead to more severe accidents. Transportation agencies could think of providing more resting stations along roadways with higher functional class. These findings can also help in calibrating driver behavior parameters for transportation-related simulation tools instead of assuming the driving behaviors are the same on all types of occasions.

There are several limitations to this research. The association rules algorithm is a great way to mine meaningful patterns. However, the process of determining the optimum value of support and confidence is somewhat subjective and deferent choices of the values could lead to slightly different results, even the main trend remains. This requires researchers or practitioners to pay more attention to this process of tuning parameters to ensure the stability of the model performance. Another limitation is the information of the participants of this naturalistic driving dataset is not yet available. Thus, the further connection between the characteristics of the drivers and their speeding behaviors cannot be established. A future study could also investigate the relationship between the identified patterns and crash data. It would present stronger evidence that the minded patterns through naturalistic driving data are important to improve roadway safety.

CRediT authorship contribution statement

Xiaoqiang Kong: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation. **Subasish Das:** Methodology, Supervision, Validation. **Kartikeya Jha:** Writing - original draft. **Yunlong Zhang:** Conceptualization, Investigation, Supervision, Validation.

Declaration of Competing Interest

None.

Acknowledgments

The authors appreciate the assistance provided by Apoorba Bibeka who helped clean and conflate large data files like SPMD and HPMS. A lot of appreciations go to three anonymous reviewers. Their thoughtful and detailed comments provided a solid base for us to improve this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105620.

References

Afghari, A.P., Haque, M.M., Washington, S., 2018. Applying fractional split model to examine the effects of roadway geometric and traffic characteristics on speeding behavior. Traffic Inj. Prev. 19 (8), 860–866. https://doi.org/10.1080/15389588.2018.1509208.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases. VLDB. pp. 487–499.

Andersen, C.S., Reinau, K.H., Agerholm, N., 2016. The relationship between road characteristics and speed collected from floating car data. In: DIVA. 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. p. 10.

Avrenli, K.A., Benekohal, R., Ramezani, H., 2013. Flagger effects in reducing the likelihood of rear-end collisions for in-platoon vehicles in work zones. Transp. Res. Rec. 2337 (1), 9–16.

Bassani, M., Dalmazzo, D., Marinelli, G., Cirillo, C., 2014. The effects of road geometrics and traffic regulations on driver-preferred speeds in northern Italy. An exploratory analysis. Transp. Res. Part F Traffic Psychol. Behav. 25, 10–26. https://doi.org/10.1016/j.trf.2014.04.019.

Ben-Bassat, T., Shinar, D., 2011. Effect of shoulder width, guardrail and roadway geometry on driver perception and behavior. Accid. Anal. Prev. 43 (6), 2142–2152. https://doi.org/10.1016/j.aap.2011.06.004.

Chevalier, A., Coxon, K., Chevalier, A.J., Wall, J., Brown, J., Clarke, E., Ivers, R., Keay, L., 2016. Exploration of older drivers' speeding behaviour. Transp. Res. Part F Traffic Psychol. Behav. 42, 532–543.

Chevalier, A., Coxon, K., Rogers, K., Chevalier, A.J., Wall, J., Brown, J., Clarke, E., Ivers, R., Keay, L., 2017. Predictors of older drivers' involvement in high-range speeding behavior. Traffic Inj. Prev. 18 (2), 124–131.

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

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. Int. J. Urban Sci. 23 (1), 30–48. https://doi.org/10.1080/12265934.2018.1431146.

Das, S., Kong, X., Tsapakis, I., 2019. Hit and run crash analysis using association rules mining. J. Transp. Saf. Secur. 0 (0), 1–20. https://doi.org/10.1080/19439962.2019.1611682.

Eksler, V., Popolizio, M., Allsop, R., 2009. How far from Zero? Benchmarking of Road Safety Performance in the Nordic Countries. European Transport Safety Council.

Elvik, R., 2008. Dimensions of road safety problems and their measurement. Accid. Anal. Prev. 40 (3), 1200–1210.

Hamilton, B.A., 2015. Safety Pilot Model Deployment-Sample Data Environment Data Handbook. US Dep. Transp. Intell. Transp. Syst. Jt. Program Off. URL https://catalog.data.gov/dataset/safety-pilot-model-deployment-data (Accessed 25 January 2020). [WWW Document].

Johnson, I., 2017. arulesCBA: Classification for Factor and Transactional Data Sets Using Association Rules.

Liu, B., Hsu, W., Ma, Y., 1998. Integrating Classification and Association Rule Mining. KDD, pp. 80–86.

Liu, S., Wang, J., Fu, T., 2016. Effects of lane width, lane position and edge shoulder width on driving behavior in underground urban expressways: a driving simulator study. Int. J. Environ. Res. Public Health 13, 10. https://doi.org/10.3390/ijerph13101010.

Lobo, A., Rodrigues, C., Couto, A.Fdo, 2018. Free-Flow Speed Model Based on Portuguese Roadway Design Features for Two-Lane Highways | Transportation Research Record: Journal of the Transportation Research Board [WWW Document]. URL https://trrjournalonline.trb.org/doi/abs/10.3141/2348-02 (Accessed 7 September 2018). .

Mobileye, 2020. Mobileye. [WWW Document]. URL https://www.mobileye.com/our-technology/ (Accessed 14 February 2020).. .

National Center for Statistics and Analysis, 2019. Traffic Safety Facts (No. DOT HS 812 687). Traffic Safety Facts.

National Highway Traffic Safety Administration, 2020. What Drives Speeding? [WWW Document]. What Drives Speeding. URL https://www.nhtsa.gov/risky-driving/speeding (Accessed 31 January 2020). .

Reagan, I.J., Bliss, J.P., Van Houten, R., Hilton, B.W., 2013. The effects of external motivation and real-time automated feedback on speeding behavior in a naturalistic setting. Hum. Factors 55 (1), 218–230.

Richard, C.M., Campbell, J.L., Lichty, M.G., Brown, J.L., Chrysler, S., Lee, J.D., Boyle, L., Reagle, G., 2012. Motivations for Speeding, Volume I: Summary Report. National Highway Traffic Safety Administration, United States.

Richard, C., Campbell, J.L., Brown, J.L., Lichty, M.G., Chrysler, S.T., Atkins, R., 2013. Investigating speeding behavior with naturalistic approaches: methodological lessons learned. Transp. Res. Rec. 2365 (1), 58–65.

Zhao, G., Wu, C., 2013. Effectiveness and acceptance of the intelligent speeding prediction system (ISPS). Accid. Anal. Prev. 52, 19–28.