Improper Passing and Lane-Change Related Crashes: Pattern Recognition Using Association Rules Negative Binomial Mining

Subasish Das, Sudipa Chatterjee, and Sudeshna Mitra

Abstract Improper passing or lane-change related traffic crash is a critical issue. To pass or change lanes, the driver is required to make several judgments based on dynamic variables and allow little room for judgmental error that can result in drastic consequences if performed improperly. We used Florida rural roadway crash data from the second Strategic Highway Research Program (SHRP2) to investigate improper passing/lane-change related crashes. We applied an unsupervised data mining technique (known as association rules negative binomial (NB) miner) to extract the knowledge pattern of co-occurrence of the significant variables. This method identified some hidden trends from the complex nature of the traffic crash database. The findings show that this algorithm is suitable for pattern identification from traffic crash data.

Keywords Safety · Improper passing/lane-change related crashes · Association rules negative binomial miner · Data mining

1 Introduction

Significant numbers of crashes occur on rural roadways, and large portions of them are due to the lack of effective countermeasures to separate opposing traffic flows. As a result, a major concern involves vehicles crossing the centerline and resulting in

S. Das (✉)
Texas A&M Transportation Institute, Bryan, TX 77807, USA
e-mail: s-das@tti.tamu.edu

S. Chatterjee
Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
e-mail: sudipa.chatterjee@iitkgp.ac.in

S. Mitra
Global Road Safety Facility, The World Bank, 1818 H Street NW, Washington, DC 20433, USA
e-mail: smitra5@worldbank.org

V. Singh et al. (eds.), *Computational Methods and Data Engineering*,
Advances in Intelligent Systems and Computing 1227,
https://doi.org/10.1007/978-981-15-6876-3_46

either sideswiping or head-on collisions. At the same time, improper passing/lane-change related crashes on divided roadways are not negligible. It is necessary to investigate the patterns of improper passing/lane-change related crashes. Driving on roadways is a multifaceted task, which requires the perception, comprehension, and projection of states of the roadway condition, as well as decision making on courses of spontaneous action and execution of driving behaviors. Passing or lane-changing on rural roadways is complex due to the roadway environment and opposite direction vehicles if the roadway is undivided. Many studies envisioned to explore the complex nature of passing to improve roadway standards by re-evaluating the patterns of improper passing crashes. Factors range from a wide variety in improper passing/lane-change crashes: the presence of curvature, curve radius, curve length, skid condition, weather, lane width, and the presence of countermeasures like raised pavement markers or edge lines, size of the vehicle, posted speed limit, vehicle speed, lighting condition, traffic volume, and time of the day. Linear regression methods usually cannot capture the clustered nature of complex crash data. Mixed effect modeling can capture the clustering nature by considering both fixed and random effects. However, it requires significant efforts to identify the data clustering using statistical methods. Data mining algorithms are helpful in determining clusters and hidden effects inside a complex dataset with less effort. We aim to mitigate the research gap by investigating a wider variety of geometric and environmental variables to identify the hidden patterns of improper passing/lane-change related crashes acquired from the second Strategic Highway Research Program (SHRP2) roadway inventory database (RID) data in Florida. To increase the precision of the knowledge extraction, we used association rules negative binomial (NB) miner algorithm.

1.1 Contexts and Research Questions

To understand the improper passing related issues, it is important to have a clear understanding on the pavement markings and scenarios for improper passing or lane-changing. Figure 1 illustrates the criteria for permitted passing and no-passing zones [1]. Figure 1a shows the permissible passing on a two-lane road with a single broken yellow centerline line. Figure 1b shows the permissible passing on a two-lane road with broken yellow line and solid yellow line marking. Figure 1c shows the improper passing on a road with similar markings. Figure 1d shows the improper passing on a road which has a solid double yellow centerline marking. Figure 1e indicates that passing zone is not allowed near the buffer zone. Usually, no-passing markings are installed in the areas with curves or low visibility ranges. Improper passing mainly indicates the pattern of passing vehicles on no-passing zones on undivided roadways. However, for divided roadways, improper lane-changing may relate to other factors. Improper passing crashes mainly involve driver cognition and attitude while passing. Earlier research shows that roadway characteristics play a key role in the likelihood

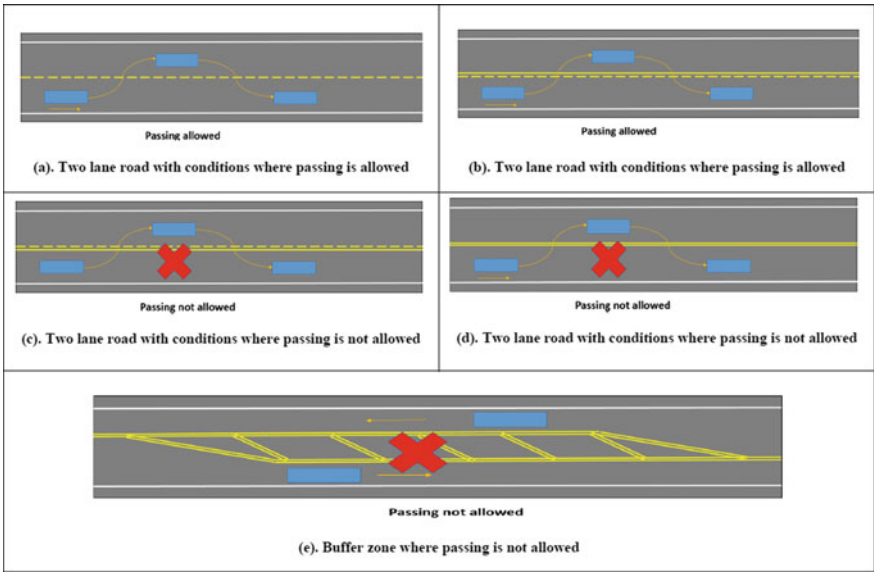

Fig. 1 Passing permitted and no-passing criteria

of injury or fatal crashes in improper passing crashes. The current research is focused on identifying the association between the key geometric features on the zones where crashes occurred due to improper passing.

The intent of this study is to address two key research questions:

- **RQ1**: What are the key contributors and association patterns in improper passing and lane-change related crashes?
- **RQ2**: Is the association rules NB miner algorithm suitable for identifying patterns from complex crash data?

2 Related Work

Many studies have attempted to understand the impact of various passing/lane-change patterns on crash severity. Historically, research has focused on refinement of roadway infrastructure guidelines, provided primarily by the American Association of State Highway and Transportation Officials (AASHTO) and the Federal Highway Administration (FHWA), which provides criteria for two-lane, two-way highways such as minimum passing sight distance and no-passing zones [2, 3]. Refinement of these guidelines included the re-evaluation and classification of passing maneuvers and the development of passing models that factored in higher speeds [2, 3]. Benefits of passing identified included: reduction of congestion, delays, and overall improvement of level of service for a roadway [4].

Current studies have explored the relationship between roadway geometry and driver behavior on rural two-lane roadways through simulation, field observation and, for a limited number of studies, naturalistic driving data. For example, the presence of a guardrail, highly visible pavement markers or the addition of centerline and shoulder rumble strips improved lane positioning of vehicles, which reduced departures from the roadway, i.e., encroachments [5, 6]. Lower overtaking speeds were found on roadways with no centerline markings and narrower lanes, while wider lanes, that offered more space, resulted in significantly higher passing speeds [7].

Roadway curvature, specifically the size of the curve, could also impact the likelihood of lane departure, left-side encroachment, and the amount of time drivers took to pass a lead vehicle, i.e., critical passing gaps where larger curves, which afford larger sight distances, had smaller passing distances compared to smaller or narrower curves [6].

Nighttime driving on rural roads can be particularly dangerous due to limited visibility and availability of lighting which are significant factors in passing behavior [8–10]. Nighttime conditions were correlated with lower traffic volumes, larger passing gaps, and larger headways [6, 8, 11]. The effect of speed on improper passing has been studied through variables including speed of subject vehicle, lead vehicle, opposing lane vehicle, effects on driver frustration, age, and gender [8, 12–14]. Higher roadway speeds were associated with smaller passing gaps, and posted speed limits also influenced driver's overtaking speed wherein lower posted speed limits correlated with lower overtaking speeds [7, 8, 12]

Many transportation safety researchers and practitioners have used association rules mining for different research problems [15–22]. One of major challenges in association rules mining is the determination of optimal support or confidence values. Association rules NB miner rather uses a model-based frequency constraint as an alternative of user-specified values. As the crash dataset is complex in nature, the identification of clusters or groups with higher precision would be an effective way of understanding the trends for which the current method has major advantages. The intent of this paper is to demonstrate an approach that can be used to better understand the factors that influence the occurrences of improper passing/lane-change crashes. This study is the first of its kind that used association rules NB miner in transportation safety research.

3 Data Description

3.1 Data Collection

The second Strategic Highway Research Program (SHRP2) project populated a roadway information database (RID) with data from the SHRP2 mobile data collection project; existing roadway data from government, public, and private sources; and supplemental data that further characterize traffic operations [23]. This database

provides good quality data that are linkable to the SHRP2 naturalistic driving study (NDS) database utilizing geographic information system (GIS) tools. The RID is the supplementary tools for safety researchers to look at data sets of selected road characteristics and study matching NDS trips to explore the relationships between driver, vehicle, and roadway characteristics. Florida RID maintains traffic crash data for six years (2005–2010). The dataset contains crash, vehicle, and person information. The database is divided into two parts: (1) local roadways and (2) state highway systems (SHS) roadways. To prepare the database for this study, this study used both roadways.

3.2 Descriptive Statistics

The primary focus of the database preparation for this study is to create a detailed database on crashes related to improper passing/lane-change. Improper passing related attributes in RID are: (1) 05 (improper lane-change), (2) (improper passing). Figure 2 illustrates the distribution of improper passing/lane-change crashes in Florida based on two major facility types: divided roadways and undivided roadways. The primary notion of this study was to investigate the passing/lane-change related crashes on rural two-lane roadways. Due to the very small sample size of the data, the research considered a broader group of facility (rural roadways) to perform the analysis. Research also directed toward diving the database into two major groups (divided and undivided roadways) to identify the safety issues associated with roadway division.

Florida RID database maintains a larger number of variables. The authors conducted a detailed literature review to investigate the significant factors associated with improper passing/lane-change crashes. A group of major roadway and environmental variables was selected from the research synthesis and was explored in Florida RID database for availability. Moreover, the authors used random forest algorithm to conduct variable importance for selecting a final group of variables. In recent years, many studies have been using random forest algorithm to determine the variable importance rather than correlation analysis due to its applicability in all variable types (discrete, continuous, ordinal, and nominal). The details of the variable selection method are not described here to make the study more focused on the current scope of the research.

Table 1 lists descriptive statistics of the final ten variables. Significant difference is visible for different variables in the divided and undivided roadways. Friction on the pavement can be measured by friction factor or skid number. A higher skid number usually indicates a higher friction factor. Maximum posted speed is higher in percentage in divided roadways while comparing with undivided roadways. Divided roadways also exhibit higher percentage in high annual average daily traffic (AADT) and high average trucking percentages. On the other hand, undivided roadways exhibit higher percentage of fatal and injury crashes.

Fig. 2 Improper passing/lane-change crashes on divided and undivided roadways in Florida

4 Methodology

We used association rules NB miner to perform the analysis. Agrawal et al. introduced the data mining on the transaction data based on the associated items using the mining association rules in 1993 [24]. This section introduces the theoretical aspects of this approach based on Hahsler study [24]. Consider $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n distinct items and $Q = \{q_1, q_2, \ldots, q_m\}$ be the transactions. Each transaction in Q contains a subject of the items in I. A rule is defined as an implication of the form *Antecedent* $\rightarrow$ *Consequent* or $M \rightarrow N$ where $M, N \subseteq I$ *and* $M \cap N = \emptyset$. A *d-itemset* has a size of d items. Support is defined on itemset $M \subseteq I$ as the proportion of transactions in which all items in Z:

$$\text{support}(M) = \frac{\text{freq}(M)}{|Q|} \tag{1}$$

Table 1 Descriptive statistics of the key variables

Category	Percentage		Category	Percentage	
Surface width (ft)	Divided (%)	Undivided (%)	AADT (vpd)	Divided (%)	Undivided (%)
0.00–10.00	0.03	0.17	0–10,000	4.97	64.17
10.01–20.00	0.78	14.21	10,000–20,000	15.17	26.24
20.01–30.00	62.60	81.95	20,000–30,000	18.82	3.92
30.01–40.00	36.05	3.49	30,000–40,000	13.21	1.92
>40	0.55	0.17	40,000–50,000	14.33	0.96
Skid number			50,000–60,000	10.40	0.87
0.00–30.00	10.49	1.66	>60,000	23.10	1.92
30.01–40.00	70.50	33.13	Lighting condition		
40.01–50.00	17.93	48.13	Daylight	70.59	70.70
>50	1.09	17.09	Dark (no streetlight)	21.11	21.01
Shoulder type			Dark (streetlight)	4.37	3.49
Paved	25.60	77.77	Dawn/dusk	3.94	4.80
Paved with warning	73.57	11.25	Weather condition		
Other	0.83	10.99	Inclement	34.04	28.68
Maximum speed (mph)			Non-inclement	65.96	71.31
0–50	3.36	13.60	Surface condition		
50–60	10.14	77.86	Dry	83.65	91.02
>60	86.50	8.54	Wet	16.35	8.98
Average truck percentage			Crash severity		
0.00–5.00	3.48	5.41	Fatal	1.49	4.80
5.01–10.00	14.97	24.59	Incapacitating injury	7.93	17.18
10.01–15.00	23.56	29.47	Possible injury	17.90	17.96
15.01–20.00	26.00	18.48	Non-incapacitating injury	13.70	17.44
>20	32.00	22.06	No injury	58.98	42.63

where freq(M) = frequency of itemset M (number of transactions in which M occurs) in Q, and $|Q|$ = number of transactions in the database.

For the rule: $M \rightarrow N$, confidence can be calculated as:

$$\text{Confidence}(M \rightarrow N) = \frac{\text{Support}(M \rightarrow N)}{\text{Support}(M)} \tag{2}$$

Relevance of the itemset Z to the user depends on the constraint: support $\geq \sigma$, where: σ is a user-specified minimum support value. Itemsets that satisfy the minimum support constraint are known as frequent itemsets. The performance measure of association rules is lift, which provides measure of the deviation from statistical independence of the relationship between M and N and is useful to identify associations that are significant deviations from the assumption of statistical independence. Lift is defined as:

$$\text{Lift}(M \rightarrow N) = \frac{\text{Support}(M \rightarrow N)}{\text{Support}(M) \times \text{Support}(N)} \tag{3}$$

Hashler [24] replaced the usage of the lift for a model aiming to evaluate the deviation for the set of all possible 1-extensions of an itemset together to evaluate a local frequency constraint for these extensions. The count distribution of itemsets is assumed to follow a NB distribution. The probability mass function of NB distribution is:

$$Pr[P = p] = (1+b)^{-t} \frac{\Gamma(t+p)}{\Gamma(p+1)\Gamma(t)} \left(\frac{b}{1+b}\right)^{p}, \; p = 0, 1, 2 \ldots \tag{4}$$

where $t =$ dispersion parameter, $b =$ mean parameter, $p =$ realization of the random variable P. Consider $\Sigma = \sigma f$, where f is the number of total transactions in the database. The count threshold is equivalent to the minimum support σ. Then, the expected number of 1-itemsets can be considered as:

$$x\, Pr[P \geq \Sigma] \tag{5}$$

where $x =$ the number of variables items. The counts for the 1-extensions of association m can be modeled by random variable P_m with the following probability mass function:

$$Pr[P_m = p] = (1+b_m)^{-t} \frac{\Gamma(t+p)}{\Gamma(p+1)\Gamma(t)} \left(\frac{b_m}{1+b_m}\right)^{p}, \quad \text{for } p = 0, 1, 2 \ldots \tag{6}$$

Consider M be the set of all 1-extension of a known association m which are generated by joining m with all candidate items, which co-occurrence with m in at least ρ transactions. For set M, the predicted precision is:

$$\text{Predicted precison of a rule} = \begin{cases} (o-e)/o \text{ if } o \geq e \text{ and } o > 0 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{7}$$

where o is the observed and e is the expected number of candidate items which have a co-occurrence frequency with itemset m of $p \geq \rho$. The observed number is calculated as the sum of items with count p by $o = \sum_{p=\rho}^{p\max} o_r$, where $p_{\max}$ is the highest observed co-occurrence. The expected number is given by the baseline model

as $e = (x - |m|)\Pr[P_m \geq \rho]$, where $x - |m|$ is the number of possible candidate items for pattern m. This method overcomes the conventional drawbacks of defining a minimum threshold support.

5 Results

We used open-source R package '*arulesNBMiner*' to perform analysis on two broader groups of unsupervised datasets based on the roadway facility types: divided and undivided roadways [25, 26]. As no response variable was preselected in both of the datasets, the learning algorithm of the model development was unsupervised in nature. It is important to note that a lower precision threshold usually increases a significant amount of increase of the generated rules. The number of itemsets is also another issue to interpret the results. Rules associated with 3 itemsets include one extra attribute than 2-itemset rules. By investigating these issues, this study used precision threshold as 0.9 and maximum number of itemset as 3. Tables 2 and 3 list the top twenty rules (based on precision scores) generated for two and three itemsets respectively. Table 2 lists the top twenty (based on higher precision) two itemset rules for divided and undivided roadways. For divided roadways, the significant attributes are paved shoulder with warning higher AADT, surface widths (in between 20 and 40 ft), surface condition, and higher speed (above 50 mph). The significant attributes for undivided roadways are higher skid number, lower AADT, narrower surface widths (in between 20 and 30 ft), average percentage of trucks, and crash severity (fatal and injury). The findings of the two itemset precision rules are consistent with the other studies that point to speed, surface widths, and traffic volume to improper passing crashes [6, 7, 9, 12].

From Tables 2 and 3, it is seen that the top twenty rules (generated from divided rural roadways) do not exhibit severity in either antecedent or consequent. On the other hand, severity is visible in the rules for undivided roadways. For example, rule number 8 for undivided roadways (Table 2) is: *Severity* = *Incapacitating Injury* → *Speed* = 50–60 mph with a precision value of 0.963. This rule indicates that the co-occurrence of incapacitating injury crashes with higher speed (50–60 mph) is over 95%.

Table 3 lists the rules for three itemsets. For divided roadways, the significant attributes are *paved shoulder with warning, higher AADT*, *weather, surface widths* (*in between* 30 *and* 40 ft), *and higher speed* (60 mph). The significant attributes for undivided roadways are *higher skid number, lower AADT, narrower surface widths* (*in between* 20 *and* 30 ft), *average percentage of trucks, and crash severity* (*non-incapacitating injury or no injury*). The findings of the three itemset precision rules are consistent with the other studies that point to speed, surface widths, and traffic volume to improper passing crashes [6, 7]. Weather exhibits a significant factor in the group of rules generated for divided roadways. For example, rule number 2 for divided roadways (Table 3) is: {*Surface Width* = 30–40 ft, *Surface Condition* = *Wet*} → *Weather* = *Inclement* with a precision value of 0.989. This rule indicates that the

Table 2 Two itemset precision rules for divided and undivided roadways

No.	Antecedent	Consequent	Precision
Divided roadways			
1	AADT[1] = 50,000–60,000	Shoulder = Paved with warning	0.965
2	Width[2] = 30.01–40.00	Shoulder = Paved with warning	0.965
3	AADT = 40,000–50,000	Shoulder = Paved with warning	0.963
4	AADT = 60,000	Shoulder = Paved with warning	0.960
5	AADT = 60,000	Surface[5] = Dry	0.958
6	Speed[3] = 50–60	Surface = Dry	0.957
7	AADT = 30,000–40,000	Width = 20.01–30.00	0.954
8	Speed = 0–50	Width = 20.01–30.00	0.948
9	AADT = 60,000	Skid = 30.01–40.00	0.947
10	AADT = 20,000–30,000	Surface = Dry	0.946
11	Speed = 50–60	Shoulder = Paved	0.946
12	AADT = 60000	Width = 30.01–40.00	0.946
13	Surface = Wet	Speed = 60	0.946
14	AADT = 50,000–60,000	Surface = Dry	0.945
15	AADT = 40,000–50,000	Skid = 30.01–40.00	0.943
16	AADT = 30,000–40,000	Surface = Dry	0.941
17	AADT = 60,000	Speed = 60	0.941
18	Shoulder[4] = Paved with warning	Speed = 60	0.941
19	AADT = 0–10,000	Shoulder = Paved	0.940
20	Width = 30.01–40.00	Surface = Dry	0.939
Undivided roadways			
1	Skid[6] = 50	AADT = 0–10,000	0.977
2	Skid = 50	Width = 20.01–30.00	0.971
3	Skid ≥ 50	Speed = 50–60	0.971
4	Skid = 40.01–50.00	Width = 20.01–30.00	0.967
5	AADT = 0–10,000	Width = 20.01–30.00	0.967
6	AADT = 0–10,000	Shoulder = Paved	0.965
7	Trucks[7] = 5.01–10.00	Width = 20.01–30.00	0.964
8	Severity[8] = Incapacitating inj.	Speed = 50–60	0.963
9	Skid = 40.01–50.00	Speed = 50–60	0.963
10	Severity = Non-incapacitating inj.	Shoulder = Paved	0.962
11	Skid = 40.01–50.00	Light[9] = Daylight	0.961
12	Severity = Fatal	Shoulder = Paved	0.959
13	Shoulder = Other	Speed = 50–60	0.959
14	Trucks = 20	Speed = 50–60	0.957

(continued)

Table 2 (continued)

No.	Antecedent	Consequent	Precision
15	Skid = 40.01–50.00	AADT = 0–10,000	0.957
16	Trucks = 5.01–10.00	Speed = 50–60	0.955
17	AADT = 0–10,000	Speed = 50–60	0.955
18	Severity = Fatal	Width = 20.01–30.00	0.955
19	Severity = Fatal	AADT = 0–10,000	0.955
20	Trucks = 20	Weather[10] = Non-inclement	0.954

Note [1]AADT—annual average daily traffic (vpd), [2]Width—surface width (ft), [3]Speed—maximum posted speed (mph), [4]Shoulder—type of shoulder, [5]Surface—surface condition, [6]Skid—skid number, [7]Trucks—average percentage of trucks, [8]Severity—crash severity, [9]Light—lighting condition, [10]Weather—weather condition

Table 3 Three itemset precision rules for divided and undivided roadways

No.	Antecedent	Consequent	Precision
Divided roadways			
1	Weather = Non-inclement, AADT = 50,000–60,000	Shoulder = Paved with warning	0.989
2	Width = 30.01–40.00, Surface = Wet	Weather = Inclement	0.989
3	Width = 30.01–40.00, Weather = Non-inclement	Shoulder = Paved with warning	0.989
4	Weather = Non-inclement, AADT = 40,000–50,000	Shoulder = Paved with warning	0.988
5	Weather = Non-inclement, AADT = 60,000	Shoulder = Paved with warning	0.987
6	Shoulder = Paved with warning, Weather = Non-inclement	Surface = Dry	0.986
7	Width = 30.01–40.00, Trucks = 20	Shoulder = Paved with warning	0.986
8	AADT = 60,000, Trucks = 15.01–20.00	Shoulder = Paved with warning	0.986
9	Width = 30.01–40.00, Trucks = 15.01–20.00	Shoulder = Paved with warning	0.986
10	Weather = Non-inclement, AADT = 60,000	Surface = Dry	0.986
11	AADT = 60,000, Trucks = 10.01–15.00	Shoulder = Paved with warning	0.986
12	AADT = 60,000, Trucks = 10.01–15.00	Speed = 60	0.986
13	Width = 30.01–40.00, Weather = Non-inclement	Speed = 60	0.986

(continued)

Table 3 (continued)

No.	Antecedent	Consequent	Precision
14	Width = 30.01–40.00, AADT = 40,000–50,000	Shoulder = Paved with warning	0.986
15	Width = 30.01–40.00, AADT = 40,000–50,000	Speed = 60	0.986
16	Weather = Non-inclement, AADT = 50,000–60,000	Surface = Dry	0.986
17	Weather = Non-inclement, AADT = 50,000–60,000	Speed = 60	0.986
18	Width = 30.01–40.00, AADT = 50,000–60,000	Shoulder = Paved with warning	0.986
19	Width = 30.01–40.00, Light = Dark (no streetlight)	Shoulder = Paved with warning	0.985
20	Shoulder = Paved with warning, AADT = 20,000–30,000	Width = 20.01–30.00	0.985
Undivided roadways			
1	Skid = 40.01–50.00, Trucks ≥ 20	Width = 20.01–30.00	0.990
2	Skid = 40.01–50.00, Trucks ≥ 20	Speed = 50–60	0.989
3	Skid ≥ 50, Shoulder = Paved	Speed = 50–60	0.989
4	Skid ≥ 50, Light = Daylight	Width = 20.01–30.00	0.988
5	Skid ≥ 50, Light = Daylight	Speed = 50–60	0.988
6	Skid = 40.01–50.00, AADT = 0–10,000	Width = 20.01–30.00	0.988
7	Severity = Non-incapacitating injury, Trucks = 5.01–10.00	Width = 20.01–30.00	0.987
8	Skid ≥ 50, Severity = No Injury	AADT = 0–10,000	0.987
9	Skid ≥ 50, Severity = No Injury	Speed = 50–60	0.987
10	Skid ≥ 50, Speed = 50–60	Width = 20.01–30.00	0.986
11	Skid ≥ 50, Speed = 50–60	AADT = 0–10,000	0.986
12	Skid = 40.01–50.00, Trucks ≥ 20	AADT = 0–10,000	0.986
13	Skid = 40.01–50.00, Trucks ≥ 20	Surface = Dry	0.986
14	Skid ≥ 50, Shoulder = Paved	AADT = 0–10,000	0.986
15	Skid ≥ 50, AADT = 0–10,000	Width = 20.01–30.00	0.986
16	Skid ≥ 50, Weather = Non-inclement	AADT = 0–10,000	0.985

(continued)

Table 3 (continued)

No.	Antecedent	Consequent	Precision
17	Severity = Non-incapacitating injury, AADT = 0–10,000	Shoulder = Paved	0.985
18	Skid = 40.01–50.00, AADT = 0–10,000	Shoulder = Paved	0.985
19	Skid = 40.01–50.00, AADT = 0–10,000	Surface = Dry	0.985
20	AADT = 0–10,000, Trucks = 5.01–10.00	Speed = 50–60	0.985

co-occurrence of these three attributes in divided roadway dataset is over 98%. On the other hand, the average percentage of trucks exhibits significant dominance on undivided roadways. For example, the top rule for undivided roadways is: {*Skid Number* = 40–50, *Average Percentage of Trucks* = 20%} → *Roadway Width* = 220–30 ft with a precision value of 0.990. This rule shows that the co-occurrence of these three attributes in divided roadway dataset is 99% (Table 3).

The following findings below summarize the contents of knowledge extraction from the top twenty precision rules (two itemset and three itemset) for both divided and undivided roadways:

- Improper passing crashes on divided roadways associate with higher AADT, wider roadways, and higher speed for two itemset rules. On the other hand, undivided roadways in corporate lower AADT, narrower roadways, low to medium percentage of trucks, and severe/fatal crashes.
- Fatal or severe injuries are dominant on undivided roadways for two itemset rules.
- For three itemset rules, inclement weather contributes significantly on improper passing crashes on divided roadways.
- For three itemset rules, average percentage of trucks contributes significantly on improper passing crashes on undivided roadways.

6 Conclusions

The current study used Florida SHRP2 RID crash data for six years (2005–2010) to identify the key issues associated with improper passing crashes. Facility types on rural roadways (divided and undivided) show significant differences in the descriptive statistics of the geometric and environmental variables. The undivided roadways show a higher likelihood of fatal and injury crashes compared to divided roadways. The research team used an unsupervised data mining method (Association Rules NB Miner) to generate two itemset and three itemset rules. This method is an alternative of conventional association rules mining with an advantage of having no need of predefined support and confidence thresholds to generating rules. The current method is robust because it identifies rules with higher precision (by using a single

performance measure) from a large dataset by identifying key co-occurrence, while the conventional association rules use three parameters as performance measures. For two itemset precision rules in divided roadways, paved shoulder with warning, high AADT, and high speed are dominant in both antecedents and consequents. On the other hand, for undivided roadways, the dominant associated factors are crash severity, low AADT, percentage of trucks, and skid number. For three itemset rules while considering divided roadways, the dominant factors are weather, high AADT, and percentage of trucks. For divided roadways, skid resistance, low AADT, and crash severity are dominant. The findings of this research are consistent with those of previous studies. The top twenty rules for both roadways clearly show that facility type (either divided or undivided) plays a significant role in the higher likelihood of injury or fatal crashes. However, divided roadways also need attention in reducing improper passing crashes on roadways with high AADT and speed. The findings of the current study will help the safety professionals in mitigating improper passing crashes and crash severities.

Based on prior study results, we recommend that the addition of passing lanes and shoulders significant safety improvement in reducing the potential number of fatal crashes on rural roads [27, 28]. Interventions that improve roadway safety have been identified as one of the top priorities for roadway users [29]. Forward collision warning technology has also shown promise in its ability to track objects in front of the vehicle and provide feedback of an impending collision, a valuable system to assist with overtaking maneuvers [12].

The reason of using RID in this study is to relate the roadways with NDS data in this ongoing project to identify the relationship between improper passing and passing behavior of the drivers. One of the limitations of this study is that it only considers geometric variables to conduct analysis. Therefore, future research should investigate at least two major issues: identify exact passing permitted and no-passing locations from the spatially coded crash information to perform a more robust analysis; use NDS data for exploring behavioral perspectives.

References

1. Manual on Uniform Traffic Control Devices (2009) For streets and highways. U.S. Dept. of Transportation, Federal Highway Administration, Washington, DC
2. Carlson P, Miles J, Johnson P (2016) Daytime high-speed passing Maneuvers observed on rural two-lane, two-way highway: findings and implications. Transp Res Rec: J Transp Res Board 1961:9–15
3. Jenkins J, Rilett L (2005) Classifying passing Maneuvers: a behavioral approach. Transp Res Rec: J Transp Res Board 1937:14–21
4. Romana M (1999) Passing activity on two-lane highways in Spain. Transp Res Rec: J Transp Res Board 1678:90–95
5. Gates T, Savolainen P, Datta T, Todd R, Russo B, Morena J (2012) Use of both centerline and shoulder rumble strips on high-speed two-lane rural roadways. Transp Res Rec: J Transp Res Board 2301:36–45

6. Hallmark S, Tyner S, Oneyear N, Carney C, Mcgehee D (2015) Evaluation of driving behavior on rural 2-lane curves using the SHRP 2 naturalistic driving study data. J Saf Res 54:17–27
7. Shackel SC, Parkin J (2014) Influence of road markings, lane widths and driver behaviour on proximity and speed of vehicles overtaking cyclists. Accid Anal Prev 73:100–108
8. Farah H, Toledo T (2010) Passing behavior on two-lane highways. Transp Res Part F: Traffic Psychol Behav 13(6):355–364
9. Papakostopoulos V, Nathanael D, Portouli E, Marmaras N (2015) The effects of changes in the traffic scene during overtaking. Accid Anal Prev 79:126–132
10. Levulis SJ, Delucia P, Jupe J (2015) Effects of oncoming vehicle size on overtaking judgments. Accid Anal Prev 82:163–170
11. Llorca, C, Moreno A, García A, Pérez-Zuriaga A (2013) Daytime and nighttime passing Maneuvers on a two-lane rural road in Spain. Transp Res Rec: J Transp Res Board 2358:3–11
12. Chen R, Kusano K, Gabler H (2015) Driver behavior during overtaking Maneuvers from the 100-car naturalistic driving study. Traffic Inj Prev 16(2):176–181
13. Kinnear N, Helman S, Wallbank C, Grayson G (2015) An experimental study of factors associated with driver frustration and overtaking intentions. Accid Anal Prev 79:221–230
14. Vlahogianni E, Golias J (2012) Bayesian modeling of the microscopic traffic characteristics of overtaking in two-lane highways. Transp Res Part F: Traffic Psychol Behav 15(3):348–357
15. Das S, Kong X, Tsapakis I (2019) Hit and run crash analysis using association rules mining. J Transp Saf Secur 1–20
16. Yu W (2019) Discovering frequent movement paths from taxi trajectory data using spatially embedded networks and association rules. IEEE Trans Intell Transp Syst 20(3)
17. Das S, Sun X, Goel S, Sun M, Rahman A, Dutta A (2020) Flooding related traffic crashes: findings from association rules. J Transp Saf Secur 1–19
18. Das S, Dutta A, Jalayer M, Bibeka A, Wu L (2018) Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: exploration using 'Eclat' association rules to promote safety. Int J Transp Sci Technol 7(2):114–123
19. Weng J, Li G (2019) Exploring shipping accident contributory factors using association rules. J Transp Saf Secur 11(1):36–57
20. Das S, Sun X, Dutta A (2019) Patterns of rainy weather crashes: applying rules mining. J Transp Saf Secur 1–23
21. Weng J, Zhu J, Yan X, Liu Z (2016) Investigation of work zone crash casualty patterns using association rules. Accid Anal Prev 92:43–52
22. Das S, Dutta A, Avelar R, Dixon K, Sun X, Jalayer M (2020) Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. Int J Urban Sci 23(1):30–48
23. SHRP2 RID. http://www.ctre.iastate.edu/shrp2-rid/. Last accessed 2020/03/13
24. Hashler M (2006) A model-based frequency constraint for mining associations from transaction data. Data Min Knowl Disc 13:137–166
25. R Development Core Team. R (2013) A language and environment for statistical computing. Version 2.10.1. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org. Last accessed 2020/03/13
26. Hashler MR package 'arulesNBMiner'. https://cran.r-project.org/web/packages/arulesNBMiner/arulesNBMiner.pdf. Last accessed 2020/03/13
27. Schrock S, Parsons R, Zeng H (2011) Estimation of safety effectiveness of widening shoulders and adding passing lanes on rural two-lane roads. Transp Res Rec: J Transp Res Board 2203:57–63
28. Brewer M, Venglar S, Fitzpatrick K, Ding L, Park B (2012) Super 2 highways in Texas. Transp Res Rec: J Transp Res Board 2301:46–54
29. Mutabazi M, Russell E, Stokes R (1998) Drivers' attitudes, understanding, and acceptance of passing lanes in Kansas. Transp Res Rec: J Transp Res Board 1628:25–33