

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/340482669>

# Automated vehicle collisions in California: Applying Bayesian latent class model

Article in *IATSS Research* · April 2020

DOI: 10.1016/j.iatssr.2020.03.001

CITATIONS

0

READS

59

3 authors:



**Subasish Das**

Texas A&M University

85 PUBLICATIONS 309 CITATIONS

[SEE PROFILE](#)



**Anandi Dutta**

University of Louisiana at Lafayette

22 PUBLICATIONS 108 CITATIONS

[SEE PROFILE](#)



**Ioannis Tsapakis**

Texas A&M University

24 PUBLICATIONS 276 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



STANDARD [View project](#)



Coauthor Network Project [View project](#)

Contents lists available at [ScienceDirect](#)

IATSS Research



## Research Article

## Automated vehicle collisions in California: Applying Bayesian latent class model

Subasish Das <sup>a,\*</sup>, Anandi Dutta <sup>b</sup>, Ioannis Tsapakis <sup>c</sup><sup>a</sup> Texas A&M Transportation Institute, 1111 RELUIS Parkway, Room 4414, Bryan, TX 77807, United States of America<sup>b</sup> Department of Computer Science, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249-0667, United States of America<sup>c</sup> Texas A&M Transportation Institute, 3500 NW Loop 410, Suite 315, San Antonio, TX 78229, United States of America.

## ARTICLE INFO

## Article history:

Received 19 December 2019

Received in revised form 17 March 2020

Accepted 17 March 2020

Available online xxxxx

## Keywords:

Automated vehicle

Bayesian model

Traffic collision

## ABSTRACT

The emerging technology of automated vehicles (AV) has been rapidly advancing and is accompanied by various positive and negative potentials. The new technology is expected to affect costs mainly by reducing the number of collisions and travel time, as well as improving fuel efficiency and parking benefits. On the other hand, safety outcomes from AV deployment is a critical issue. Ensuring the safety of AVs requires a multi-disciplinary approach that monitors every aspect of these vehicles. The California Department of Motor Vehicles has mandated that AV collision reports be made public in recent years. This study collected the scanned collision reports filed by different manufacturers that are assessing AVs in California (September 2014 to May 2019). The collected data offers critical information on AV collision frequencies and associated contributing factors. This study provides an in-depth exploratory analysis of the critical variables. We demonstrated a variational inference algorithm for Bayesian latent class models. The Bayesian latent class model identified six classes of collision patterns. Classes associated with turning, multi-vehicle collisions, dark lighting conditions with streetlights, and sideswipe and rear-end collisions were also associated with a higher proportion of injury severity levels. The authors anticipate that these results will provide a significant contribution to the area of AV and safety outcomes.

© 2020 International Association of Traffic and Safety Sciences. Production and hosting by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

An automated vehicle (AV) utilizes artificial intelligence (AI) and mechanics that can assist human drivers or operators by examining the adjacent situation (e.g., other roadway users, traffic signs and signals, pavement markings) and monitoring a vehicle's speed and its steering control. Because the performance of certain control functions like accelerating or decelerating depends on the signals and inputs that are gathered from the neighboring settings (e.g., a traffic light turning red from orange), both functions are undeniably associated and dependent on one another. The Society of Automotive Engineers (SAE) characterizes six levels of autonomy that examines the degree to which the automated technology is able to provide assistance and support for the driving tasks [1].

The California Department of Motor Vehicles (CA DMV) requires that trained human operators should remain behind the wheel while assessing AVs on public roads, irrespective of the autonomy levels of

the vehicles to promote safety. In addition to human drivers, the CADMV mandated that the recent years of AV collision reports must be made public. The first type of reporting is a brief list of all occurrences of AV disengagements (during failure or difficult to control events, human operator will take control by putting the automated feature of the car disengaged). The second type of report supplies a thorough summary of events in which a crash/collision/event and/or damage to property and injuries take place. The current study is limited to the second type of reports on AV collisions.

To ensure the safety requirement, there should be careful strategies at the present time by enacting more rigid regulations. It is very important for the policy makers to have a clear understanding of the safety concerns associated with AV collisions. An extensive analysis on the available AV collision data can help in transforming the regulations more effective. California based AV collision data provides crucial information on AV collisions including key contributing factors. The intent of this paper is to carefully investigate the AV collision data to shed further light on heterogeneity effects in roadway geometric features, human interactions, and other attributes with respect to occurrences of AV collisions. Out of different methodological frameworks, we found that Bayesian clustering technique is suitable for the current research context.

\* Corresponding author.

E-mail addresses: [s-das@tti.tamu.edu](mailto:s-das@tti.tamu.edu) (S. Das), [anandi.dutta@utsa.edu](mailto:anandi.dutta@utsa.edu) (A. Dutta),[i-tsapakis@tti.tamu.edu](mailto:i-tsapakis@tti.tamu.edu) (I. Tsapakis).

Peer review under responsibility of International Association of Traffic and Safety Sciences.

<https://doi.org/10.1016/j.iatssr.2020.03.001>0386-1112/© 2020 International Association of Traffic and Safety Sciences. Production and hosting by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

The paper is organized as follows. The next section provides a brief overview of the ongoing AV collision and disengagement related studies. Section three provides a short overview of Bayesian clustering theory. The next section supplies details of data collection and exploratory analysis. Section five explains the key findings from Bayesian clustering and text mining. In the last section, we provide conclusions.

## 2. Literature review

One of the major causes for disengagement in AVs is the driver's lack of attention because of the overreliance on vehicle automation and his/her use of the automation incompatible with the guidance and alerts from the manufacturer [2]. In this regard, Trovato [3] provided a structure that summarizes the fundamental behavior of a robot and automatically computes intelligent maneuvers for collision-free maneuvering and control of an AV. The capacity of AVs to improve safety and riding experience is normally regarded with apprehension. Shim et al. [4] generated a collision-avoidance system for an AV application and determined its efficient performance collision avoidance actions. Based on the evaluation of driving behavior and collision risk, Tak et al. [5] proposed a spacing policy based on Asymmetric Collision Risk (ACR). Compared to other spacing policies, the findings concluded similar patterns in the ACR spacing policy with a human driver with less acceleration/deceleration actions and smoother trajectories. Using the information gathered by a laser-scanner sensor, Jimenez et al. [6] developed a collision-avoidance system capable of taking two actions in case of danger. When executed in a vehicle and tested with other vehicles and non-motorists traveling along a private test track, the system delivered satisfactory outcomes. To analyze possible hazards on straight or curved roads, Cao et al. [7] generated a comprehensive architecture of an active collision-avoidance system for AVs. Under various conditions, the simulation results determined the success of a host vehicle capable of making a collision avoidance maneuver without human driver interference.

When lateral control was delegated to automation, Navarro et al. [8] analyzed the unexpected obstacle avoidance maneuvers by conducting a simulation study. In comparison to driving without automation, drivers returning to manual control from automatic steering were found to be less effective at maneuvering around obstacles. Dixit et al. [9] observed a significant correlation between the number of collisions, the traveled AV miles, and the vehicle's reaction time to take control in the event of disengagement and found an average distribution of 0.83 s across different companies.

To encourage safety and transparency for customers, the CA DMV has ordered that accounts of collisions involving AVs be drawn up and rendered open to the public. Correspondingly, Favaro et al. [10] generated a detailed assessment of the collision records submitted by different manufacturers studying AVs in California. The data provided significant information about the dynamics of AV collisions linked to the most common kinds of collisions and effects, frequencies of collisions, and other contributing variables. Additionally, Favaro et al. [11] analyzed AV disengagement trends for identification of timely and safe return of vehicle control to the human driver. The study provided an inclusive outline of the fragmented data such trends of disengagement reporting, average mileage driven before failure, associated frequencies, etc. acquired from AV manufacturers testing on California public roads from 2014 to 2017.

Poland et al. [12] evaluated the relationship between the driver and the Society of Automobile Engineers' (SAE) level 2 AV, including the vehicle damage, restrictions inflicted by the vehicle on the driver using scene evidence, vehicle data, and information from both drivers such as experience, medical information, phone records, experience, and computer systems. To improve automated response time, Roldan et al. [13] tested a theory using two driving simulator studies that enabled participants to drive simulated in a controlled environment

using cooperative adaptive cruise control (CACC) vehicles that directly transmit vehicle-to-vehicle data. The goal of the experiment was to evaluate the driver workload while using CACC and adaptive cruise control (ACC) technologies and determine whether CACC improves or lowers collision prevention when driver action is essential. Boggs et al. [14] established a distinctive database from the CA DMV 66 manufacturer-reported Traffic Collision Reports (OL 316) that included responses to close-ended collision issues and text mining narratives. The findings showed that most AV collisions occurred in completely autonomous mode (65.2%), and the likelihood of AVs being hit was greater than the effect before car takeovers and conventionally powered cars.

Despite the great chance AVs have to enhance the safety of traffic, they also present some major concerns. While AVs may decrease human error-induced accidents, they still encounter sensing and technology failures as well as mixed traffic environment decision-making errors. Khattak et al. [15] combined and analyzed both disengagements data and accidents with a rigorous modeling strategy. The findings suggested that disengagements are a part of the safe performance of AVs, and the activation of disengagement alerts may prevent certain existing technology errors. Lee et al. [16] developed a hazard predictive collision prevention system (RPCAS) and evaluated its effect on the safety of pedestrians and vehicles. Relative to current CASs, the findings showed that the RPCAS can effectively decrease the danger of rear-ending collision with less severe handling.

Xu et al. [17] used descriptive statistical analysis to examine the trends and features of the connected and autonomous car (CAV) involved accidents. The findings indicated that the primary factors adding to the severity stage of CAV related accidents were the CAV driving mode, roadside parking, collision location, one-way road, and rear-end collision. Lodinger and DeLucia [18] assessed time-to-collision (TTC) judgments between driving modes (manual and autonomous driving) to verify whether the automation only affected responses (i.e., speed change) or also affected visual perception (i.e., TTC estimation). Yu et al. [19] investigated the impact on the safety and effectiveness of AVs in scenarios with mixed traffic and AV-only vehicle environments. When the market penetration rate of AVs is small, the researchers found that AV-only routes experience an increase in effectiveness and safety. Under Infrastructure-to-Vehicle (I2V) and Vehicle-to-Vehicle (V2V), Rahman et al. [20] studied the safety effect of Connected Vehicles (CV) and Connected Vehicles with Lower Automation Level (CVLLA) Communication Technologies. A substantial increase in safety was a result of the implementation of CV and CVLLA methods in both sections and arterial intersections. Using three-car designs to anticipate the vehicle's behavior in a randomized situation, Rao et al. [21] simulated the longitudinal behavior of AVs in traffic jam scenarios.

As both contextual and circumferential factors should be regarded simultaneously, the evaluation of real-time threat for strategic and operational automated driving is extremely difficult. Under the collective structure of Dynamic Bayesian Networks (DBN) and interaction-aware movement models, Katrakazas et al. [22] developed a new risk assessment method that incorporates a network-level crash prediction with a vehicle-based real-time threat estimation. Results showed that a finely-tuned classification of crash prediction provides a vital indication for better risk assessment by AVs.

The literature review reveals that the previous exploratory study conducted by Favaro et al. [10] analyzed only 26 AV collision data. Two recent studies have explored CA DMV disengagement data [24,25]. However, the current study is focused on AV collisions, not AV disengagement reports. Our study is an extension of Favaro et al. [10], which mainly focused on the exploratory nature of the AV collisions. The uniqueness of this study is to perform an extensive study by unearthing the hidden trends and associated factors with the usage a larger set of AV collision data.

3. Methodology

3.1. Bayesian clustering

Ahlmann-Eltze and Yau proposed a variant of the latent class model (LCM) structure where the individuals are clustered into  $K$  classes [23]. The model can be summarized as follows:

$$\lambda | \alpha \sim \text{Dirichlet}(\alpha) \text{ or } \text{DirichletProcess}(\alpha) \tag{1}$$

$$z_i | \lambda \sim \text{Multinomial}(\lambda) \tag{2}$$

$$U_{j,k} | \beta \sim \text{Dirichlet}(\beta) \tag{3}$$

$$X_{i,j} | U_j, z_i = k \sim \text{Multinomial}(U_{j,k}) \tag{4}$$

Here,  $\alpha$  and  $\beta$  are hyper-parameters that are interpreted externally and manage model sparsity. Eq. (1) defines that class size, which is governed by a Dirichlet (in the case of a simple LCM) or by a Dirichlet Process. First, the derivation for the simple LCM is explained, then the way to expand it to the nonparametric case is presented.  $z$  is a vector which incorporates the latent class assignment for each entry. Here,  $U$  is a 3-way tensor of size  $J \times K \times R$  and has the probability for response  $r$  from an individual from class  $k$  for  $j$ . Eq. (4) stipulates the response of an individual  $i$  (belongs to class  $k$  from a Multinomial distribution according to the probability vector  $U_{j,k}$ ). The equation of joint distribution of the model is:

$$p(\lambda, z, U, X | \alpha, \beta) = p(\lambda | \alpha) \prod_{i=1}^I p(z_i | \lambda) \prod_{j=1}^J \prod_{k=1}^K p(U_{j,k} | \beta) \times \prod_{i=1}^I \prod_{j=1}^J \prod_{k=1}^K p(X_{i,j} | U_{j,k})^{I(z_i=k)} \tag{5}$$

For this model, solving for the maximum likelihood solution would lead to an EM algorithm. A variational inference (VI) method is generated to accurately propagate uncertainty and to conclude an suitable number of latent classes through the model. The idea of VI is to establish a probability model  $q$  and adjust its parameters to provide estimates of the original model  $p$ . The objective of choosing  $q$  as the mean-field approximation of  $p$ , permits the user to record the variational distribution:

$$q(\lambda, z, U) = q(\lambda)q(z)q(U), q(\lambda, z, U) = q(\lambda; \omega) \prod_{i=1}^I q(z_i; \zeta_i) \prod_{k=1}^K \prod_{j=1}^J q(U_{j,k}; \phi_{j,k}) \tag{6}$$

where  $\omega$ ,  $\zeta$  and  $\phi$  are the free variational parameters. It can be written as:

$$q(\lambda; \omega) = \text{Dirichlet}(\omega) \\ q(z_i = k; \zeta_i) = \zeta_{i,k} \\ q(U_{j,k}; \phi_{j,k}) = \text{Dirichlet}(\phi_{j,k}) \tag{7}$$

The Kullback-Leibler (KL) divergence has been utilized to measure the approximation, which allows the user to maximize the evidence lower bound (ELBO). The ultimate goal is to maximize the ELBO by minimizing the KL divergence.

4. Data collection and analysis

4.1. Data collection

We collected the crash report pdfs provided by the CA DMV. According to the DMV website (<https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/testing>): ‘under the testing regulations, manufacturers are required to provide DMV with a Report of Traffic Collision Involving an AV within 10 days after the collision’. We developed a structured database from the information of the scanned pdfs of AV collisions in California during September 2014–May 2019. The total number of reported collisions used in this study was 151. Fig. 1 shows the cumulative number of collisions or collisions during 2014–2019. The companies that deployed AVs are also shown in the plot. The graph shows that on October 24, 2014, Delphi was the first manufacturer to experience an AV collision in California. The figure illustrates the trend of a slow increase in the cumulative number of AV collisions from October 2014 until October 2017. After October 2017, there was a sharp increase in AV collisions as a greater number of companies deployed AVs.

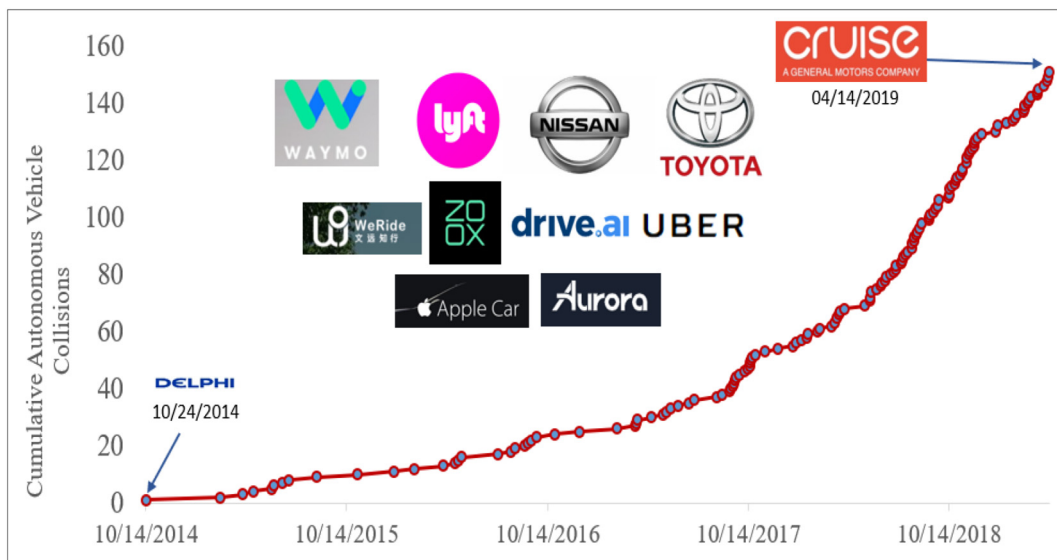












Fig. 1. Cumulative AV collisions in California.



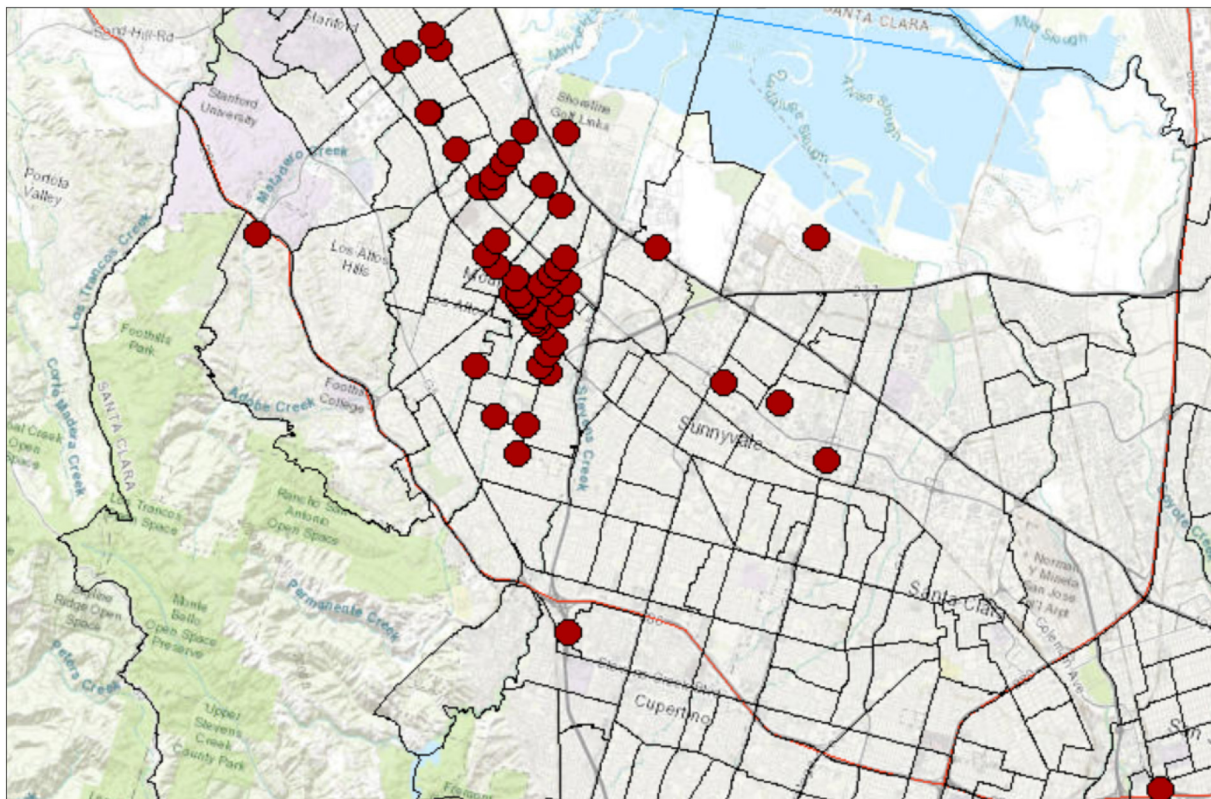
**Table 1**  
Number of collisions and automated miles by the AV companies.

Company	Traffic Collisions	Autonomous Miles
 CRUISE A GENERAL MOTORS COMPANY	71	131,676
 WAYMO	55	352,545
	7	2,245
	4	--
<b>UBER</b>	3	--
 Aurora	3	--
 TOYOTA	2	--
 Apple Car	2	--
 Delphi Technologies	1	1,820
 NISSAN	1	5,007
<b>drive.ai</b>	1	6,572
 WeRide	1	--

#### 4.2. Exploratory data analysis

Table 1 shows lists the number of traffic collisions and the number of automated miles driven for twelve companies since their AV deployment date. Waymo had the greatest number of automated miles

(352,545 miles) with the second-highest number of traffic collisions ( $N = 55$ ). Cruise had the most traffic collisions ( $N = 71$ ), and it had the second greatest number of automated miles (131,676 miles). Delphi, Nissan, Drive.ai, and WeRide all only had one traffic collision reported. In general, a greater number of automated miles was



**Fig. 2.** Location of AV collisions in Santa Clara.

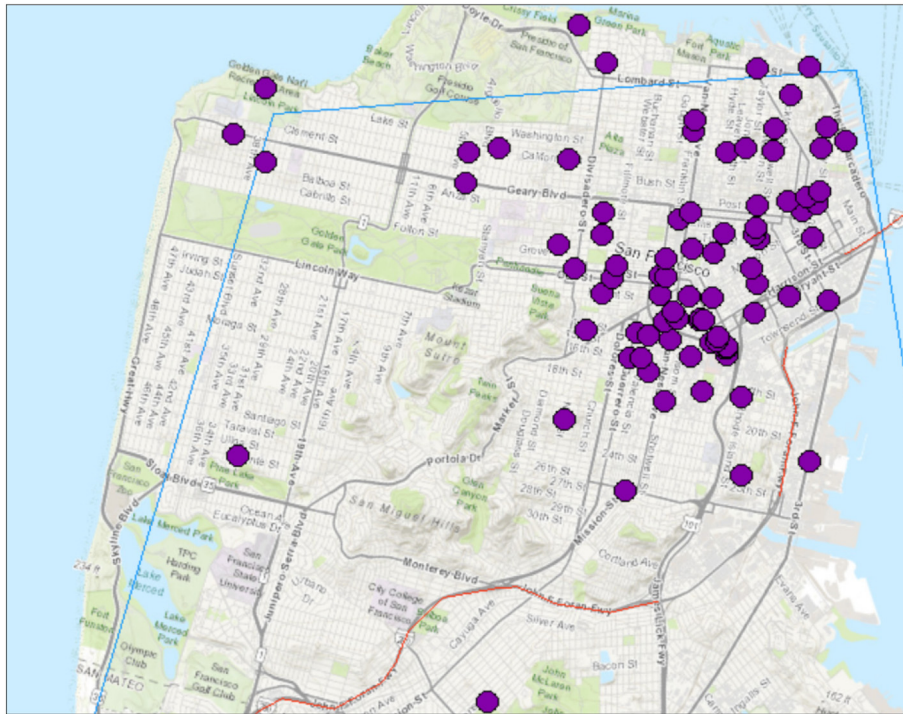


Fig. 3. Location of AV collisions in San Francisco.

associated with a greater number of traffic collisions; however, it is important to note that the data for automated miles contained missing values for half of the companies listed.

Fig. 2 illustrates a visual representation of where AV collisions took place around Santa Clara. CA DMV reports do not provide exact spatial locations, in the form of latitude and longitude, of the collisions. The report includes the intersection or cross street information. We used Google Map Application Programming Interface (API) to determine the locations of the collisions. The spatial maps will clarify the spatial distribution patterns of the AV collisions. Each red dot is one collision occurrence. As shown in the figure, most of the collisions occurred near the Mountain View area.

Fig. 3 provides a visual representation of where AV collisions took place around San Francisco. Each purple dot represents the location of one collision. As shown in the figure, the majority of the collisions were concentrated in the northeastern area.

Fig. 4 shows the frequency of AV collisions for different times of day and days of the week. Times that had a higher frequency of collision occurrences are represented by larger circles. The times and days that had the highest frequencies of collision occurrences are the mornings (07–12 pm) of Thursdays (16 collisions), 1–6 pm of Fridays (14 collisions), and 1–6 pm of Wednesdays (11 collisions). One finding from this table is that 1–6 am consistently has a very low frequency of collision occurrences for all days of the week.

Table 2 illustrates the frequency of AV collisions for each month from January 2014 to April 2019. Months that had a higher frequency of collision occurrences are represented by a darker shade of red. The months with the highest frequencies were November 2018 (12 collisions), August 2018 (10 collisions), July 2018 (8 collisions), September 2018 (8 collisions), and October 2018 (8 collisions). The table shows an overall trend of collision frequency increasing over time. This is likely due to the increasing prevalence of AVs on the roads.

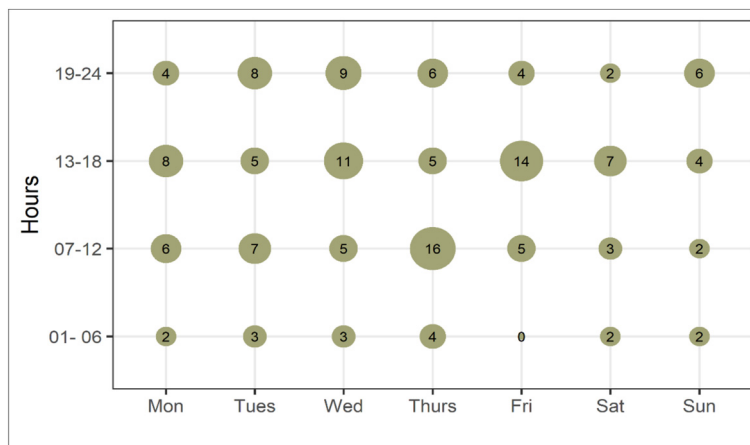


Fig. 4. AV collisions by day of week and hours.

**Table 2**  
Number of collisions by month.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2014	0	0	0	0	0	0	0	0	0	1	0	0
2015	0	1	0	2	1	2	1	1	0	0	1	0
2016	1	1	0	2	2		1	2	4	1	0	1
2017	0	1	3	1	3	2	1	2	7	7	1	1
2018	5	2	6	1	6	6	8	10	8	8	12	3
2019	4	6	6	6								

Fig. 5 shows nine bar plots to represent the distribution of AV collisions for different variables. Most the collision types were a rear-end collision (58 collisions), and the second most common type was side swipe collisions (17 collisions). The damage level of the vehicle was most commonly minor (63 collisions) or moderate (17 collisions); only two collisions were reported as major vehicle damage. A vast majority of collisions (143 collisions) did not cause severe injuries to the vehicle operator. Of the collisions studied, a majority of them (89

collisions) involved a vehicle that was in automated driving mode at the time of the collision. Furthermore, a majority of the collisions (128 collisions) involved two vehicles, rather than a single-vehicle or multi-vehicle collision. Most of the collisions occurred in daylight (68 collisions) and during clear weather (72 collisions). A majority of collisions occurred when the vehicle involved was moving (93 collisions), and the most common prior event to the collision was being at a stop (43 collisions).

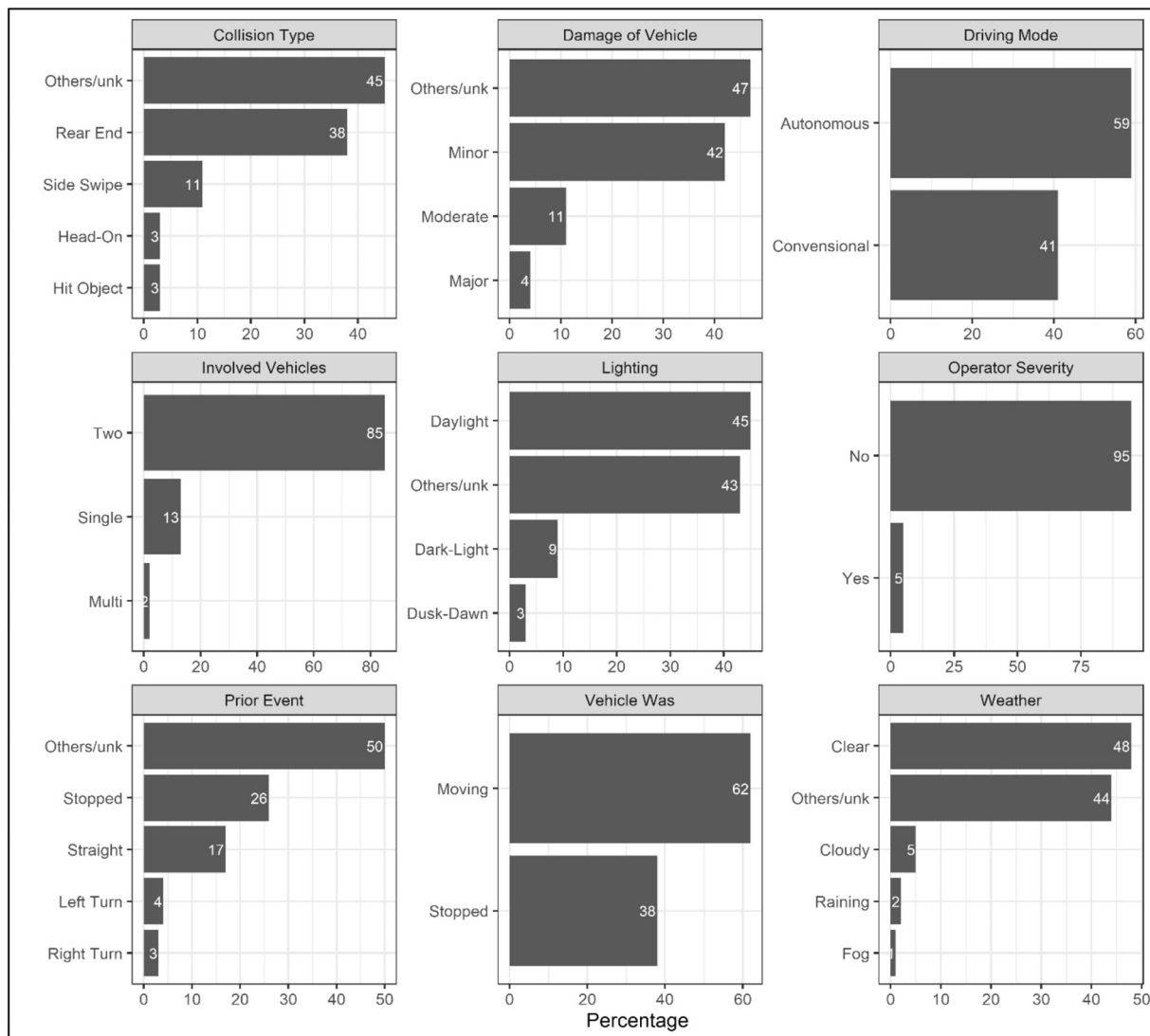


Fig. 5. Bar plot showing the distribution of AV collisions by different variable categories.

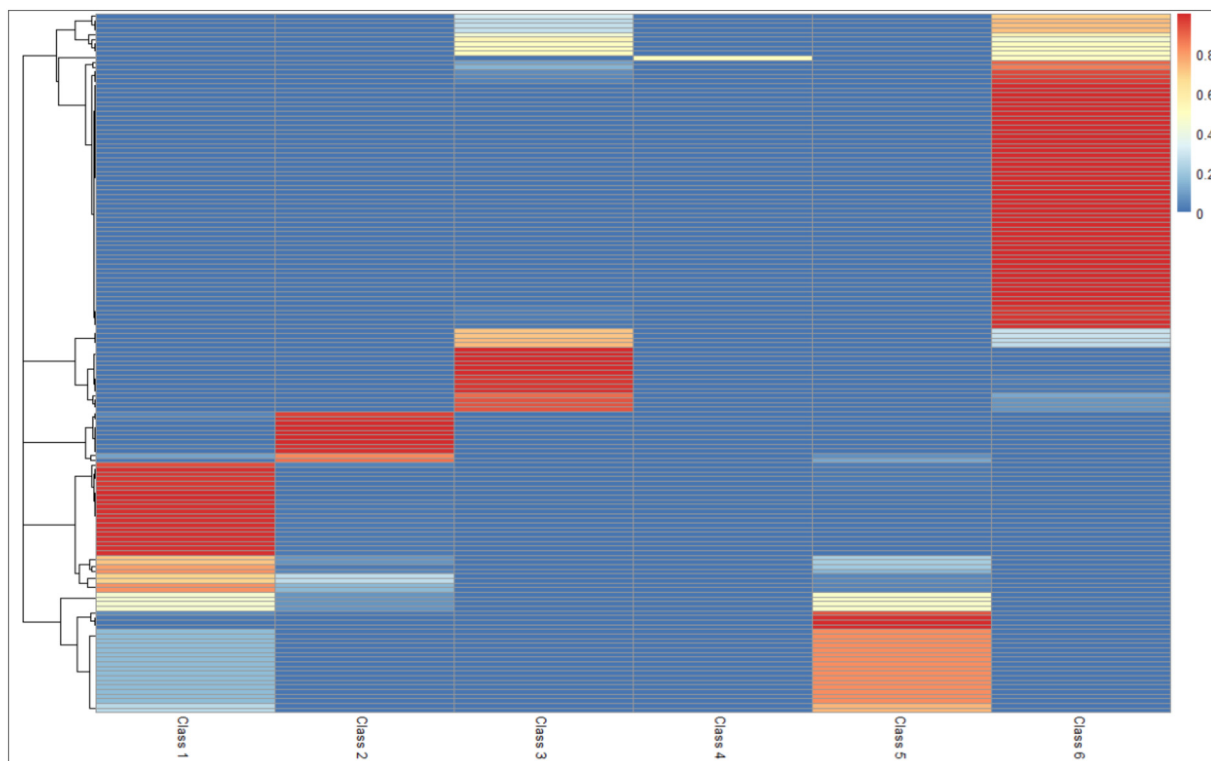


Fig. 6. Heatmap showing six classes of participant response groups.

## 5. Results and discussions

### 5.1. Results of Bayesian clustering

To perform the analysis, we used ‘mixdir’ [23], an open-source R package. We applied a hierarchical Dirichlet Process mixture of multinomial distributions. We used probabilistic latent class model (LCM) as it can assemble individuals into latent classes and reduce the dimensionality of hierarchical data. Moreover, we were able to deduce a suitable number of latent classes. These unique features of Bayesian clustering can accomplish the research goal of this study. Fig. 6 is a heat map that illustrates the six classes of participant response groups. The rows in this plot indicate each of the AV collision events. Based on the combinations of attributes among these collisions, the classes are developed. The blue color in the cells indicates low probability scores, yellow indicates mid-range probability scores, and red indicates the highest probability scores. This method is also useful because it produces probabilistic assignments of individual collisions to the latent classes. This method clusters the AV collision data and uncovers interesting latent structure. The values of Table 3 are used for model interpretations. It shows the ‘between Class’ proportions of the exogenous variables. Some of the key findings are below:

- Class 1 consists of 28 reports; most of the collisions were associated with automated driving mode and no operator severity. This class shows the highest percentage of collisions that occurred when the vehicles were moving and during the weekdays. Moreover, two-vehicle collisions occurred higher than in other conditions. The most frequent weather condition was clear and most frequent lighting condition was daylight. Also, the collision type is reported as unknown.
- Class 2 consists of 11 reports; these reports had higher counts of conventional vehicles collisions compared to automated ones. Most of the incidents were single-vehicle collisions with no severe injuries to the operator.
- Class 3 consists of 21 reports in which most collisions were

accompanied by severe operator injuries and left and right turn and straight as their prior event. Furthermore, most of the collisions took place in dark lighting. The percentage of conventional vehicle collisions was also higher than the percentage of automated ones. Moreover, class 1 and 3 are the only classes in which the highest number of collisions were side swipe collisions and in which a majority of the collisions occurred on weekends.

- Class 4 is a limited cluster, containing only one report. The report indicated a multi-vehicle collision resulting in major damages.
- Class 5 shows the highest percentages of multi-vehicle collisions that have occurred while moving. The most frequent weather condition for this cluster was ‘unknown,’ and the percentage of the operator severity is more than two times higher than no severity condition.
- Class 6 consists of 64 reports, making it the largest class. This cluster is highly associated with conventional two-vehicle collisions, which occurred while the prior event was ‘being stopped.’ It also includes the highest percentages in minor vehicle damage, and cloudy or foggy weather condition, with property damage only (PDO) collisions. Furthermore, head on and rear end were reported as the first and the second highest percentages among all collision types, respectively.

### 5.2. Findings from collision narratives

This study also gathered police collision narratives to perform text mining. The text mining pipeline (stop word and redundant word removal, word stemming, and word token development) was used to determine a set of n-grams (word groups that are in a sequence in a sentence). After exploring several n-grams, we developed trigrams from two corpora that are developed based on the vehicle automation mode during the collision event. The odds of word  $w$  in group  $i$ 's usage

can be written as  $O_{kw}^{(i)} = f_{kw}^{(i)} / (1 - f_{kw}^{(i)})$  [where,  $f_{kw}^{(i)} = \frac{y_{kw}^{(i)}}{n_k^{(i)}}$ ]. The term

$y_{kw}^{(i)}$  denotes the  $W$ -vector of word counts from documents of class  $i$  in topic  $k$ . The odds ratio between these groups can be expressed as



**Table 3**  
Distribution of variable attributes by in between classes.

Attribute	Count	Class 1 (28)	Class 2 (11)	Class 3 (21)	Class 4 (1)	Class 5 (26)	Class 6 (64)
<b>Driving Mode</b>							
Autonomous	89	21.35	4.49	12.36	0	24.72	37.08
Conventional	62	14.52	11.29	16.13	1.61	6.45	50
<b>Operator Severity</b>							
No	143	19.58	7.69	11.19	0.7	16.08	44.76
Yes	8	0	0	62.5	0	37.5	0
<b>Prior Event</b>							
Left Turn	6	16.67	0	50	0	0	33.33
Right Turn	5	0	0	60	0	0	40
Stopped	43	0	0	2.33	0	0	97.67
Straight	25	0	0	56	4	0	40
Other/Unknown	72	37.5	15.28	0	0	36.11	11.11
<b>Vehicle Was</b>							
Moving	93	10.75	9.68	21.51	1.08	27.96	29.03
Stopped	58	31.03	3.45	1.72	0	0	63.79
<b>Involved Vehicles</b>							
Single	20	0	50	10	0	0	40
Multi	3	0	0	33.33	33.33	33.33	0
Two	128	21.88	0.78	14.06	0	19.53	43.75
<b>Damage Vehicle</b>							
Major	2	0	0	0	50	0	50
Minor	63	0	0	17.46	0	0	82.54
Moderate	17	0	0	52.94	0	0	47.06
Other/Unknown	69	40.58	15.94	1.45	0	37.68	4.35
<b>Day of Week</b>							
Weekday	123	17.89	8.13	10.57	0.81	18.7	43.9
Weekend	28	21.43	3.57	28.57	0	10.71	35.71
<b>Weather</b>							
Clear	72	0	0	29.17	1.39	0	69.44
Cloudy	8	0	0	0	0	0	100
Fog	2	0	0	0	0	0	100
Raining	3	0	0	0	0	0	100
Other/Unknown	66	42.42	16.67	0	0	39.39	1.52
<b>Lighting Condition</b>							
Dark-Light	13	0	0	53.85	0	0	46.15
Daylight	68	0	0	19.12	1.47	0	79.41
Dusk-Dawn	4	0	0	25	0	0	75
Other/Unknown	66	42.42	16.67	0	0	39.39	1.52
<b>Collison Type</b>							
Head-On	5	0	0	0	0	0	100
Hit Object	4	0	0	0	25	0	75
Rear End	58	1.72	0	22.41	0	0	75.86
Side Swipe	17	0	0	47.06	0	0	52.94
Other/Unknown	67	40.3	16.42	0	0	38.81	4.48

$\theta_{kw}^{(M-P)} = O_{kw}^{(M)} / O_{kw}^{(P)}$ . Fig. 7 shows the log odds ratios of the top trigrams, the continuous sequence of three words from a document, from the collision reports for collisions in which the AV was in either autonomous or conventional mode prior to collision occurrence. In the present data, a collision report associated with autonomous as the prior mode is 1.45 times more likely to use a variant of 'av made contact' than a collision report associated with conventional as prior mode. It is important to note that the narrative texts from the collision reports are not detailed enough to separate the report types by prior driving mode in many cases. As the AV operators switch conditions from conventional to autonomous and vice versa, the narrative texts provide both autonomous and conventional driving mode information in the narratives, and it is difficult to distinguish which mode the vehicle was in prior to the collision. The current findings call for more detailed collision narrative documentation in collisions involving AVs.

**6. Conclusions**

AVs are expanding their market quickly, and with this expansion, some safety-related concerns raise significantly. Operators in fully AVs can be involved in non-driving tasks. However, if the automatic system fails, or becomes limited, the operators must take control of driving the vehicle through an appropriate and timely reaction. To understand the safety-related factors, it is essential to obtain enough data regarding the collision history and the contributing factors in AV collisions [9,11]. In this study, a comprehensive analysis was conducted using the data including the collision reports filed by various manufacturers from September 2014 to May 2019. We demonstrated a variational inference algorithm for Bayesian latent class models. They also applied the clustering algorithm to complex AV collision data, yielding good and interpretable results. The Bayesian latent class model identified

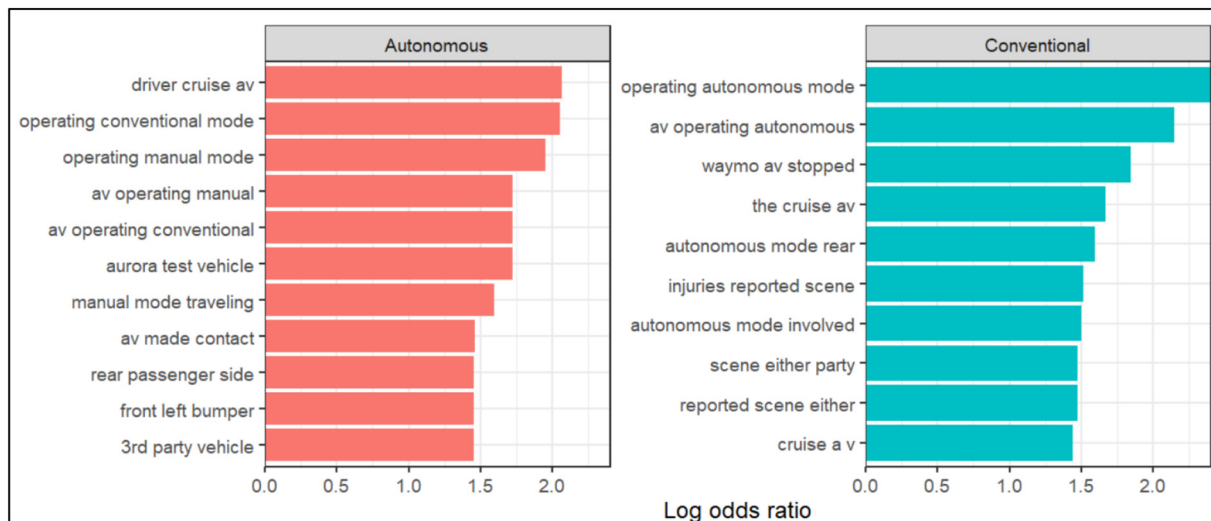


Fig. 7. Log odds ratio of the trigrams generated from the collision reports.

six classes of collision patterns based on different variables and collision traits. The variables included collision type, damage to the vehicle, operator injury severity, lighting conditions, the number of vehicles involved, weather conditions, the event prior to the collision, and whether the vehicle was moving or stopped. Classes associated with turning, multi-vehicle collisions, dark lighting conditions with streetlights, and sideswipe and rear-end collisions were also associated with a higher proportion of injury severity level. A significant finding demonstrated by Class 6 is that when a vehicle was in autonomous mode, there was a high likelihood of adverse weather collision occurrences when the vehicle's prior condition was stopped.

We also investigated collision narrative texts from police collision reports to determine whether they can be used to accurately identify the mode of the AVs prior to the collision. The calculated log odds ratio values showed that the current narrative documentation structure is not sufficient in determining the driving mode; this is because AV collisions are complex and distinctive in nature. There is a need for more advanced and robust collision narrative reporting in order to better investigate the association of automation levels with collision likelihood.

This unique study highlights the complexity and challenges of identifying key risk factors associated with AV collisions. This study extended the study conducted by Favaro et al. [10] with the inclusion of additional recent AV collisions and applied clustering method to identify the key clusters. The study is not without limitations. One limitation of this study is that the nonparametric extension, Dirichlet Process, used to overestimate the true number of latent classes. Further studies should aim to refine the algorithms to limit overestimation and mitigate this limitation.

## References

- [1] SAE International, Updated Visual Chart for Its "Levels of Driving Automation" Standard for Self-Driving Vehicles, <https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles> Accessed: July 2019.
- [2] National Transportation Safety Board, Highway Accident Report: Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016, 2017 63p.
- [3] K.I. Trovato, Collision-free maneuvering and control of an automated vehicle, *Advanced Vehicles and Infrastructure Systems 1997*, pp. 189–208.
- [4] T. Shim, G. Adireddy, H. Yuan, Automated vehicle collision avoidance system using path planning and model-predictive-control-based active front steering and wheel torque control. *Proceedings of the institution of mechanical engineers, part D, J. Automob. Eng.* 226 (6) (2012) 767–778.
- [5] S. Tak, H. Yeo, Chalmers University of Technology, SAFER Vehicle and Traffic Safety Centre. Asymmetric Collision Risk Spacing Policy for Longitudinal Control of Automated Driving Vehicle, 2015.
- [6] F. Jiménez, J.E. Naranjo, Ó. Gómez, Automated collision avoidance system based on accurate knowledge of the vehicle surroundings, *IET Intell. Transp. Syst.* 9 (1) (2015) 105–117.
- [7] H. Cao, X. Song, Z. Huang, L. Pan, Simulation research on emergency path planning of an active collision avoidance system combined with longitudinal control for an automated vehicle. *Proceedings of the institution of mechanical engineers, part D, J. Automob. Eng.* 230 (12) (2016) 1624–1653.
- [8] J. Navarro, M. François, F. Mars, Obstacle avoidance under automated steering: impact on driving and gaze behaviours, *Transport. Res. F: Traffic Psychol. Behav.* 43 (2016) 315–324.
- [9] V.V. Dixit, S. Chand, D.J. Nair, Automated vehicles: disengagements, accidents and reaction times, *PLoS One* 11 (12) (2016) <https://doi.org/10.1371/journal.pone.0168054>.
- [10] F.M. Favaro, N. Nader, S.O. Eurich, M. Tripp, N. Varadaraju, Examining accident reports involving automated vehicles in California, *PLoS One* 12 (9) (2017), e0184952. <https://doi.org/10.1371/journal.pone.0184952>.
- [11] F. Favaro, S. Eurich, N. Nader, Automated vehicles' disengagements: trends, triggers, and regulatory limitations, *Accid. Anal. Prev.* 110 (2018) 136–148, <https://doi.org/10.1016/j.aap.2017.11.001>.
- [12] K. Poland, M.P. McKay, D. Bruce, E. Becic, Fatal crash between a Car operating with automated control systems and a tractor-semitrailer truck, *Traffic Inj. Prev.* 19 (sup2) (2018) S153–S156.
- [13] S.M. Roldan, V.W. Inman, S.A. Balk, B.H. Philips, Semi-automated connected vehicle safety systems and collision avoidance: findings from two simulated cooperative adaptive cruise control studies, *IET J.* 88 (6) (2018) 30–35.
- [14] A. Boggs, A.J. Khattak, B. Wali, Analyzing Automated Vehicle Collisions in California: Application of a Bayesian Binary Logit Model, *Transportation Research Board 98th Annual Meeting*, Washington DC, 2019.
- [15] Z.H. Khattak, M.D. Fontaine, B.L. Smith, Transportation Research Board, An Exploratory Investigation of Disengagements and Collisions in Automated Vehicles, 2019.
- [16] D. Lee, S. Tak, S. Choi, H. Yeo, Development of risk predictive collision avoidance system and its impact on traffic and vehicular safety, *Transp. Res. Rec.* 2673 (7) (2019) 454–465.
- [17] C. Xu, Z. Ding, C. Wang, Transportation Research Board, Investigating the Characteristics of Connected and Automated Vehicle Involved Collisions, 2019.
- [18] N.R. Lodinger, P.R. DeLucia, Does automated driving affect time-to-collision judgments? *Transport. Res. F: Traffic Psychol. Behav.* 64 (2019) 25–37.
- [19] H. Yu, S. Tak, M. Park, H. Yeo, Impact of automated-vehicle-only lanes in mixed traffic conditions, *Transp. Res. Rec.* 2673 (9) (2019) 430–439.
- [20] M.S. Rahman, M. Abdel-Aty, J. Lee, M.H. Rahman, Safety benefits of arterials' crash risk under connected and automated vehicles, *Transp. Res. C* 100 (2019) 354–371.
- [21] S.J. Rao, T. Seitz, V.R.R. Lanka, G. Forkenbrock, Analysis and mathematical modeling of car-following behavior of automated vehicles for safety evaluation, Presented at the SAE Technical Paper, 2019.
- [22] C. Katrakazas, M. Qudus, W.H. Chen, A new integrated collision risk assessment methodology for automated vehicles, *Accid. Anal. Prev.* 127 (2019) 61–79.
- [23] C. Ahlmann-Eltze, C. Yau, MixDir: scalable Bayesian clustering for high-dimensional categorical data, *IEEE 5th International Conference on Data Science and Advanced Analytics*, Turin, Italy, 1–4 October, 2018, 2018.
- [24] S. Wang, L. Zhixia Li, Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data, *Accid. Anal. Prev.* 129 (2019) 44–54.
- [25] A.M. Boggs, R. Arvin, A.J. Khattak, Exploring the who, what, when, where, and why of automated vehicle disengagements, *Accid. Anal. Prev.* 136 (2020).