

1 **Traffic Collisions Involving Autonomous Vehicles in California: Bayesian**
2 **Model Based Clustering**

3
4 **Subasish Das, Ph.D.**

5 (Corresponding author)

6 Associate Transportation Researcher, Texas A&M Transportation Institute

7 1111 RELIS Parkway, Room 4414, Bryan, TX 77807

8 Email: s-das@tti.tamu.edu

9 ORCID ID: 0000-0002-1671-2753

10
11 **Anandi Dutta, Ph.D.**

12 Computer Science and Engineering Dept., Ohio State University

13 2015 Neil Ave, Columbus, OH 43210

14 E-mail: dutta.34@osu.edu

15
16 **Ioannis Tsapakis, Ph.D.**

17 Texas A&M Transportation Institute

18 3500 NW Loop 410, Suite 315

19 San Antonio, TX 78229

20 phone: 210-321-1217; email: i-tsapakis@tti.tamu.edu

21
22
23
24
25
26 **TOTAL WORDS: 5,694 words**

27 4,694 words = text (including abstract and references)

28 1,000= 4 table
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

ABSTRACT

The emerging technology of autonomous 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 crashes 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 safety of AVs requires a multi-disciplinary approach which monitors every aspect of these vehicles. To promote safety, the California Department of Motor Vehicles has mandated that autonomous car crash reports be made public in recent years. This study collected all crash reports filed by different manufacturers that are testing autonomous vehicles in California (September 2014 to May 2019). The data provides important information on autonomous vehicles crash frequencies and associated contributing factors. This study provides an in-depth exploratory analysis of the critical variables. The research team 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 level. The authors anticipate that these results will provide a significant contribution to the area of AV and safety outcomes.

Keywords: autonomous vehicle, Bayesian model, traffic collision, traffic crash.

1 INTRODUCTION

2 An autonomous vehicle uses artificial intelligence (AI) and mechanics that can serve a human
3 driver in 1) controlling a vehicle (controlling its speed and its main functions of steering) and 2)
4 observing the surrounding environment (e.g., other vehicles/pedestrians, road markings, traffic
5 signals, etc.). Given that the performance of particular control functions (e.g., accelerating or
6 decelerating) will depend on inputs and signals which are received from the surrounding
7 environment (e.g., a traffic light turning red), the two functions are undoubtedly linked and
8 dependent on each other. The Society of Automotive Engineers (SAE) defines six levels of
9 autonomy which considers the degree to which the autonomous technology is capable of
10 providing support and assistance the driving tasks (1).

11 The California Department of Motor Vehicles (CA DMV) requires that trained human
12 drivers should remain behind the wheel while testing on public roads, regardless of the autonomy
13 levels of the vehicles to promote safety. In addition to human drivers, the California DMV
14 mandated that the recent years of autonomous car crash reports must be made public. The first
15 type of reporting is a brief list of all occurrences of AV disengagements (during failure or
16 difficult to control events, human driver will take control by putting the autonomous feature of
17 the car disengaged). The second type of report supplies a thorough summary of events in which a
18 collision and/or damage to property and injuries take place.

19 To ensure the safety requirement, there should be careful strategies at the present time by
20 enacting more rigid regulations. However, there is a need for performing rigorous analysis on the
21 available AV crash data. California based AV crash data provides crucial information on AV
22 crashes including key contributing factors. The intent of this paper is to carefully investigate the
23 AV crash data to shed further light on heterogeneity effects in roadway geometric features,
24 human interactions, and other attributes with respect to occurrences of AV crashes.

26 LITERATURE REVIEW

27 One of the major causes for disengagement in autonomous vehicles is the driver's lack of
28 attention because of the overreliance on vehicle automation and his/her use of the automation
29 inconsistent with the guidance and warnings from the manufacturer (2). In this regard, Trovato
30 (3) provided a structure that summarizes the fundamental behavior of a robot and automatically
31 computes intelligent maneuvers for collision-free maneuvering and control of an autonomous
32 vehicle. The capacity of autonomous vehicles to improve safety and riding experience is
33 normally regarded with apprehension. Shim et al. (4) developed a collision-avoidance system for
34 an autonomous vehicle application and concluded its effective performance collision avoidance
35 maneuvers. Based on the analysis of collision risk and driving behavior, Tak et al. (5) proposed
36 an Asymmetric Collision Risk (ACR)-based spacing policy. In comparison to other spacing
37 policies, the results showed pattern similarities in the ACR spacing policy with the human driver
38 with a smoother trajectory and less acceleration/deceleration actions. According to the
39 information provided by a laser-scanner sensor, Jiménez et al. (6) generated a collision-
40 avoidance system in which two actions could be taken in case of danger. The system produced
41 satisfactory results when implemented in a vehicle and tested with pedestrians and vehicles
42 circulating along a private test track. Moreover, Cao et al. (7) developed a comprehensive
43 architecture of an active collision-avoidance system for an autonomous vehicle to decipher
44 potential hazards on a straight or curved road. Under various situations, the simulation results
45 indicated the success of a host vehicle to make a collision avoidance maneuver without the
46 intervention of a human driver.

1 When lateral control was delegated to automation, Navarro et al. (8) analyzed the
2 unexpected obstacle avoidance maneuvers by conducting a simulation study. In comparison to
3 driving without automation, drivers returning to manual control from automatic steering were
4 found to be less effective at maneuvering around obstacles. Dixit et al. (9) observed a significant
5 correlation between the number of crashes, the traveled autonomous miles, and the vehicle's
6 reaction time to take control in the event of disengagement and found an average distribution of
7 0.83 seconds across different companies.

8 To encourage safety and transparency for customers, the California Department of Motor
9 Vehicles has ordered that accounts of crashes involving autonomous vehicles be drawn up and
10 rendered open to the public. Correspondingly, Favarò et al. (10) generated a detailed assessment
11 of the crash records submitted by different manufacturers studying autonomous vehicles in
12 California. The data provided significant information about the dynamics of autonomous vehicle
13 crashes linked to the most common kinds of collisions and effects, frequencies of crashes, and
14 other contributing variables. Additionally, Favarò et al. (11) analyzed the safety-critical role of
15 AV disengagement which required the timely and safe return of vehicle control to the human
16 driver. The study provided an inclusive outline of the fragmented data such trends of
17 disengagement reporting, average mileage driven before failure, associated frequencies, etc.
18 acquired from AV manufacturers testing on California public roads from 2014 to 2017.

19 Poland et al. (12) examined the interaction between the Society of Automobile
20 Engineers' (SAE) level 2 automated vehicle and the driver, including the vehicle damage,
21 limitations imposed by the vehicle on the driver using scene evidence, recorded data available
22 from the vehicle, and information from both drivers such as experience, phone records, computer
23 systems, and medical information. To improve automated response time, Roldan et al. (13) tested
24 a theory using two driving simulator studies that enabled participants to drive simulated in a
25 controlled environment using cooperative adaptive cruise control (CACC) vehicles that directly
26 transmit vehicle-to-vehicle data. The goal of the experiment was to evaluate the driver workload
27 while using CACC and adaptive cruise control (ACC) technologies and determine whether
28 CACC improves or lowers collision prevention when driver action is essential. Boggs et al. (14)
29 established a distinctive database from the California Department of Motor Vehicles (DMV) 66
30 manufacturer-reported Traffic Collision Reports (OL 316) that included responses to close-ended
31 collision issues and text mining narratives. The findings showed that most AV crashes occurred
32 in completely automated mode (65.2 percent), and the likelihood of AVs being hit was greater
33 than the effect before car takeovers and conventionally powered cars.

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

44 Xu et al. (17) used descriptive statistical analysis to examine the trends and features of
45 the connected and autonomous car (CAV) involved accidents. The findings indicated that the
46 primary factors adding to the severity stage of CAV related accidents were the CAV driving

mode, roadside parking, crash location, one-way road, and rear-end collision. Lodinger and DeLucia (18) compared time-to-collision (TTC) judgments between manual and automated driving to determine whether the automation only affected responses (i.e., braking) or also affected visual perception (i.e., TTC estimation). Yu et al. (19) investigated the impact on the safety and effectiveness of autonomous cars (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 automated cars in traffic jam scenarios.

As both contextual and circumferential factors should be regarded simultaneously, the evaluation of real-time threat for strategic and operational autonomous 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 methodology that incorporates a network-level crash estimate with a vehicle-based real-time threat estimate. Results showed that a well-calibrated classification of crash prediction provides a vital indication for better risk perception by autonomous vehicles.

The literature review reveals that the previous exploratory study conducted by Favaro et al. (10) analyzed only 26 AV crash data. This suggests a need to examine a comprehensive AV crash data to unearth the hidden trends and associated factors. This study aims to conduct an in-depth study with inclusion of 151 AV crashes in California.

METHODOLOGY

Bayesian Clustering

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

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

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

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

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

α and β are hyper-parameters which are defined externally and govern the sparsity of the model. Equation 1 defines that the size of the classes is governed by a Dirichlet (in the case of a simple LCM) or by a Dirichlet Process (in the case of a nonparametric LCM). First, the derivation for the simple LCM is described, then the way to extend it to the nonparametric case is presented. z is a vector which includes the latent class assignment for each individual. 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 question j . Equation 4 specifies that the response of an individual i that belongs to class k is a

draw from a Multinomial distribution according to the probability vector $U_{j,k}$. The joint distribution of the model can be defined as:

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

For this model, finding the maximum likelihood solution would lead to an EM algorithm. This algorithm is similar to the one described by Linzer and Lewis (28). However, a variational inference (VI) method is developed to properly propagate uncertainty through the model and to infer an appropriate number of latent classes. The idea of VI is to define a simplified probability model q and tune its parameters to approximate the original model p . The purpose of choosing q as the mean-field approximation of p , allows the user to write down the variational distribution:

$$q(\lambda, z, U) = q(\lambda)q(z)q(U), \quad (6)$$

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

where ω , ζ and ϕ are the free variational parameters, that are subsequently optimized. It is also defined that

$$\begin{aligned} q(\lambda; \omega) &= \text{Dirichlet}(\omega) \\ q(z_i = k; \zeta_i) &= \zeta_{i,k} \\ q(U_{j,k}; \phi_{j,k}) &= \text{Dirichlet}(\phi_{j,k}) \end{aligned} \quad (7)$$

The Kullback-Leibler (KL) divergence has been utilized to measure the approximation, which allows the user to maximize the evidence lower bound (ELBO). It is found that iterating between the following equations maximizes the ELBO and thus also minimizes the KL divergence.

DATA DESCRIPTION

Data Collection

The research team developed a database that provides descriptive and detailed reports of AV crashes in California during 2014-2019. The total number of reported crashes used in this study was 151. Figure 1 shows the cumulative number of collisions or crashes 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 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.

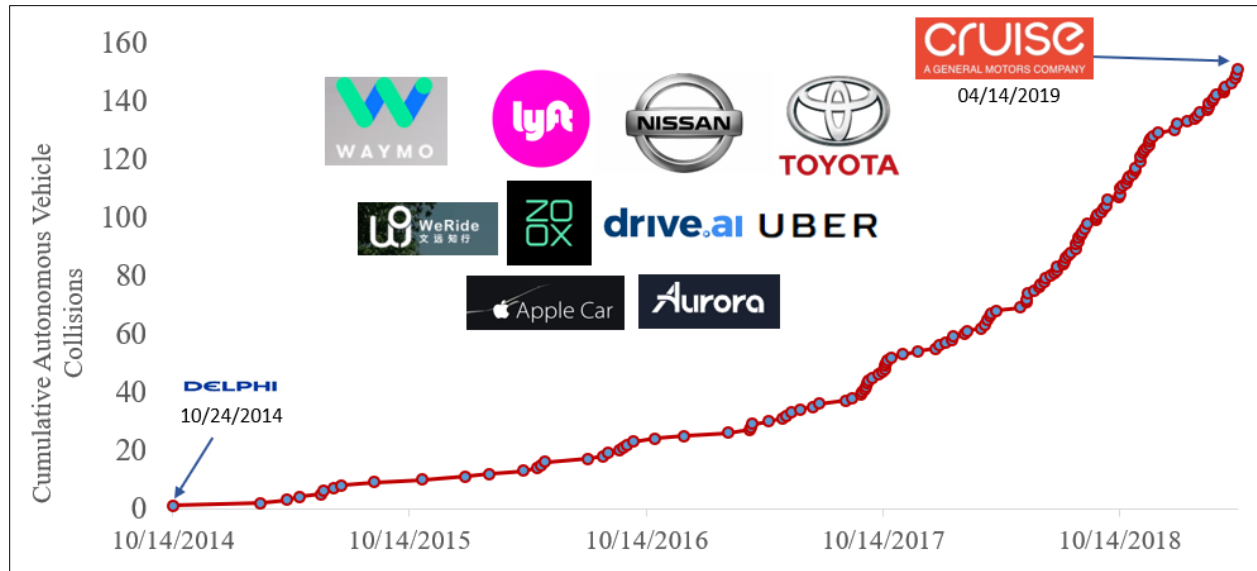


Figure 1 Cumulative AV collisions in California.

Exploratory Data Analysis

Table 1 shows lists the number of traffic collisions and the number of autonomous miles driven for twelve companies since their AV deployment date. Waymo had the greatest number of autonomous miles (352,545) with the second-highest number of traffic collisions (55). Cruise had the most traffic collisions (71), and it had the second greatest number of autonomous miles (131,676). Delphi, Nissan, Drive.ai, and WeRide all only had one traffic collision reported. In general, a greater number of autonomous miles was associated with a greater number of traffic collisions; however, it is important to note that the data for autonomous miles contained missing values for half of the companies listed.

Table 1 Number of Collisions and Autonomous Miles by the AV Companies












Company	Traffic Collisions	Autonomous Miles
	71	131,676
	55	352,545
	7	2,245
	4	--
UBER	3	--
	3	--
	2	--
	2	--
	1	1,820
	1	5,007
	1	6,572
	1	--

Figure 2 provides a visual representation of where AV crashes took place around Santa Clara, California. Each red dot represents one collision occurrence. As shown in the figure, a majority of the crashes were concentrated in the Mountain View area.

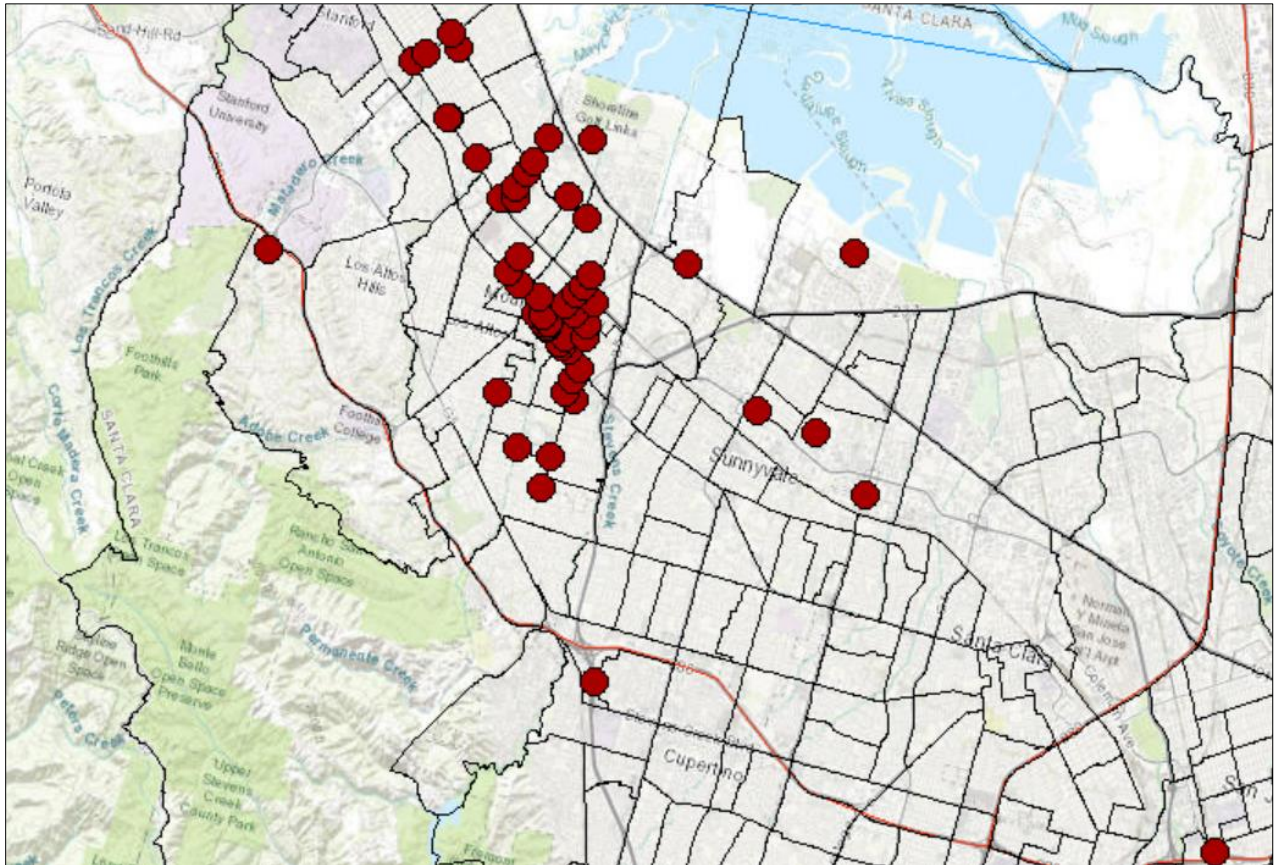


Figure 2 Location of AV crashes in Santa Clara.

Figure 3 provides a visual representation of where AV crashes took place around San Francisco, California. Each red dot represents one crash occurrence. As shown in the figure, the majority of the crashes were concentrated in the northeastern area.

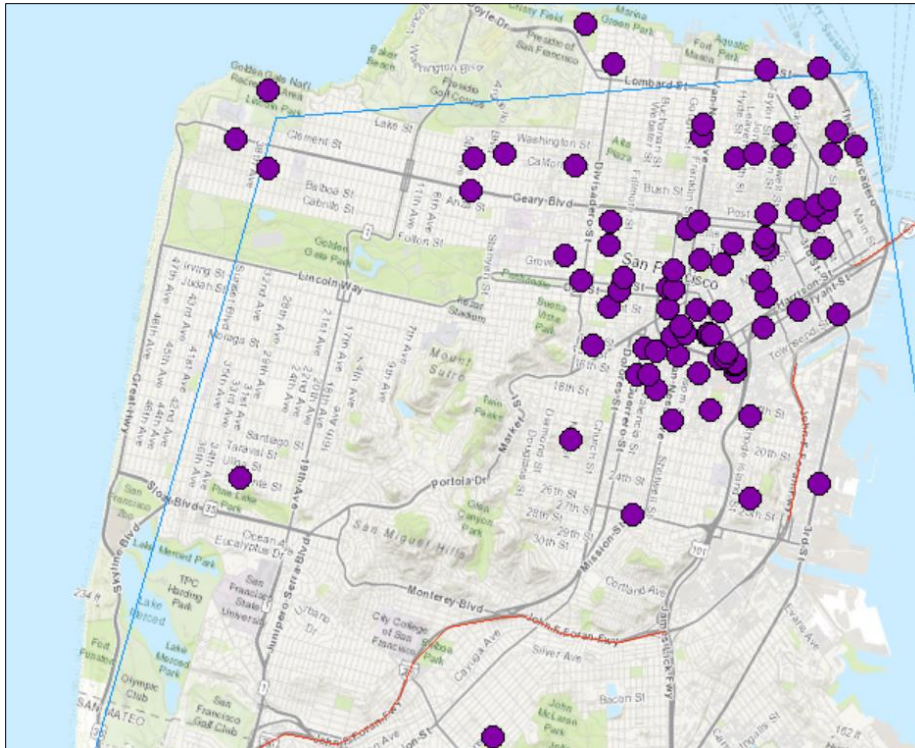


Figure 3 Location of AV crashes in San Francisco.

Table 2 lists the frequency of AV collisions for different times of day and days of the week. Times that had a higher frequency of crash occurrences are represented by a darker shade of red. The times and days that had the highest frequencies of crash occurrences are Thursdays from 6-12 PM (16), Fridays from 1-6 PM (14), and Wednesdays from 1-6 PM (11). One finding from this table is that the time period of 1-6 AM consistently has a very low frequency of crash occurrences for all days of the week.

Table 2 Number of Collisions by Time of the Day

Day of Week	1- 6 AM	6- 12 PM	1- 6 PM	7-12 AM
Saturday	2	3	7	2
Sunday	2	2	4	6
Monday	2	6	8	4
Tuesday	3	7	5	8
Wednesday	3	5	11	9
Thursday	2	16	5	6
Friday	0	5	14	4

Table 3 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), August 2018 (10), July 2018 (8), September 2018 (8), and October 2018 (8). 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.

Table 3 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								

Figure 4 shows nine bar plots to represent the distribution of AV collisions for different variables; 151 total crashes were studied. A majority of 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 2 collisions were reported as major vehicle damage. A vast majority of collisions (143 collisions) did not cause severe injuries to the driver. Of the crashes studied, a majority of them (89 collisions) involved a vehicle that was in autonomous driving mode at the time of the crash. Furthermore, a majority of the crashes (128 collisions) involved two vehicles, rather than a single vehicle or multi-vehicle collision. Most of the crashes occurred in daylight (68 collisions) and during clear weather (72 collisions). A majority of crashes

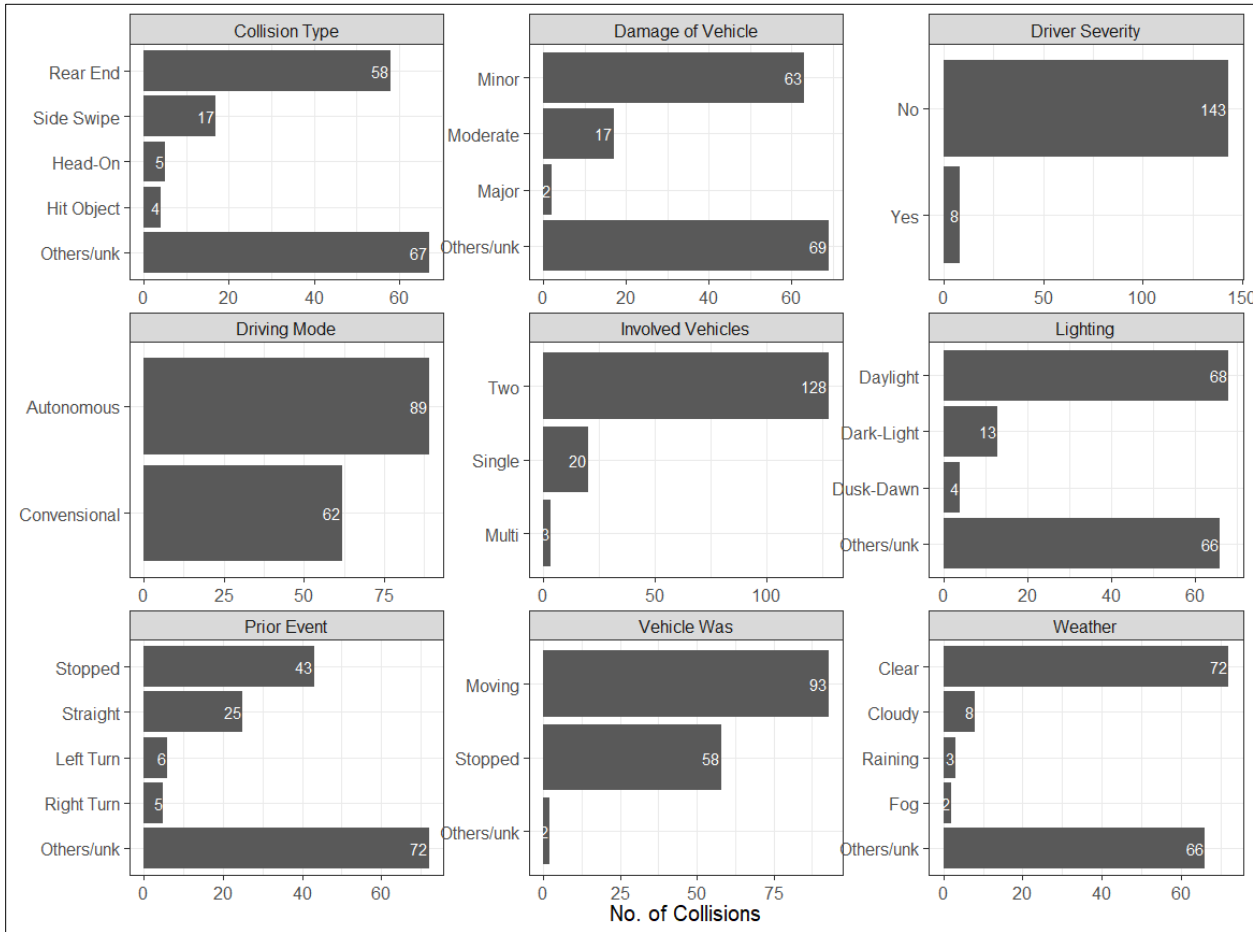


Figure 4 Bar plot showing the distribution of AV crashes by different variable categories.

occurred when the vehicle involved was moving (93 collisions), and the most common prior event to the crash was being at a stop (43 collisions).

RESULTS AND DISCUSSIONS

This study used *mixdir* (24), an open-source R package, to perform the analysis. The research team employed a hierarchical Dirichlet Process mixture of multinomial distributions. The package is a probabilistic latent class model (LCM), so it can be used to reduce the dimensionality of hierarchical data and cluster individuals into latent classes. Furthermore, it can be used to infer an appropriate number of latent classes. Figure 5 is a heat map that provides a visual representation of the six classes of participant response groups. The rows in this plot indicate each of the AV crash events. Based on the combinations of attributes among these crashes, 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 crashes to the latent classes. This method clusters the AV crash data and uncovers interesting latent structure. The values of Table 4 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; a majority of the crashes were associated with autonomous driving mode and no driver severity. This class indicates the highest percentage of crashes that occurred when the vehicles were moving and during the weekdays. Moreover, two-vehicle crashes 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.

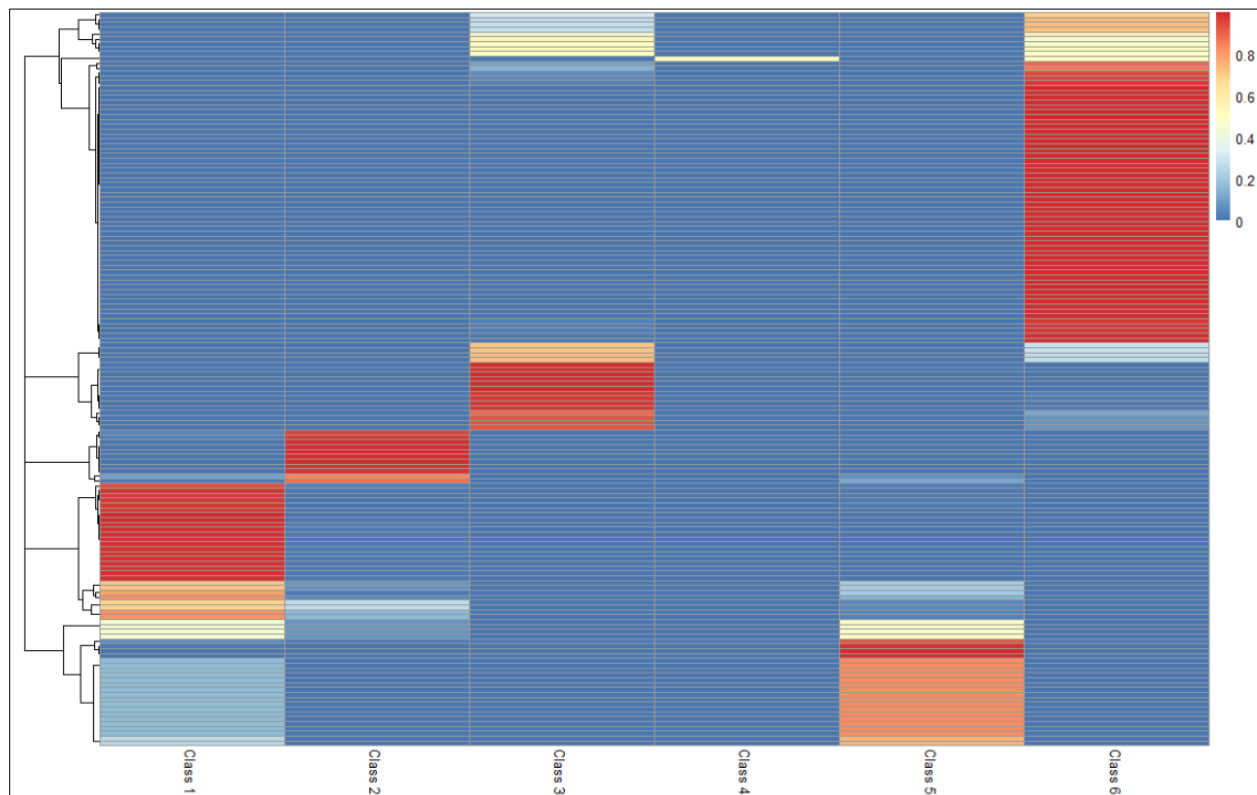


Figure 5 Heatmap showing six classes of participant response groups.

- Class 2 consists of 11 reports; these reports contained higher counts of conventional vehicles crashes compared to autonomous ones. A majority of the incidents were single-vehicle crashes with no severe injuries to the driver.
- Class 3 consists of 21 reports in which the majority of crashes were accompanied by severe driver injuries, and left and right turn and straight as their prior event. Furthermore, most of the crashes took place in dark lighting. The percentage of conventional vehicle crashes was also higher than the percentage of autonomous ones. Moreover, class 1 and 3 are the only classes in which the highest number of crashes were side swipe collisions and in which a majority of the crashes occurred on weekends.
- Class 4 is a limited cluster, containing only one report. The report indicated a multi-vehicle crash resulting in major damages.
- Class 5 shows the highest percentages of multi-vehicle crashes that have occurred while moving. The most frequent weather condition for this cluster was ‘unknown,’ and the percentage of the driver 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 crashes, 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) crashes. Furthermore, head on and rear end were reported as the first and the second highest percentages among all collision types, respectively.

Table 4 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)
Driving Mode							
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
Driver Severity							
No	143	19.58	7.69	11.19	0.7	16.08	44.76
Yes	8	0	0	62.5	0	37.5	0
Prior Event							
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
Vehicle Was							
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
Involved Vehicles							
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
Damage Vehicle							
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
Day of Week							
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
Weather							
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
Lighting Condition							
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
Collision Type							
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

This study also gathered police crash 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, the research team developed trigrams from two corpora that are developed based on the vehicle automation mode during the crash event. The *odds* of word w in group i 's usage can be defined 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 frequencies from documents of class i in topic k . The odds ratio between the two groups can be expressed as $\theta_{kw}^{(M-P)} = O_{kw}^{(M)} / O_{kw}^{(P)}$. This is generally presented for single words in isolation or as a metric for ranking words. Figure 6 shows the log odds ratios of the top trigrams, the continuous sequence of three words from a document, from the crash reports for crashes in which the AV was in either autonomous or conventional mode prior to crash occurrence. In the present data, a crash report associated with autonomous as the prior mode is 1.45 times more likely to use a variant of 'av made contact' than a crash report associated with conventional as prior mode. It is important to note that the narrative texts from the crash reports are not detailed enough to separate the report types by prior driving mode in many cases. As the autonomous drivers 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 crash. The current findings call for more detailed crash narrative documentation in crashes involving AVs.

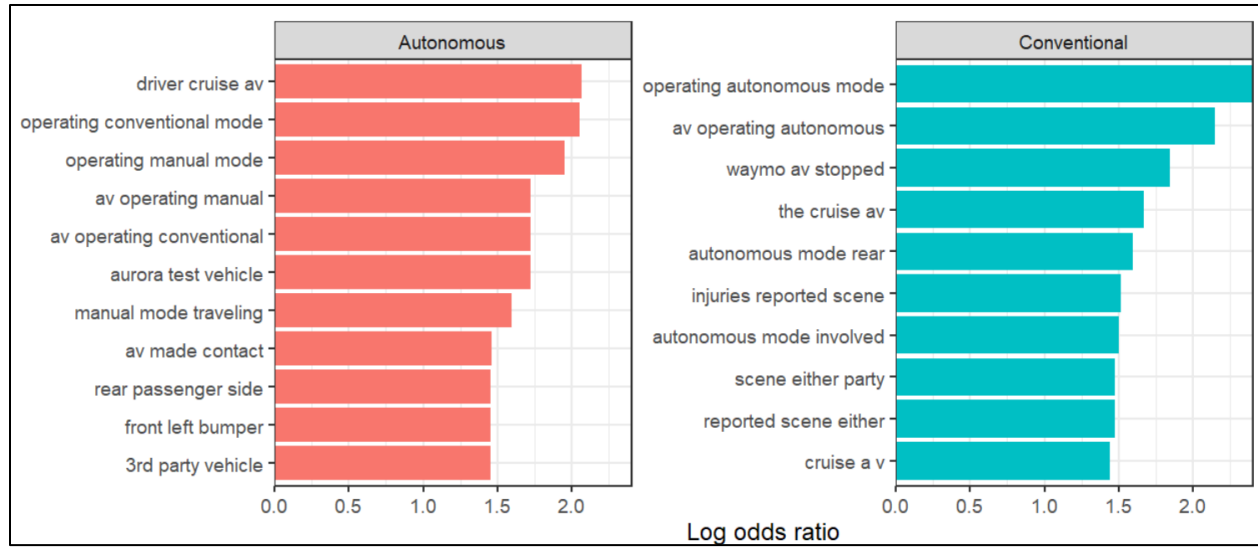


Figure 6 Log odds ratio of the trigrams generated from the crash reports.

CONCLUSIONS

AVs are expanding their market quickly, and with this expansion, some safety-related concerns raise significantly. Drivers in fully AVs can be involved in non-driving tasks. However, if the automatic system fails, or becomes limited, the drivers 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 crash history and the contributing factors in AV crashes (9, 11). In this study, a comprehensive analysis was conducted using the data including the crash reports filled by various manufacturers from September 2014 to May 2019. The research team 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 six classes of collision patterns based on different variables and crash traits. The variables included collision type, damage to the vehicle, driver injury severity, lighting conditions, the number of vehicles involved, weather conditions, the event prior to the crash, 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 crash occurrences when the vehicle's prior Districondition was stopped.

The research team also investigated crash narrative texts from police crash reports to determine whether they can be used to accurately identify the mode of the AVs prior to the crash. 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 crashes are complex and distinctive in nature. There is a need for more advanced and robust crash narrative reporting in order to better investigate the association of automation levels with collision likelihood.

This study is unique in that it highlights the complexity and challenges of identifying key risk factors associated with AV crashes. 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.

ACKNOWLEDGMENT

The authors are extremely grateful to Sirajum Munira, Xiaoqiang Kong, Apoorba Bibeka, Sadia Najneen and Kartikeya Jha for preparing the datasets. The authors appreciate the assistance provided by the students on this manuscript preparation: Bitu Maraghehpour, Magdalena Theel and Ly-Na Tran.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm the contribution to the paper as follows: study conception and design: Subasish Das; data collection: Subasish Das; analysis and interpretation of results: Subasish Das; draft manuscript preparation: Subasish Das, Anandi Dutta, and Ioannis Tsapakis. All authors reviewed the results and approved the final version of the manuscript.

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, p. 63p.
3. Trovato, K. I. Collision-Free Maneuvering and Control of an Autonomous Vehicle. *Advanced vehicles and infrastructure systems*, 1997, pp. 189–208.
4. Shim, T., G. Adireddy, and H. Yuan. Autonomous 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: Journal of Automobile Engineering*, Vol. 226, No. 6, 2012, p. pp 767-778.
5. Tak, S., H. Yeo, and Chalmers University of Technology. SAFER Vehicle and Traffic Safety Centre. Asymmetric Collision Risk Spacing Policy for Longitudinal Control of Autonomous Driving Vehicle, 2015.
6. Jiménez, F., J. E. Naranjo, and Ó. Gómez. Autonomous Collision Avoidance System Based on Accurate Knowledge of the Vehicle Surroundings. *IET Intelligent Transport Systems*, Vol. 9, No. 1, 2015, p. pp 105-117.
7. Cao, H., X. Song, Z. Huang, and L. Pan. Simulation Research on Emergency Path Planning of an Active Collision Avoidance System Combined with Longitudinal Control for an Autonomous Vehicle. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, Vol. 230, No. 12, 2016, p. pp 1624-1653.
8. Navarro, J., M. François, and F. Mars. Obstacle Avoidance under Automated Steering: Impact on Driving and Gaze Behaviours. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 43, 2016, p. pp 315-324.
9. Dixit, V. V., S. Chand, and D. J. Nair. Autonomous Vehicles: Disengagements, Accidents and Reaction Times. *PLOS ONE*, Vol. 11, No. 12, 2016, p. e0168054. <https://doi.org/10.1371/journal.pone.0168054>.
10. Favaro, F. M., N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju. Examining Accident Reports Involving Autonomous Vehicles in California. *PLOS ONE*, Vol. 12, No. 9, 2017, p. e0184952. <https://doi.org/10.1371/journal.pone.0184952>.

11. Favarò, F., S. Eurich, and N. Nader. Autonomous Vehicles' Disengagements: Trends, Triggers, and Regulatory Limitations. *Accident Analysis & Prevention*, Vol. 110, 2018, pp. 136–148. <https://doi.org/10.1016/j.aap.2017.11.001>.
12. Poland, K., M. P. McKay, D. Bruce, and E. Becic. Fatal Crash Between a Car Operating with Automated Control Systems and a Tractor-Semitrailer Truck. *Traffic Injury Prevention*, Vol. 19, No. sup2, 2018, p. pp S153-S156.
13. Roldan, S. M., V. W. Inman, S. A. Balk, and B. H. Philips. Semi-Autonomous Connected Vehicle Safety Systems and Collision Avoidance: Findings from Two Simulated Cooperative Adaptive Cruise Control Studies. *ITE Journal*, Vol. 88, No. 6, 2018, p. pp 30-35.
14. Boggs, A., A. J. Khattak, B. Wali, and Transportation Research Board. Analyzing Automated Vehicle Crashes in California: Application of a Bayesian Binary Logit Model, 2019.
15. Khattak, Z. H., M. D. Fontaine, B. L. Smith, and Transportation Research Board. An Exploratory Investigation of Disengagements and Crashes in Autonomous Vehicles, 2019.
16. Lee, D., S. Tak, S. Choi, and H. Yeo. Development of Risk Predictive Collision Avoidance System and Its Impact on Traffic and Vehicular Safety. *Transportation Research Record: Journal of the Transportation Research Board*, 2019.
17. Xu, C., Z. Ding, C. Wang, and Transportation Research Board. Investigating the Characteristics of Connected and Autonomous Vehicle Involved Crashes, 2019.
18. Lodinger, N. R., and P. R. DeLucia. Does Automated Driving Affect Time-to-Collision Judgments? *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 64, 2019, p. pp 25-37.
19. Yu, H., S. Tak, M. Park, and H. Yeo. Impact of Autonomous-Vehicle-Only Lanes in Mixed Traffic Conditions. *Transportation Research Record: Journal of the Transportation Research Board*, 2019.
20. Rahman, M. S., M. Abdel-Aty, J. Lee, and M. H. Rahman. Safety Benefits of Arterials' Crash Risk under Connected and Automated Vehicles. *Transportation Research Part C: Emerging Technologies*, Vol. 100, 2019, p. pp 354-371.
21. Rao, S. J., T. Seitz, V. R. R. Lanka, and G. Forkenbrock. Analysis and Mathematical Modeling of Car-Following Behavior of Automated Vehicles for Safety Evaluation. Presented at the SAE Technical Paper, 2019.
22. Katrakazas, C., M. Quddus, and W. H. Chen. A New Integrated Collision Risk Assessment Methodology for Autonomous Vehicles. *Accident Analysis & Prevention*, Vol. 127, 2019, p. pp 61-79.
23. Ahlmann-Eltze, C., and C. Yau. MixDir: Scalable Bayesian Clustering for High-Dimensional Categorical Data. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, Turin, Italy, 1-4 October 2018.