https://doi.org/10.1016/j.amar.2021.100165

Stop or go decisions at the onset of yellow light in a connected environment: A hybrid approach of decision tree and panel mixed logit model

Yasir Ali [a, b], Md. Mazharul Haque [a,*], Zuduo Zheng [c], Michiel C. J. Bliemer [b],

[a]*School of Civil & Environment Engineering, Faculty of Engineering, Queensland University of Technology, Australia.*

[b]*Institute of Transport and Logistics Studies, University of Sydney Business School, The University of Sydney, Australia.*

[c]*School of Civil Engineering, The University of Queensland, Australia.*

** corresponding author*

Abstract

Driver decisions at the onset of yellow traffic lights are often critical, as inaccurate decisions may result in traffic conflicts and collisions. A future connected environment where vehicles can communicate with traffic lights is expected to minimize the uncertainty associated with a driver's decision-making at signalized intersections by providing advance information related to traffic light changes. The effectiveness of such a connected environment, however, remains unexplored due to the paucity of relevant data. This study examines driver decisions at the onset of yellow traffic lights when they are assisted with advance information about traffic light changes. Seventy-eight participants with diverse backgrounds performed driving experiments on an urban route with a signalized intersection simulated in the CARRS-Q Advanced Driving Simulator. The experiment consisted of two randomized driving conditions: baseline (without advance information aids) and connected environment (with advance information aids). Contrary to the existing literature, this study employs a hybrid approach, leveraging the combined benefits of data mining to identify *a priori* relationships and a panel mixed logit model (more specifically, correlated grouped random parameters logit with heterogeneity-in-means approach) to account for unobserved heterogeneity as well as the correlation among random parameters. Our analysis shows that drivers in the connected environment are less likely to proceed through intersections at the onset of yellow light compared to the baseline condition. However, at the individual driver-level, the connected environment's impact on driver decisions is mixed. Female drivers have been found to have a higher propensity for yellow light running in the connected environment than that of male drivers. Overall, the connected environment assists drivers in making safer decisions at the onset of yellow light.

Keywords: Connected environment; road safety; yellow light; driving behavior; correlated grouped random parameters model; machine learning.

1. Introduction

Advancements in communication technologies like connected vehicles have shown promise to address massive transport issues related to traffic congestion, road safety, and greenhouse gas emissions. Although recent research efforts have emphasized the effectiveness of a connected environment in minimizing crash risk and improving traffic flow conditions on motorways, it is imperative to assess the efficacy of a connected environment in more complex interactions such as in case of urban road networks. This study focuses on the impact of a connected environment in an urban context with signalized intersections.

Intersections are susceptible to conflicting movements of various road users from different directions. Specifically, traversing through a signalized intersection is a complex driving maneuver that requires significant cognitive capability and visual-manual attention (Lu et al., 2015, Caird et al., 2007). Due to this complexity, intersections are associated with high crash risk (Choudhary and Velaga, 2019). For instance, during 2010, the U.S. National Highway Traffic Safety Administration reported about 35% of crashes occurred at intersections (Choi, 2010). In 2018, 179 and 48 drivers were killed in intersection-related crashes in New South Wales and Queensland, Australia, respectively (DTMR, 2019, TfNSW, 2019).

Traversing through a signalized intersection when a traffic light changes from green to yellow is an attentive task, requiring a driver to quickly decide whether to stop or cross the intersection and thus regarded as a critical interval (Elmitiny et al., 2010, Papaioannou, 2007). Drivers may end up in the dilemma zone, where one cannot safely stop before the stop line nor proceed through the intersection during a yellow interval, and then often make a decision based on their driving speeds, distance to the stop line, and their position in traffic stream (Elmitiny et al., 2010). Elmitiny et al. (2010) observed two frequent behaviors: (i) aggressive driving, where one is far away from the stop line but decides to proceed through an intersection, tending to run the red light; and (ii) conservative driving, where one is close to the stop line and could safely pass through an intersection, but decides to stop. Due to driver heterogeneity, conflicting decisions may arise from a following vehicle, leading to an increased probability of rear-end or angle crashes at intersections. It is reported that about 139,000 and 846 people were respectively injured and killed in red light running crashes in the U.S. during 2018 (IIHS, 2020).

An inappropriate and risky decision of crossing an intersection at the onset of yellow light often leads to a red light violation as well as conflicts with leading vehicles that have decided to stop at the intersection and conflicts with vehicles from other directions of travel (Elmitiny et al., 2010). Along this line, Baguley (1988) classified driving behavior at intersections in three groups: (a) drivers who are likely to clear the intersection before the red light but are either hindered by a slow-moving leader or their own indecisiveness (Retting et al., 2002); (b) uncertain drivers in the dilemma zone; and (c) drivers deliberately running the red light knowing that they could cross without any safety hazards (NHTSA, 2006). This study attempts to understand driver decisions collectively from these three groups at the onset of yellow light and their contributing factors.

To study the cause-and-effect relationship of driver decisions with its contributory factors, most of the previous studies solely apply traditional statistical models (mostly commonly binary logistic models). These models require an analyst to specify main effects and potential interactions among them based on their prior knowledge and do not account for unobserved heterogeneity. This problem further aggravates while specifying higher-order interactions typically unknown to an analyst and when there exist multiple (or repeated) driving conditions that add an additional layer of potential correlation between repeated observations (Mannering et al., 2016). Ignoring these issues during the model development process may lead to model misspecification issues (see Mannering et al. (2016) for a detailed discussion). To this end, Mannering et al. (2020) pointed out that "*there is a clear need in the safety field to ground* *intrinsically predictive models within causal frameworks, while also taking insights from intrinsically predictive models (especially from big data) to improve upon causal structures through insights from associations involving variables not typically available in traditional safety data. One promising direction for future research would be a hybrid modeling approach of data-driven and statistical methods (with strong consideration to causal elements)*". Following the recommendation of Mannering et al. (2020), there is a clear need to combine two approaches (data mining for obtaining prior knowledge about underlying relationships and advanced econometric modeling for capturing unobserved heterogeneity) to better understand driver decisions in a connected environment (more details are provided in Section 3).

The objective of this study is to investigate the impact of a connected environment on driver decisions at the onset of yellow light at signalized intersections. We address the following research questions: (1) can a connected environment reduce (or even eradicate) yellow light running?; (2) do drivers use advance information aids provided by a connected environment in a conservative manner to stop before the stop line?; (3) do drivers utilize such information in a counterproductive manner to safely proceed through an intersection?; and (4) does a connected environment result in a monotonous effect on driver decisions or there is a differential impact based on driver demographics? To answer these research questions, we employ a hybrid approach of data mining and advanced econometric modeling using real trajectory data collected from the advanced driving simulator experiment designed to mimic driving conditions in a connected environment.

To this end, the rest of the paper is structured as follows: Section 2 reviews the relevant literature. Section 3 explains the experimental plan, including the driving simulator, scenario design, participant details, data collection procedure, data processing, and the hybrid modeling approach adopted in this study. Modeling results are presented in Section 4, and Section 5 discusses the impact of the connected environment on driver decisions. Finally, Section 6 concludes the study and provides an outlook for future research.

2. Previous work on driver behavior at signalized intersections in traditional and connected environments

A synthesis of the literature reveals abundant research studies related to driving behavior at signalized intersections. These studies can be classified into two streams: studies related to driver decisions at signalized intersections in a traditional environment and identifying the factors affecting their decisions (Lu et al., 2015, Elmitiny et al., 2010, Caird et al., 2007, Papaioannou, 2007, Retting et al., 2002, Porter and England, 2000, Newton et al., 1997, Baguley, 1988, Mahalel and Prashker, 1987, Sheffi and Mahmassani, 1981), and studies related to distracted driver decisions at signalized intersections (Choudhary and Velaga, 2019, Eluru and Yasmin, 2016, Haque et al., 2016a, Xiong et al., 2016). For instance, using field data, Elmitiny et al. (2010) developed a binary logit model for estimating the probability of stopping at or crossing the stop line as a function of approaching speed, distance to the stop line, driver demographics such as gender, age group, and the presence or absence of a dilemma zone. Along similar lines, younger and older driver decisions were predicted using a binary logistic model as a function of the time to stop line using a moderate-fidelity simulator (Caird et al., 2007). In another driving simulator study, a new traffic light change anticipation system was tested and compared with a regular traffic light system and found that the new system reduced red light running violations compared to the baseline system. On the other hand, Haque et al. (2016a) investigated how mobile phone use affects driver stop/go decisions at signalized intersections using a driving simulator. This study found that running a yellow light while distracted is a function of driver demographic and speed at the onset of yellow light. More specifically, distracted young and middle-aged drivers showed a lower probability of yellow right running, reflecting risk compensation behavior. Similarly, in another driving simulator-based study, it was found that time to the stop line, maneuver type, and distraction caused by a mobile phone and a music player had a significant impact on the probability of crossing the intersection at the onset of yellow light (Choudhary and Velaga, 2019). Although these, as well as other relevant, studies have substantiated the significance of research related to driver decisions at signalized intersections during undistracted and distracted driving conditions in a traditional environment, it is yet unclear what the impact is of advance information provided by a connected environment on driver decisions at signalized intersections. This research gap motivates the present study.

Using vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications, a connected environment provides event-based as well as advance information aids, which will assist in reducing the uncertainty associated with decision-making and thereby alleviating (if not completely suppressing) traffic safety issues (Ali, 2020). A thorough literature review suggests that most of the existing studies related to a connected environment are based on numerical simulations and focus on analyzing macroscopic (or network-wide) benefits of a connected environment (Njobelo et al., 2018, Xiang et al., 2016, Sam et al., 2015, Lee and Park, 2012, Chang et al., 2009). For example, Xiang et al. (2016) reported the effectiveness of auditory warning messages on brake response time to a red-light running vehicle provided by a connected environment and found reduced collision rates at intersections when warning messages were available. Similarly, another study found that an advanced stop assist system in a connected environment lowered hard braking events by about 50% at signalized intersections (Sam et al., 2015). Although these studies provide evidence of the potential benefits of a connected environment using network simulations, an important component—the human factor—is not accounted for in these studies, which is vital for the success of a connected environment (Sharma et al., 2017). Realizing this limitation of microsimulations, recent simulator-based studies have shown a positive impact of real-time driving aids in a connected environment in improving safety during car-following (Sharma et al., 2020b, Sharma et al., 2019) and lane-changing (Ali et al., 2020a, Ali et al., 2020b, Ali et al., 2020c, Ali et al., 2020e, Ali et al., 2019a, Ali et al., 2019b, Ali et al., 2018)—note that these studies focused on motorways. However, it is still unclear how advance information in a connected environment affects driver decisions in an urban road context with signalized intersections.

3. Methodology

3.1 Design of experiment

Given the novelty of a connected environment and paucity of relevant data, this study designed an experiment to collect high-quality vehicle trajectory data. As data related to driver decisions at the onset of yellow light at a signalized intersection can be difficult and unsafe to obtain from field experiments, the Centre for Accident Research and Road Safety-Queensland (CARRS-Q) Advanced Driving Simulator (shown in Figure 1(a)) was utilized to provide a controlled driving environment and flexibility of collecting data without safety concerns. Participants were asked to drive in a city environment in two randomized driving conditions: baseline driving (without advance information aids; same as the traditional driving environment) and connected environment (with advance information about traffic light changes). The baseline driving condition is considered the '*default*' driving condition to which the driving performance is compared.

3.1.1 *Advanced Driving Simulator*

The CARRS-Q Advanced Driving Simulator (Figure 1(a)) consists of a Holden Commodore car with fully functioning controls, fitted with three projectors displaying a 180° field of view. The simulator is also attached to a flexible rotating base that can provide six degrees-of-freedom, mimicking real driving features like acceleration, deceleration, braking, cornering, and road surface. In addition, the simulator car produces simulated engine noises, vehicle-road interaction noises, and sounds of other traffic interactions. The simulator uses SCANeR™ studio software that connects eight computers for controlling the simulator car dynamics and virtual environment, and records basic operational variables (speeds, accelerations, positions, etc.) at every 0.05 s.

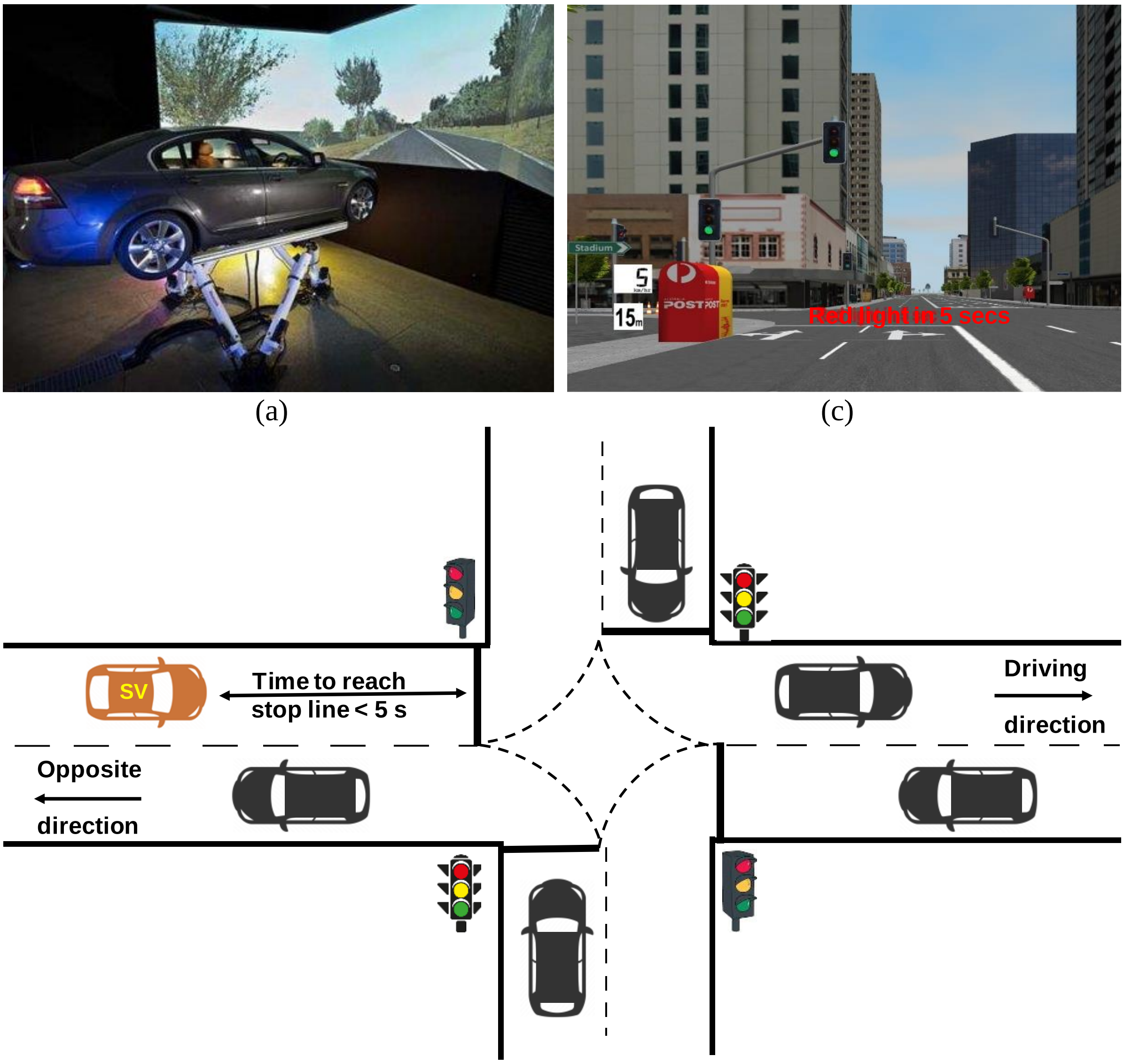


Fig. 1. Experiment design: (a) the Advanced Driving Simulator; (b) schematic of the designed driver-traffic signal interaction; and (c) a snapshot of the advance information displayed on the windscreen in the connected environment.

3.1.2 Participants

To ensure the diversity and representativeness of participants recruited for this study, we advertized our experiment at various public places and social media platforms. In total, 78 participants were recruited for this study, and their descriptive statistics are presented in Table 1. The mean age of the participants was 30.8 years (standard deviation [SD] 11.70 years), with 64% of them being male. The mean ages for male and female participants were respectively 34.1 (SD 12.6) years and 24.9 (SD 6.7) years. About 80% of the participants possessed an open driving licence, and their average driving experience was 12.2 (SD 11.5) years. About 10% of the participants self-reported that they were involved in a traffic crash in the past one year. About 58% of the participants reported that they had heard about connected vehicles previously. To compensate for their time of volunteer participation in the experiment, each participant received AUD 75 after completing the experiment.

Table 1 Descriptive statistics of the participants

Driver characteristics	**Mean**	**SD**	**Count**	**Percentage**
Driver's age (years)	30.8	11.7	—	—
Young drivers	22.11	2.44	38	48.72
Middle-aged drivers	35.34	3.36	32	41.03
Older drivers	58	4.08	8	10.26
Gender				
Male	—	—	50	64.1
Female	—	—	28	35.9
Education				
Primary	—	—	2	2.5
Junior (Grade 10)	—	—	1	1.3
Senior (Grade 12)	—	—	18	23.1
TAFE or Apprenticeship	—	—	9	11.5
University	—	—	48	61.6
Licence type				
Open	—	—	62	79.5
Provisional	—	—	16	20.5
Years of driving	12.2	11.5	-	-
Kilometers driven in a typical year				
0-5,000 km	—	—	10	12.8
5,001-10,000 km	—	—	19	24.4
10,000-15,000 km	—	—	15	19.2
15,001-20,000 km	—	—	18	23.1
20,001-25,000 km	—	—	6	7.7
> 25,000 km	—	—	10	12.8
Crash involvement in last one year				
Involved	—	—	8	10.3
Not involved	—	—	70	89.7
Frequency of driving per week				
Less than 2 times	—	—	5	6.4
2-4 times	—	—	28	35.9
5-6 times	—	—	16	20.5
7-8 times	—	—	7	9.0
More than 8 times	—	—	22	28.2
Prior information about Connected Vehicles				
Yes	—	—	33	42.3
No	—	—	45	57.7

3.1.3 Design of traffic signals

To satisfy the study needs, the Brisbane Central Business District (CBD) area and its surrounding environment were simulated in the driving simulator with high accuracy with traffic signs and road marking designed according to Australian road standards. The posted speed limit in the city was 40 km/h. The interaction with a traffic signal was judiciously placed on two intersections along a city route. Prior to interacting with traffic signals, drivers drove in the simulated city environment to familiarize themselves with city driving. When approaching a signalized intersection, a driver was required to respond to the change in the traffic light turning from green to yellow. In the experiment, the driver interacted with one of the two traffic signals in each drive while the other traffic signal was green when drivers approached the intersection. The selection of the intersection for yellow light interactions in a drive was randomized among the participants.

The driver-traffic signal interaction event was scripted in such a way that the traffic light turned from green to yellow when the driver was 5 s away from the stop line (see Figure 1(b)) based on the speed and distance to the stop line of the subject vehicle. Following the guidelines of the Department of Transport and Mains Roads, Queensland, the yellow interval between the red and the green light was set as 3 s, which means that participants had 2 s to read and interpret the message. Although previous studies have used this time period as an explanatory variable in the model (Choudhary and Velaga, 2019, Haque et al., 2016a) due to variability in the design of traffic lights, in our study, we use a fixed time period of 5 s to minimize confounding factors as otherwise, it would be difficult to determine whether the change in driving behavior is caused by different time periods or due to the connected environment. We also avoided having lead vehicles or ambient traffic near the two intersections to further avoid confounding the data. The simulated environment is the same for both the baseline and connected environment scenario with the exception of advance traffic light information shown in the simulated connected environment.

3.1.4 Design of the connected environment

Using simulated vehicle-to-infrastructure (V2I) communications between the traffic light and the subject vehicle, advance information was disseminated to the participants in the connected environment driving condition. For the design of the advance information, a thorough literature review was conducted, and designs of major car manufacturers were reviewed to determine how information is disseminated to drivers. By utilizing this knowledge, the information in the driving simulator was provided in two forms: visually (a text message) and auditory (a beep sound). The visual information was displayed on the windscreen resembling the heads-up display equipped in some of the recent vehicle models. Figure 1(c) illustrates an example of advance information showing the message “Red light in 5 secs” when the participant was 5 s away from the stop line.

Prior to the two actual experiment drives, participants performed a familiarization drive to get acquainted with the driving environment, simulator car, and designed interactions. Once they felt confident about their driving, they were allowed to participate in the actual experiment. Several strategies were implemented to minimize learning effects (and resulting bias) caused by repeated driving. First, the order of the two scenarios (baseline and connected environment) was randomized. Second, the intersection where advance information was received was also randomized in each drive. Third, the surrounding environment (including cars and buildings) was changed for each drive while keeping the signalized intersections identical. Fourth, although the scope of this paper is limited to investigating driver-traffic signal interaction in the city, each drive consisted of several other tasks such as car-following, lane-changing, and interactions with a pedestrian crossing, which are presented elsewhere (Ali et al., 2020c). Each of the two drives took on average 10-12 mins, and the entire experiment finished in about 50 mins.

3.1.5 Participant experiment protocol

At the CARRS-Q Advanced Driving Simulator facility, participants were briefed about the driving simulators and the objective of the experiment, including a detailed explanation about advance traffic light information in the simulated connected environment using schematics presented in Figure 1. Participants were instructed to obey the posted speed limit and drive to the speed limit as close as possible. Before starting the two experiment drives, participants were asked to complete a pre-driving questionnaire, including questions related to demographics, driving history, and driving behavior, and to perform a familiarization drive consisting of interactions with traffic light changes. Participants were tested for motion sickness using the standard instrument of motion sickness assessment adapted from Brooks et al. (2010), and workloads were assessed after each drive using the NASA TLX questionnaire. After completing the experiment, the participants received their fixed monetary reward.

3.2 Data collection

Driver decisions are extracted from the driving simulator data as a binary dependant variable for econometric modeling purposes, where 1 means that the participant proceeded through the intersection at the onset of yellow light, while 0 means that the participant stopped before the stop line. Explanatory variables are classified into traffic operational variables, driver demographics, and driving conditions. Traffic operational variables include driving speed, distance to the stop line at the onset of yellow light, and acceleration noise (or variation) prior to the onset of yellow light. Driver demographics contain age, gender, driving experience, licence type, and education. The driving condition variable has two categories, baseline and connected environment.

Table 2. Summary statistics of operational and response data for each driving scenario

Variable	Description of variables	Mean (SD)		Count (Percentage)	
		Baseline	CE	Baseline	CE
Driving condition					
Baseline	Driving without information aids (reference)	—	—	78	100
CE	Driving with information aids (dummy)	—	—	78	100
Traffic operational variables					
Speed	The speed of drivers at the onset of a yellow light (m/s)	10.16 (1.41)	9.65 (1.07)	—	—
Acceleration noise	The standard deviation of acceleration/deceleration of a driver prior to the onset of a yellow light (m/s^2)	0.64 (0.25)	0.25 (0.15)	—	—
Distance	The distance from the stop line at the onset of a yellow light (m)	30.23 (4.15)	50.70 (5.84)	—	—
Response variable					
Decisions	Driver decisions to proceed through a yellow light	—	—	51 (65.38)	26 (33.33)

CE: connected environment

Seventy-eight participants performed two drives, resulting in 156 decisions at the onset of yellow interval at signalized intersections. Each driver encountered two interactions with a traffic signal in repeated driving, thereby forming a panel dataset. Summary statistics of the explanatory and response variables are presented in Table 2. Among the 156 encounters with yellow lights, 51 drivers decided to proceed through the intersection in baseline conditions, while 26 drivers did so in the connected environment (this difference is also statistically significant, Fisher's exact test: p-value < 0.001). The mean driving speeds during the baseline and connected environment driving conditions are respectively 10.16 m/s (SD 1.41 m/s) and 9.65 m/s (SD 1.07 m/s). Acceleration noise—measured as the standard deviation of acceleration during the roadway segment prior to the onset of yellow light—is considered as an indicator of reckless driving (Ali et al., 2020e), and its values for the baseline and connected environment driving conditions are 0.64 m/s^2 (SD 0.25 m/s^2) and 0.25 m/s^2 (SD 0.15 m/s^2), respectively.

3.3 Data analysis techniques

Let y_{ij} be the indicator variable for the decision of driver i in scenario $j \in \{\text{baseline, connected}\}$, which equals 1 if the driver has proceeded through the intersection at a traffic signal and is 0 otherwise. Let $\mathbf{x}_{ij}$ and $\mathbf{z}_i$ be column vectors of corresponding driver/scenario-specific values of the traffic operational variables and driver-specific values of the sociodemographic variables, respectively. Further, let $\mathbf{w}_{ij}$ be a column vector of relevant interaction terms between traffic operational variables and sociodemographic variables obtained via decision tree analysis (see Section 4.1). The systematic utility V_{ij} that driver i attaches to proceeding through the intersection in scenario j, relative to stopping, is assumed to be described by a linear-additive relationship of operational and sociodemographic variables,

$$V_{ij} = \alpha + \boldsymbol{\beta}_i'\mathbf{x}_{ij} + \boldsymbol{\gamma}'\mathbf{z}_i + \boldsymbol{\delta}'\mathbf{w}_{ij}, \tag{1}$$

where, α is a constant, $\boldsymbol{\gamma}$ is a column vector of coefficients for the sociodemographic variables to describe unobserved heterogeneity across drivers, and $\boldsymbol{\delta}$ is a column vector of coefficients associated with the interaction terms. We also include random parameters as suggested in the literature (e.g., Sharma et al. (2020a), Fountas et al. (2018a), Fountas and Anastasopoulos (2017), Mannering et al. (2016)), namely $\boldsymbol{\beta}_i$ is a column vector of driver-specific random parameters for the operational variables defined as

$$\boldsymbol{\beta}_i = \boldsymbol{\mu} + \boldsymbol{\Psi}\mathbf{z}_i + \boldsymbol{\Omega}\boldsymbol{\varphi}, \tag{2}$$

where $\boldsymbol{\varphi}$ is a column vector of independent standard normally distributed random variables. As a result, the mean of the distribution of $\boldsymbol{\beta}_i$ is equal to $\boldsymbol{\mu} + \boldsymbol{\Psi}\mathbf{z}_i$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ are a column vector and a matrix of coefficients, respectively, the latter describing unobserved heterogeneity across drivers with respect to the sensitivity towards traffic operational conditions. The covariance of the distribution of $\boldsymbol{\beta}_i$ (to describe unobserved heterogeneity) is equal to matrix $\boldsymbol{\Omega}\boldsymbol{\Omega}'$, where $\boldsymbol{\Omega}$ is a lower triangular matrix in the Cholesky decomposition that contains information about variances as well as covariances to explicitly account for correlations in the coefficients (Fountas et al., 2018b, Greene, 2012). Let $\boldsymbol{\sigma}$ denote the non-zero elements in matrix $\boldsymbol{\Omega}$.. The standard deviation of the k-th random parameter in vector $\boldsymbol{\beta}_i$, denoted by β_{ki}, can be obtained as $\sqrt{\operatorname{var}(\beta_{ki})} = \sqrt{\sigma_{k,k}^2 + \sigma_{k,k-1}^2 + \sigma_{k,k-2}^2 + \cdots + \sigma_{k,1}^2}$, where the indices refer to positions in matrix $\boldsymbol{\Omega}$.

Both probit and logit model formulations were tested, and the logit formulation was found to outperform its counterpart in terms of statistical fit, namely McFadden's pseudo ρ^2 and Akaike Information Criterion (AIC). Considering the logit model, the probability of observing $y_{ij} = 1$ in the data is given by

$$p_{ij} = \Pr(y_{ij} = 1) = \frac{1}{1 + e^{-V_{ij}}}. \tag{3}$$

Given that there are two interactions with a traffic light by the same driver, it is likely that behavioral responses to both interactions are similar (Pantangi et al., 2019). To account for repeated observations of the same participant, also referred to as panel data, we explicitly consider correlations across observations in the baseline and connected environment scenarios by taking the same draw from the distribution of $\boldsymbol{\beta}_i$ for both scenarios (Huo et al., 2020, Sharma et al., 2020b).

Maximum likelihood estimates for coefficients $(\alpha, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\Psi}, \boldsymbol{\sigma})$ are obtained by maximizing the following loglikelihood function:

$$L = \sum_i \int_{\boldsymbol{\beta}_i} \ln\left(\prod_j p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \right) f(\boldsymbol{\beta}_i \mid \boldsymbol{\mu}, \boldsymbol{\Psi}, \boldsymbol{\sigma}) d\boldsymbol{\beta}_i, \tag{4}$$

where the product over the scenarios accounts for the panel nature of the data (see Revelt and Train (1998)) and the integral considers the expectation over all possible values of $\boldsymbol{\beta}_i$ where f is the probability density function of the corresponding multivariate normal distribution that depends on distributional coefficients. We use Monte Carlo simulation with 1000 quasi-random Halton draws to numerically approximate the integral. We also tested log-normal, Weibull, uniform, and triangular distributions, but the normal distribution density function outperformed others in terms of statistical fit and interpretation, which corroborates the safety literature (Eker et al., 2019, Pantangi et al., 2019, Mannering and Bhat, 2014).

As this study finds more than one random parameter to be statistically significant (see Section 4.3), the potential correlation between the random parameters is captured. Note that such a modeling approach is frequently used in the safety literature and called a *correlated grouped random parameters logit with heterogeneity-in-means* approach, whereas the correlation between random parameters can be obtained using the elements in variance-covariance matrix $\boldsymbol{\Omega}\boldsymbol{\Omega}'$ as $\operatorname{corr}(\beta_{ki}, \beta_{li}) = \operatorname{cov}(\beta_{ki}, \beta_{li}) / \left(\sqrt{\operatorname{var}(\beta_{ki}) \operatorname{var}(\beta_{li})}\right)$ where k and l refer to rows in $\boldsymbol{\beta}_i$.

To evaluate the validity of the parameter estimates and provide easy and straightforward interpretation of each explanatory variable on the probability of yellow light running, (point) elasticities for continuous variables and (arc) pseudo-elasticities for categorical variables are calculated using the fitted model. Note that the mathematical formulations of these elasticities are omitted for brevity purpose, and interested readers are referred to Washington et al. (2020) for more details. While the elasticity measure indicates the percentage effect of 1% change in a continuous explanatory variable on the yellow light running probability, the pseudo-elasticity is an arc elasticity that explains the percentage effect on the yellow light running probability of an indicator variable when its value changes from zero to one.

In an effort to justify the superiority of the adopted approach (i.e., a correlated grouped random parameters logit model with heterogeneity-in-means) over its competing approaches (e.g., an uncorrelated grouped random parameters logit model with heterogeneity-in-means and a fixed parameters logit model), likelihood ratio tests are conducted, whose statistic can be calculated as $\chi^2 = -2[L_1 - L_2]$ (Washington et al., 2020), where L_1 and L_2 are the loglikelihood values at the convergences of two competitive models. χ^2 is chi-squared distribution with degrees-of-freedom corresponding to the difference of explanatory parameters between the two competitive models. In addition, goodness-of-fit measures such as AIC, which penalizes for additional parameters in the model, and McFadden pseudo ρ^2 are employed for model comparison.

Specifying the best subset of explanatory variables that often includes main effects and potential interactions among them is challenging primarily because of limited prior knowledge of the underlying relationships. To this end, the analyst selects *a priori* second- and higher-order interaction effects and non-linearities associated with main effects in conventional approaches before the model estimation. However, it is practically impossible to cater for all the possible combinations of main effects and potential higher-order interaction effects that tend to grow geometrically and exponentially, respectively, with the number of ordinal and nominal variables (Haque et al., 2016a). This issue poses a problem of judiciously selecting and omitting variables in a model, which may lead to misspecification issues like omitted variable bias (Washington et al., 2020, Mannering et al., 2016, Mannering and Bhat, 2014).

To overcome this problem, this study employs a hybrid approach of data mining (i.e., decision tree) and advanced econometric modeling approach (i.e., panel mixed logit model, more specifically, *the correlated grouped random parameters logit model with heterogeneity-in-means*). At the first level, a decision tree analysis, which is a non-parametric method to obtain possible interactions by classifying the observations in the predictor space in an iterative process, is performed. Various potential predictors can be identified during this decision tree classification, where each predictor receives various cut-off values. However, it is often reported that decision trees are often associated with *type I* error due to this multiplicity, making it hard to obtain proper inferences about the underlying relationships. Nevertheless, decision tree analysis can be used to obtain *a priori* knowledge obtained from tree branches and can assist in determining which interaction effects to include in the logit model. At the second stage, the model is estimated by considering relevant interactions from the decision tree. This combined approach allows the consideration of higher-order interaction effects (using decision tree) and makes inferences about model output (Washington et al., 2020, Haque et al., 2016a).

4. Results

4.1 Decision tree

We use a hybrid approach to determine relevant higher-order interaction effects in a systematic way using a data mining technique. More specifically, this study employed a decision tree based on Chi-Squared Automatic Interaction Detection (CHAID) algorithm using the '*CHAID*' library in Python (Ramotowski and Fitzgerald, 2020). This algorithm constructs a tree from various possible combinations and divisions on the basis of chi-square tests with a corresponding *p*-value of less than 0.05. While the dependent variable is a binary outcome (i.e., a driver deciding to stop at a signalized intersection or not) in the decision tree, the input variables are driving conditions, traffic operational variables, and driver demographics, as shown in Table 2. To construct the tree, a *k*-fold validation was performed, where we consider $k = 10$ to divide the entire dataset into 10 unique subsets, and each subset was used to assess the tree structure. In this process, each cycle used nine-tenths of the data to train the decision tree. The decision tree correctly classified 79% of cases using 19 numbered leaves (terminal nodes), see Figure 2. Driving condition reveals the highest information gain, and thus is located at the extreme left (or top) of the tree. Table 3 presents these 19 terminal nodes that serve as potential interaction terms for inclusion in vector **w** in our econometric model.

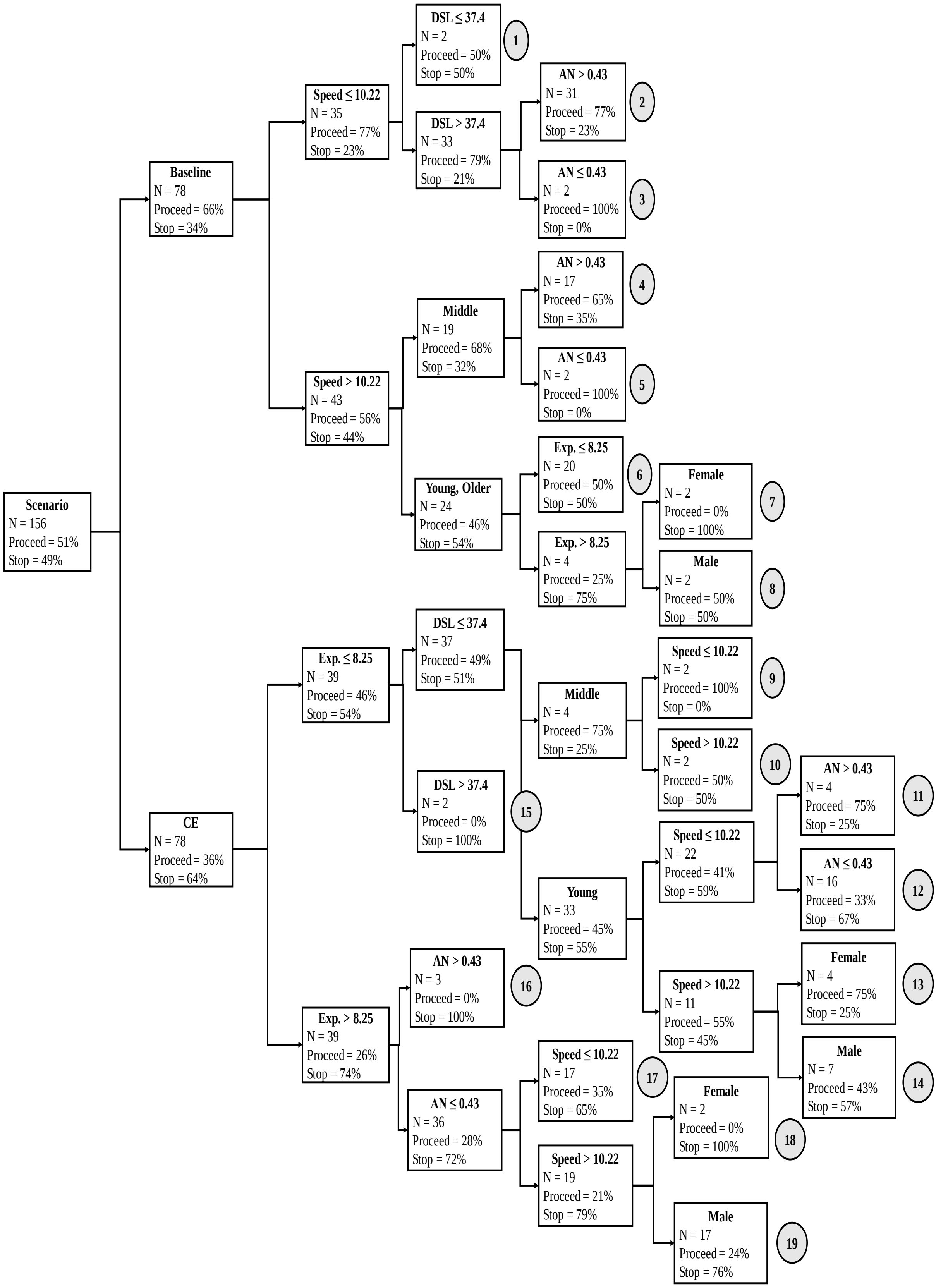
The decision tree classifies the driver decisions to stop or proceed by dividing the data into 37 smaller and homogenous groups, and their corresponding statistics are presented within each node (see Figure 2). The total number of cases reaching each node (*N*) and the percentage of stopping and proceeding at the intersection for that particular node are listed in Figure 2. For instance, terminal node 1 indicates that about 50% of drivers in the baseline condition with speed ≤ 10.22 m/s and distance to the stop line ≤ 37.4 m are likely to proceed. Similarly, terminal node 8 implies that 50% of young (or older) male drivers in the baseline condition with experience > 8.25 years are likely to stop. Terminal node 18 shows that 100% of female drivers in a connected environment with experience > 8.25 years, acceleration noise ≤ 0.43 m/s^2, and speed > 10.22 m/s will stop.

Table 3. Interaction effects obtained from the decision tree and their descriptions

No.	Description
1	Drivers in the baseline condition with speed ≤ 10.22 m/s and DSL ≤ 37.4 m
2	Drivers in the baseline condition with speed ≤ 10.22 m/s, DSL > 37.4 m, and AN > 0.43 m/s^2
3	Drivers in the baseline condition with speed ≤ 10.22 m/s, DSL > 37.4 m, and AN ≤ 0.43 m/s^2
4	Middle-aged drivers in the baseline condition with speed > 10.22 m/s and AN > 0.43 m/s^2
5	Middle-aged drivers in the baseline condition with speed > 10.22 m/s and AN ≤ 0.43 m/s^2
6	Young (or older) drivers in the baseline condition with experience ≤ 8.25 years
7	Young (or older) female drivers in the baseline condition with experience > 8.25 years
8	Young (or older) male drivers in the baseline condition with experience > 8.25 years
9	Middle-aged drivers in CE with experience ≤ 8.25 years, speed ≤ 10.22 m/s, and DSL ≤ 37.4 m
10	Middle-aged drivers in CE with experience ≤ 8.25 years, speed > 10.22 m/s, and DSL ≤ 37.4 m
11	Young drivers in CE with experience ≤ 8.25 years, DSL ≤ 37.4 m, speed ≤ 10.22 m/s, and AN > 0.43 m/s^2
12	Young drivers in CE with experience ≤ 8.25 years, DSL ≤ 37.4 m, speed ≤ 10.22 m/s, and AN ≤ 0.43 m/s^2
13	Young female drivers in CE with experience ≤ 8.25 years, DSL ≤ 37.4 m, speed > 10.22 m/s, and AN > 0.43 m/s^2
14	Young male drivers in CE with experience ≤ 8.25 years, DSL ≤ 37.4 m, speed > 10.22 m/s, and AN > 0.43 m/s^2
15	Drivers in CE with experience ≤ 8.25 years and DSL > 37.4 m
16	Drivers in CE with experience > 8.25 years and AN > 0.43 m/s^2
17	Drivers in CE with experience > 8.25 years, AN ≤ 0.43 m/s^2, and speed ≤ 10.22 m/s
18	Female drivers in CE with experience > 8.25 years, AN ≤ 0.43 m/s^2, and speed > 10.22 m/s
19	Male drivers in CE with experience > 8.25 years, AN ≤ 0.43 m/s^2, and speed > 10.22 m/s

CE, DSL, AN, Exp., and speed respectively denote connected environment, distance to the stop line at onset of a yellow light (m), acceleration noise (m/s^2), experience (years), and speed at onset of a yellow light (m/s).

Logit models with and without interaction effects were compared, and it was found that although both the models possess a reasonable explanatory power, the model with interaction effects outperformed the counterpart model based on goodness-of-fit statistics (AIC and McFadden's pseudo ρ^2). Thus the model with interaction effects is considered hereafter.



Note that numbers in circle indicate the interaction term; C.E., DSL, AN, Exp., and speed respectively denote connected environment, distance to the stop line at the onset of a yellow light (m), acceleration noise (m/s^2), experience (years), and speed at the onset of a yellow light (m/s).

Fig. 2. Decision tree schematic for the stop/proceed decision model

4.2 Model selection

Apart from estimating the correlated grouped random parameters logit with heterogeneity-in-means (CGRPLHM) model, this study also estimated an uncorrelated grouped random parameters logit with heterogeneity-in-means (UGRPLHM) model and a fixed parameters (FP) model to evaluate the best performing model. To assess the overall statistical performance of the estimated models, following metrics are used: the log-likelihood value at convergence, L (the higher the better), the log-likelihood value with only a constant, L_0, AIC (the smaller the better), and McFadden's pseudo ρ^2 (the larger the better).

Table 4 presents the statistical model fits for all three models. The AIC value is 190.4 in the FP model, which is decreased to 189.1 and 184.1, respectively, in the UGRPLHM and CGRPLHM models, while the McFadden ρ^2 is increased from 0.059 in the FP model to 0.153 in the CGRPLHM model. These statistics reflect the better performance of the CGRPLHM model, which is also confirmed by performing likelihood ratio tests, and results are presented in Table 4. Following observations can be made from these results: (a) both the UGRPLHM and CGRPLHM models show better performance compared to the FP model at a 95% confidence level; and (b) a higher χ^2 statistics (i.e., 9.1) is obtained when comparing the UGRPLHM and CGRPLHM models, implying the superior performance of the CGRPLHM model (the critical value is 5.99 with two degrees-of-freedom), further ensuring the appropriateness of the CGRPLHM model for this study.

Table 4. Summary of statistical fits of the models considered in this study

Candidate model	L_0	L	*df*	χ^2	AIC	McFadden's pseudo ρ^2
Fixed parameters model (FP)	-92.13	-88.20	7	7.86	190.4	0.042
Uncorrelated grouped random parameters logit with heterogeneity-in-means model (UGRPLHM)	-92.13	-82.56	12	19.14	189.1	0.104
Correlated grouped random parameters logit with heterogeneity-in-means model (CGRPHM)	-92.13	-78.01	14	28.24	184.1	0.153
Comparisons (H_0 = simpler model is better)			*df*	χ^2	*p*-value	Remark
FP versus UGRPLHM			5	11.28	0.045	UGRPLHM is superior
FP versus CGRPLHM			7	20.38	0.004	CGRPLHM is superior
CGRPLHM versus UGRPLHM			1	9.10	0.002	CGRPLHM is superior

df: degrees-of-freedom

4.3 Model interpretation

Table 5 presents the estimation results for the correlated grouped random parameters logit with heterogeneity-in-means model fitted to the driver decisions of proceeding through the intersection at the onset of yellow light. The dummy variable for the connected environment and distance to the stop line variables are found to be random and normally distributed, which is consistent with the literature (Ali et al., 2020d, Fountas et al., 2019). Moreover, unobserved heterogeneity in the connected environment is associated with gender. The non-random parameters in the model are speed at the onset of yellow light, acceleration noise, dummy variables for young and older drivers. Systematic utility function (1) can be written as

$$
\begin{aligned}
V = & -0.946 \\
& + \beta_{CE} \times \text{CE} + 0.224 \times \text{speed} - 1.685 \times \text{acc. noise} + \beta_{DSL} \times \text{distance to the stop line} \\
& + 0.6 \times \text{YoungDriver} - 0.972 \times \text{OlderDriver} \\
& + 0.914 \times \text{Interaction Term 2} - 1.262 \times \text{Interaction Term12},
\end{aligned} \tag{5}
$$

where the first line contains the constant, the second line the traffic operational variables, the third line the sociodemographic variables, the fourth line the interaction terms, and where

$$
\begin{pmatrix} \beta_{\text{CE}} \\ \beta_{\text{DSL}} \end{pmatrix} = \begin{pmatrix} -0.889 \\ -0.024 \end{pmatrix} + \begin{pmatrix} 0.841 \\ 0 \end{pmatrix} \times \text{FemaleDriver} + \begin{pmatrix} 2.531 & 0 \\ -0.26 & 0.058 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} \tag{6}
$$

is the specified correlation structure between random parameters with φ_1 and φ_2 be the independent standard normally distributed random variables.

The diagonal and below diagonal elements of Cholesky matrix for each random parameter are given in Table 5. The standard deviation of each random parameter can be calculated as the square root of the variance (elements on the diagonal of the variance-covariance matrix, which can be calculated as $\mathbf{\Omega\Omega}'$). For instance, the standard deviations of the connected environment and distance to the stop line variables are calculated as $\sqrt{6.405} = 2.531$ and $\sqrt{0.071} = 0.266$, respectively.

Table 5. Estimation results of the correlated grouped random parameters logit with heterogeneity-in-mean model

						95% CI of estimate	
Variable	*estimate*	*s.e.*	*z*-value	*p*-value	elasticity	lower	upper
Non-random parameters							
Constant	-0.946	0.329	-2.87	0.004	—	—	—
Speed at onset of a yellow light	0.224	0.101	2.20	0.027	1.607	0.026	0.421
Acceleration noise	1.685	0.742	2.27	0.023	4.376	0.231	3.139
Young driver	0.600	0.303	1.98	0.048	0.189	0.001	1.193
Older driver	-0.972	0.421	-2.31	0.021	-0.131	-1.797	-0.148
Interaction term 2	0.914	0.404	2.26	0.023	0.156	0.122	1.705
Interaction term 12	-1.262	0.627	-2.01	0.044	-0.080	-2.490	0.033
Random parameters							
Connected env. (mean)	-0.889	0.440	-2.02	0.043	-0.346	-1.751	-0.026
Distance to stop line (mean)	-0.024	0.012	-2.00	0.045	-0.453	-0.048	-0.001
Diagonal values in Cholesky matrix							
Connected env.	2.531	1.145	2.21	0.027	—	—	—
Distance to stop line	0.260	0.082	3.17	<0.001	—	—	—
Below diagonal values in Cholesky matrix							
Distance to the stop line: connected environment	0.058	0.014	4.14	<0.001	—	—	—
Heterogeneity in mean of connected environment							
Female	0.841	0.403	2.08	0.037	0.125	0.051	1.631

L = -78.01; L_0 = -92.13; Likelihood ratio = 28.24 (p-value < 0.001); McFadden pseudo ρ^2 = 0.153; AIC = 184.1; No. of observations = 156; No. of groups = 78; Group size = 2

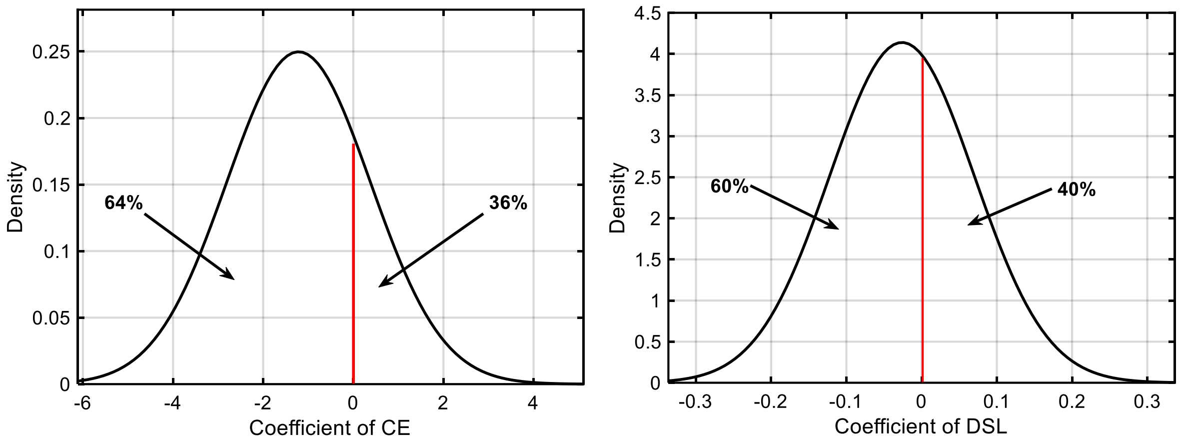
Speed at the onset of the yellow light is a significant predictor and positively associated with driver decisions (Table 5). The model suggests that drivers are more likely to proceed through the intersection with an increased speed. More specifically, with every one percent increase in the speed, the probability of proceeding through the intersection at the onset of the yellow light increases by 1.61%. This finding is intuitive because drivers with higher speeds often think that they can cross the intersection without red light violations and do not want to disrupt their continuous motion by braking hard to stop before the stop line.

Acceleration noise (or variation) has a significant and positive impact on driver decisions at the onset of the yellow light, as shown in Table 5. Results reveal that drivers with higher acceleration noise are more likely to proceed at the onset of the yellow light, with the probability of yellow light running increases by 4.38% with every one percent increase in the acceleration noise. This result implies that reckless drivers, indicated by higher acceleration noise, have a higher tendency to proceed through the intersection at the onset of the yellow light.

Young and older drivers have a higher and a lower propensity of proceeding through the intersection at the onset of the yellow light compared to middle-aged drivers (Table 5), respectively. More specifically, compared to middle-aged drivers, the probability of yellow light running increases and decreases by 0.19% and 0.13%, respectively, for young and older drivers. As it is well-known that younger drivers are often risk-takers, and older drivers are conservative (Ali et al., 2019a), our findings are intuitive and in line with the literature. We further elaborate on these findings in detail in the next section.

Apart from the main effects, the developed model contains two interaction terms. *Interaction term 2* shows that drivers in the baseline condition with speed ≤ 10.22 m/s, distance to the stop line > 37.4 m, and acceleration noise ≤ 0.43 m/s^2 are more likely to proceed through the yellow light at the intersection, with the corresponding increase in the probability of 0.16% (Table 5). Similarly, *interaction term 12* indicates that young drivers in the connected environment with speed ≤ 10.22 m/s, distance to the stop line ≤ 37.4 m, acceleration noise > 0.43 m/s^2, and experience ≤ 8.25 years have a lower propensity of proceeding through the intersection at the onset of yellow light.

Table 5 indicates that the mean (z-stats = -2.02; p-value = 0.043) and standard deviation (z-stats = 3.89; p-value < 0.001) of the connected environment dummy variable are statistically significant. Figure 3(a) shows the distribution of the connected environment's coefficients, reflecting a significant heterogeneity in driver decisions in the connected environment where according to Figure 3(a), the probability of yellow light running decreases for most drivers (64%), but not necessarily for all. This finding implies that driver decisions in the connected environment are not monotonous, as there exist two classes of drivers: one who stops before the stop line and one who proceeds through the intersection at the yellow light. This result suggests that not all the drivers use the advance information provided by the connected environment in the same way, as some drivers use this information to stop prior to the stop line, reflecting their safer behavior, while others use it in a counterproductive manner and proceed through the intersection.



(a) (b)

Fig. 3. Distributions of coefficients of (a) the connected environment; (b) Distance to stop line; note that the distribution of coefficients is obtained by keeping one of the two random parameters fixed at the mean value

We also find that heterogeneity in driver decisions in the connected environment is related to gender (Table 5). More specifically, female drivers reveal a higher likelihood of proceeding through the intersection at the onset of the yellow light in the connected environment with the probability of 0.13% compared to male drivers.

Table 5 also reveals that not just the mean of the distance to the stop line is statistically significant (*z*-stats = -2.00; *p*-value = 0.045), but also its standard deviation (*z*-stats = 4.10; *p*-value < 0.001), indicating a significant heterogeneity in driver decisions corresponding to the distance to the stop line. Figure 3(b) shows the existence of heterogeneity where the probability decreases for most drivers (60%), but not for all. The negative sign of the distance to the stop line variable implies that when this distance increases, the probability of proceeding through the intersection decreases, which is intuitive because drivers have a large distance to safely stop before the stop line. On the other hand, this probability increases for some drivers who tend to be aggressive and often accelerate to proceed through the intersection at the onset of yellow light.

Table 5 also presents the diagonal and below diagonal elements of Cholesky matrix, which can be used to calculate variance-covariance of the correlated random parameters and thereby assist in calculating the correlation coefficient between two random parameters (see detailed calculations in Section 3.3). We find that distance to the stop line and connected environment are statistically correlated at a 5% significance level (*t*-stats = 3.13; *p*-value = 0.001) with a covariance of -0.66 and a correlation coefficient of 0.27. Note that *t*-stats is calculated following the post-estimation technique presented in (Fountas et al., 2018a), and for mathematical details, we refer interested readers to their study. It is worth noting that the correlation between random parameters suggests the existence of interactions of unobserved characteristics associated with the explanatory variables with correlated random parameters (Huo et al., 2020). More specifically, a positive correlation of random parameters implies homogeneous effects of unobserved characteristics of driver decisions to stop or proceed, whereas a negative correlation of random parameters suggests mixed effects of unobserved characteristics on driver decisions. In this study, we find a positive correlation between distance to the stop line and connected environment, reflecting a homogeneous effect of the unobserved characteristics associated with driver decisions related to distance to the stop line in the connected environment. In other words, an increase in the effect of the distance to the stop line (represented by β_{DSL}) in the connected environment would increase the probability of running the yellow light because of unobserved heterogeneity associated with these two variables. This result implies that drivers receive information in advance about traffic light change and use such information to navigate through the intersection, reflecting that they are aware of the time left for the signal to turn red, and they decide to cross the intersection in the given time without causing a red light violation.

5. Discussion

5.1 Driver decisions in the connected environment

Driver decisions and subsequent actions approaching a signalized intersection are regarded as critical because of their direct implications on traffic safety (Papaioannou, 2007). An uncertainty in driver decisions may cause a rear-end collision (if a driver decides to stop and applies sudden hard braking) or angle collision (if the driver decides to proceed). This uncertainty mainly arises when a traffic light suddenly changes, and the driver finds him/herself in the dilemma zone. To this end, a connected environment provides advance information that is expected to minimize (if not completely eliminate) the uncertainty associated with driver decision-making. As such, the developed model can provide insights into the probabilities of drivers' running the yellow light as a function of driving condition, traffic operational variables, and driver demographics. More specifically, the probabilities can be calculated using the parameter estimates reported in Table 5 together with the mean values of the continuous explanatory variables and reference category for categorical variables. Note that the probabilities obtained from Equations (7) and (8) and depicted in Figure 4 are calculated for the reference category participants in the baseline and connected environment driving conditions, reflecting the average probabilities for middle-aged male drivers. The predicted probability for drivers' running the yellow light in the baseline (without advance information) for a driving speed of 9 m/s can be computed as follows:

$$p_{\text{Baseline}} = \frac{\exp(-0.946+0.224\times 9+1.685\times 0.45+0.6\times 0-0.972\times 0-0.024\times 40.47+0.914\times 0-1.262\times 0+(-0.889\times 0+0.841\times 0))}{1+\exp(-0.946+0.224\times 9+1.685\times 0.45+0.6\times 0-0.972\times 0-0.024\times 40.47+0.914\times 0-1.262\times 0+(-0.889\times 0+0.841\times 0))}=0.70 \tag{7}$$

Similarly, the corresponding probability for the connected environment can be computed as follows:

$$p_{\text{CE}} = \frac{\exp(-0.946+0.224\times 9+1.685\times 0.45+0.6\times 0-0.972\times 0-0.024\times 40.47+0.914\times 0-1.262\times 0+(-0.889\times 1+0.841\times 0))}{1+\exp(-0.946+0.224\times 9+1.685\times 0.45+0.6\times 0-0.972\times 0-0.024\times 40.47+0.914\times 0-1.262\times 0+(-0.889\times 1+0.841\times 0))}=0.49 \tag{8}$$

The probabilities of running the yellow light for the speed of 9 m/s (or, approximately 30 km/h) are respectively 70% and 49% for the baseline and connected environment (Figure 4(a)), suggesting a 21% reduction in the probability, which is attributed to the advance availability of the traffic signal information in the connected environment. This result further highlights the benefits of the connected environment in assisting drivers to make safer and informed decisions. Interestingly, the benefit of advance information is found to be a function of driver's approaching speed, i.e., the lower the speed, the higher the benefit (in other words, a higher reduction in the probability of yellow light running, see Figure 4(a)). This can be explained by the fact that when drivers are driving at higher approaching speeds, they tend to utilize the information presented by the connected environment to traverse the intersection, keeping in mind the time left for the signal to turn red from green. A similar interpretation can made for the relationship of acceleration noise with the probability of yellow light running (see Figure 4(b) for more details).

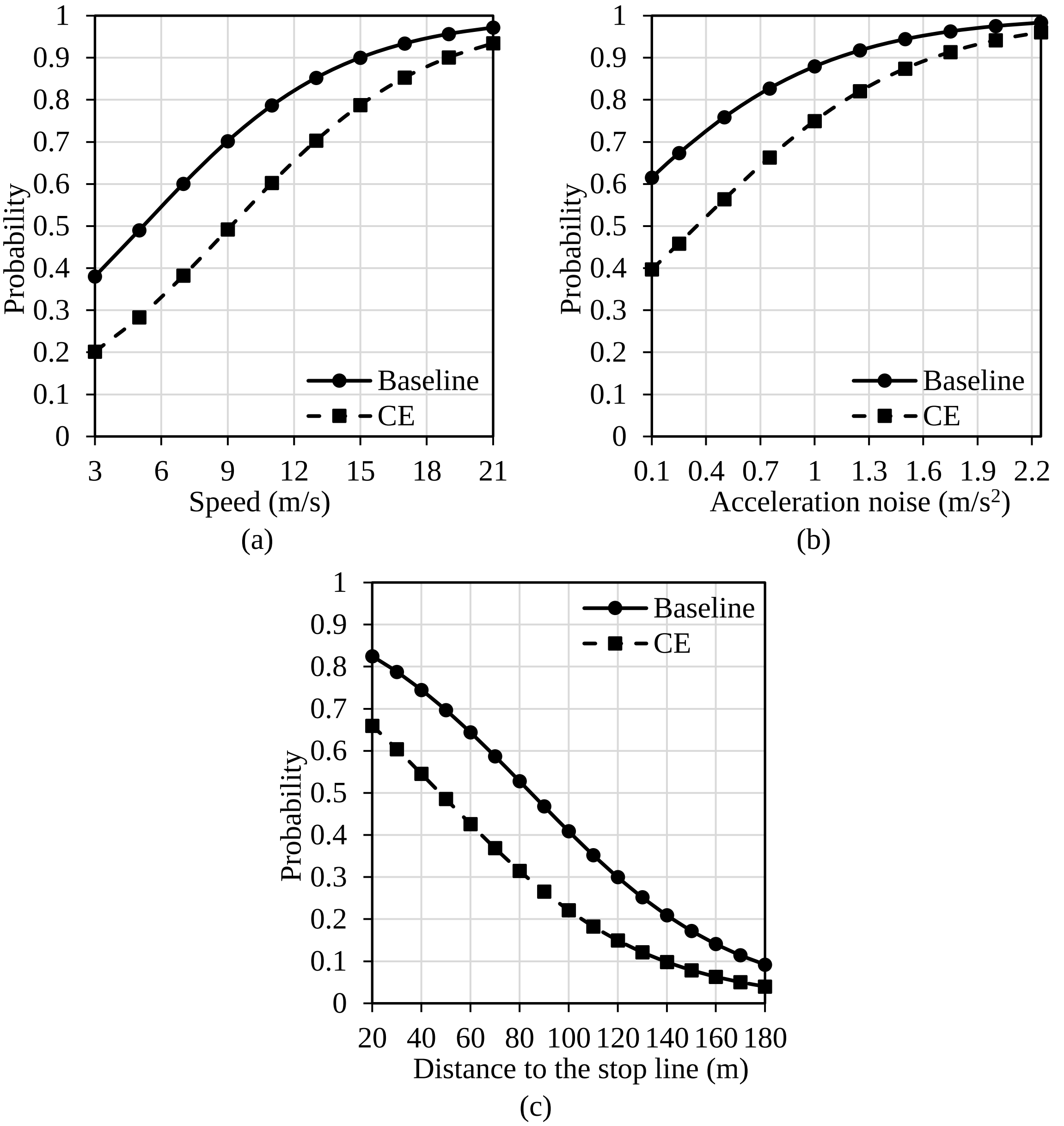


Fig. 4. Probabilities of running the yellow light in different conditions as a function of (a) speed at the onset of the yellow light; (b) acceleration noise; and (c) distance to the stop line

Some previous studies also highlighted the benefits of the connected environment. For instance, Sharma et al. (2020a) reported that advance information disseminated via a connected environment provided additional time to drivers in a hard-braking event, where drivers were found to decelerate smoothly. In another study where drivers were given advance information about congestion ahead, it was found that drivers performed discretionary lane-changing earlier with a higher safety margin in a connected environment (Ali et al., 2020c). In line with these studies, we also observe that the connected environment assists most of the drivers to comply with the traffic lights.

Drivers' approaching speeds have been repeatedly noted in the literature as a contributing factor to their decisions of stopping or proceeding at the onset of the yellow light. In general, drivers tend to drive as close as possible to the posted limit, but in some cases, they may violate the posted speed limit, and in turn, find themselves in a dilemma to stop or proceed. On the contrary, the connected environment provides event-based warning information whenever a driver exceeds the posted speed limit, which may result in selecting a lower driving speed. To examine whether any speed reduction is observed in our dataset, drivers' approaching speeds to the signalized intersection are tested and compared between two driving conditions using a paired t-test, as used in our previous studies (Ali et al., 2018, Haque et al., 2016b). Results reveal that the difference in the approaching speed at the onset of the yellow light is statistically significant ($t = 3.56$, p-value = 0.03) between two driving conditions. More specifically, the mean speeds at the onset of the yellow light in the baseline and connected environment driving conditions are respectively 10.16 m/s and 9.65 m/s. Drivers, on average, are found to drive 0.5 m/s slower while driving in the connected environment.

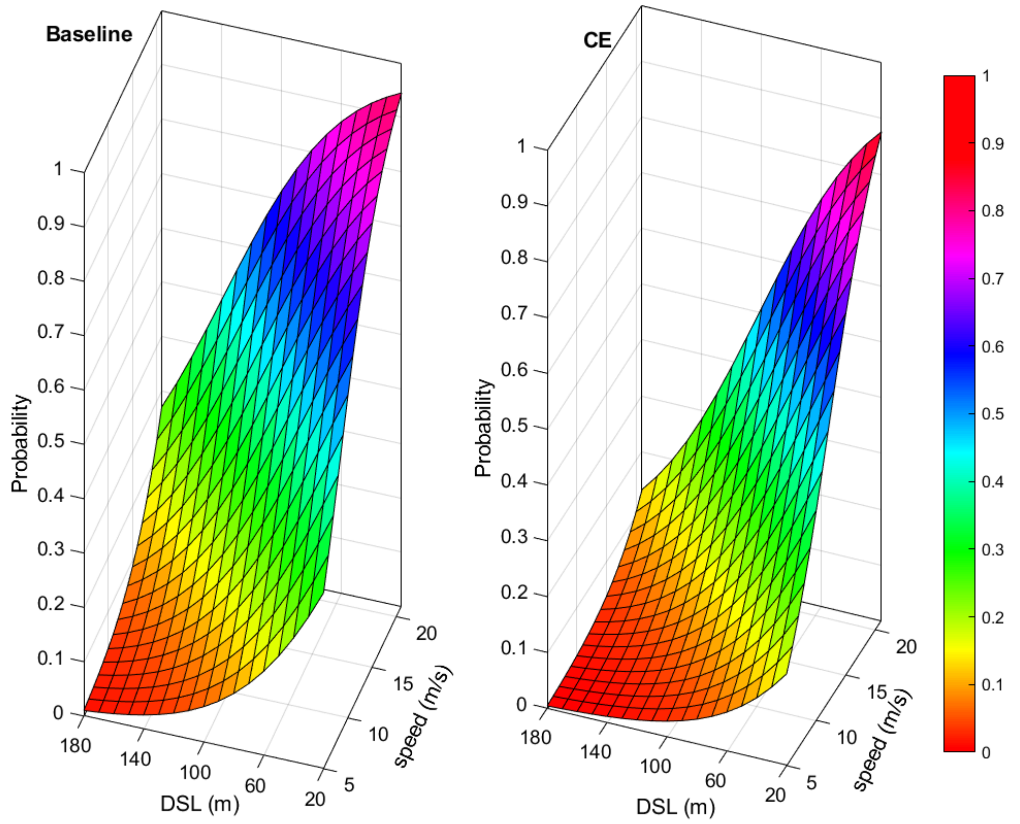
Table 6. Approaching speed selection of different driver groups at the onset of a yellow light

Speed (m/s)	Driving condition		Significance by a paired t-test	Remark
	Baseline	CE		
All drivers	10.12	9.66	$t = 3.56$, p-value = 0.03	Significant
Age group				
Young	10.05	9.76	$t = 1.01$, p-value = 0.21	Not significant
Middle-aged	10.38	9.44	$t = 4.41$, p-value = 0.02	Significant
Older	9.65	10.07	$F = 0.61$, p-value = 0.30	Not significant
Gender				
Male	10.27	9.62	$t = 3.25$, p-value = 0.04	Significant
Female	9.95	9.73	$t = 0.53$, p-value = 0.32	Not significant

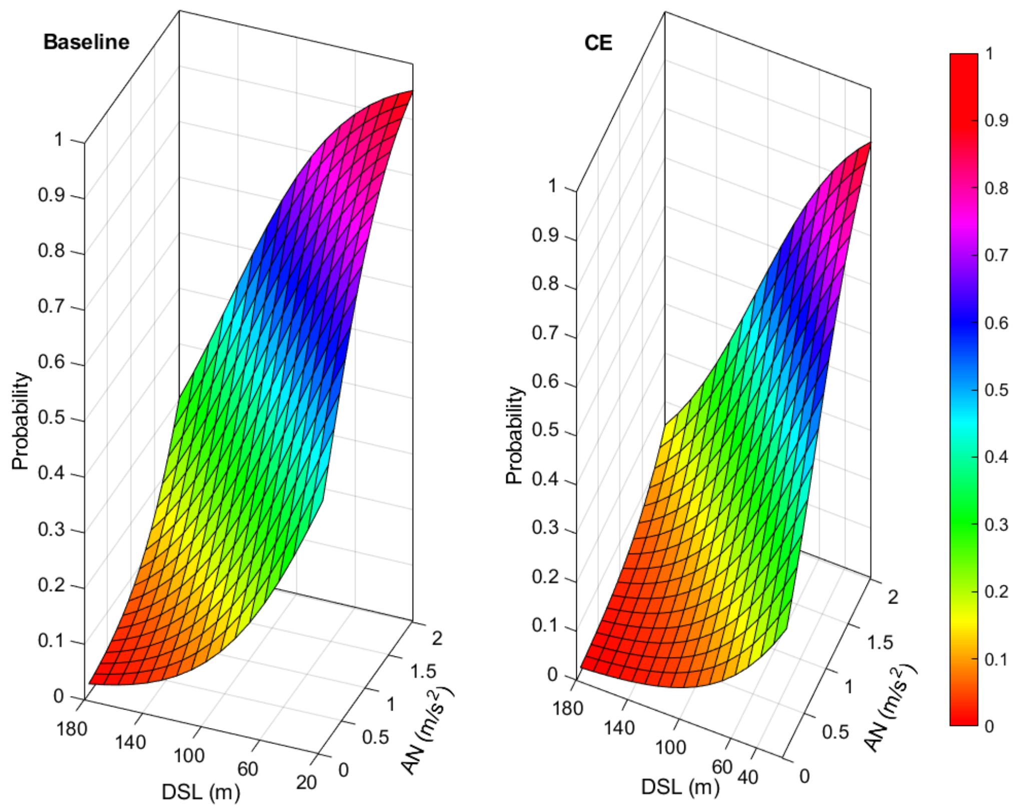
CE: connected environment

Figures 4(b) and 4(c) display the probability of yellow light running in the baseline and connected environment driving conditions corresponding to different acceleration noise and distance to the stop line values, respectively, and these probabilities can be interpreted in a similar manner.

Furthermore, using the developed model, probability surface plots are generated as a function of distance to the stop line, speed at the onset of yellow light, and acceleration noise, while controlling for other exogenous variables. To illustrate the impact of driving conditions, probability surfaces specific to the driving condition (e.g., baseline and connected environment) are developed and presented in Figure 5. These plots clearly highlight how under the connected environment condition, the probability of yellow light running drops significantly with higher distance and lower speeds, compared to that of the baseline condition (Figure 5(a)). Similarly, with lower distance and higher acceleration noise, the probability of yellow light running reduces significantly in the connected environment compared to the baseline condition (Figure 5(b)). These results imply the effectiveness of the connected environment in reducing the likelihood of yellow light running, thereby improving safety.



(a) Speed and distance to the stop line (DSL)



(b) Acceleration noise (AN) and distance to the sop line (DSL)

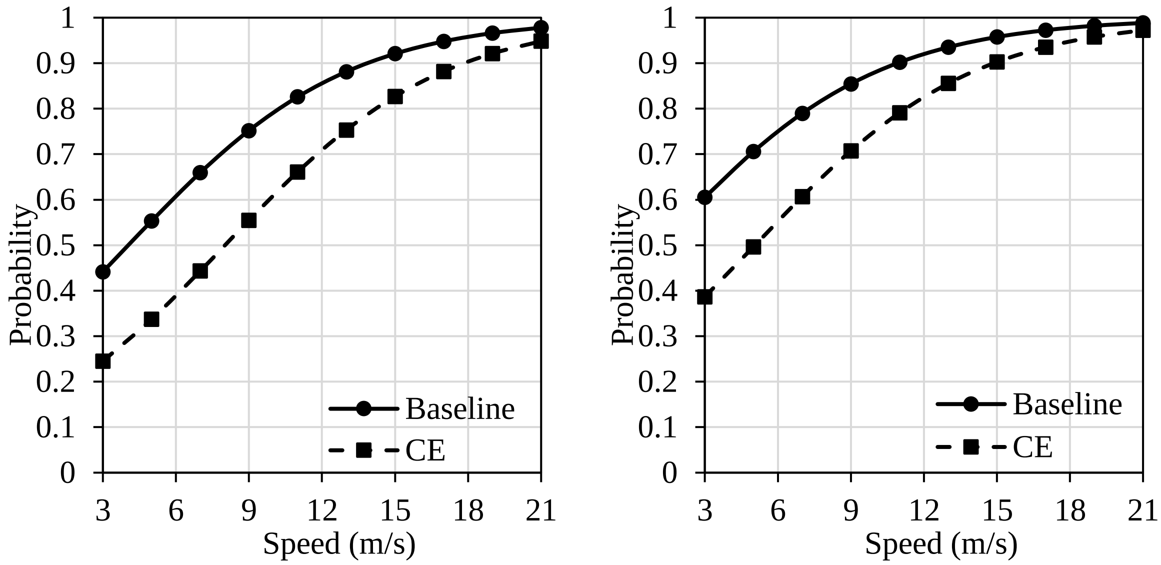
Fig. 5. The impact of interaction effects on the probability of yellow light running

5.2 Impact of driver demographics on the probability of yellow light running

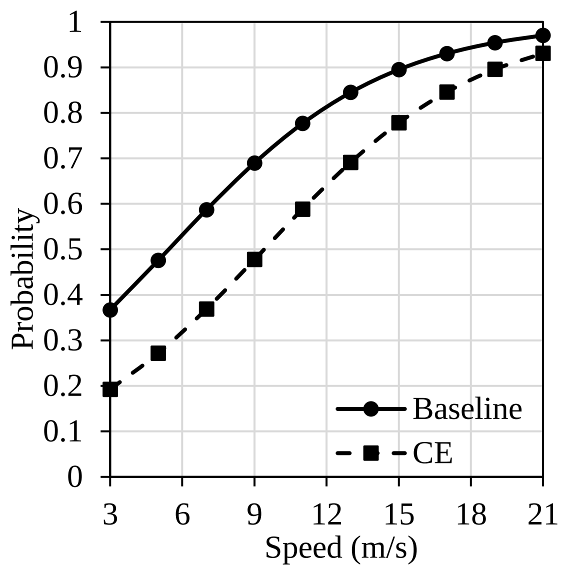
5.2.1 Driver age

Figure 6 displays the probability of yellow light running across all age groups. It can be observed that the probabilities tend to increase for all age groups with the increase in approaching speed for both the scenarios, while a higher likelihood of running the yellow light is found in the baseline driving condition. For instance, the probability of yellow light running for young drivers in the baseline driving condition at 9 m/s is 75% (Figure 6(a)), while at the same approaching speed, the probability of yellow light running for young drivers in the connected environment is 55%, suggesting a 20% reduction in yellow light running. This reduction in the probability can be attributed to the slower approaching speed selection of young drivers in the connected environment. More specifically, young drivers' approaching speeds were 0.3 m/s lower (but not statistically significant) in the connected environment compared with the speed in the baseline condition (Table 6). In general, young drivers have repeatedly been noted as risky drivers in the literature (Montgomery et al., 2014, Leung and Starmer, 2005), as they have the propensity to proceed through the intersection at the onset of the yellow light either by increasing their speed or causing a red light violation (Yang and Najm, 2006). The connected environment, however, has been found to reduce such risky behavior of young drivers by providing advance information related to traffic light change, as found in this study.

Middle-aged drivers appear to have a lower propensity for yellow light running in the connected environment. In particular, we find that middle-aged drivers show about a 15% reduction in the probability of yellow light running at the speed of 9 m/s in the connected environment compared to the baseline condition (Figure 6(b)). In line with this finding, the approaching speed of middle-aged drivers in the connected environment is found to be 1 m/s lower, which could be one of the reasons for this age group of drivers' decreased yellow light running probabilities. As noted in Khatoon et al. (2013), middle-aged drivers are less risky compared to young drivers, and they are more likely to take better advantage of the available information (Ali et al., 2019a). Consistent with the literature, this study finds that middle-aged drivers' probability of yellow light running is further decreased in the connected environment.



(a) Young drivers (b) Middle-aged drivers



(c) Older drivers

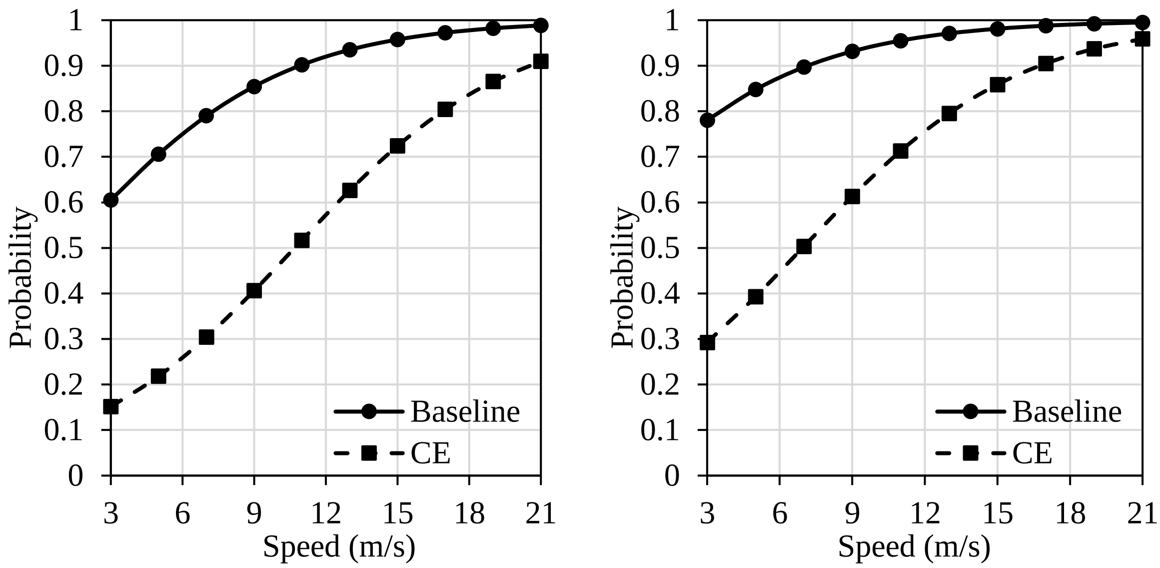
Fig. 6. Impact of the connected environment (CE) on the probability of yellow light running of different age groups

As reported in the literature, older drivers, in general, take more time in processing information, deciding, and taking safe actions to avoid potential safety-critical events (Preusser et al., 1998). Also, when driving without advance information, these drivers are more likely to proceed through an intersection at the onset of the yellow light (Caird et al., 2007). However, this study demonstrates that the probability of yellow light running of older drivers can be significantly decreased in the connected environment when they are assisted with advance information. Such information provides additional time to older drivers, which in turn, they use for making better and safer decisions related to when they should stop, they stop safely, and when they should pass through the intersection, they pass through it efficiently. Although the approaching speed of older drivers is about 0.4 m/s higher (but not statistically significant) in the connected environment, they appear to utilize the information and stop before the stop line more often. For instance, the probability of yellow light running in the baseline condition at the speed of 9 m/s is 68%, while the corresponding probability in the connected environment is 47% (Figure 6(c)), implying a 21% reduction in the yellow light running.

In summary, the connected environment has shown to reduce the probability of yellow light running across all age groups, whereas the older age group has been found to take the most advantage of available information compared to other age groups. To support this argument, the area between curves for the baseline and connected environment is calculated for each age group. The areas for young, middle-aged, and older drivers are 1.52, 0.76, and 1.81, respectively, implying that older drivers benefits more from the connected environment. This finding corroborates with some of the existing literature (Caird et al., 2008, Kramer et al., 2007), suggesting that older drivers are likely to benefit more from the in-vehicle information systems compared to other age groups.

5.2.2 Drivers' gender

Figure 7 represents the probability of yellow light running for both male and female drivers. Note that these probabilities are calculated using Equations (7) and (8), but the only difference is that the heterogeneity in the connected environment is defined for female drivers by setting the dummy variable for female as one. As shown in Figure 7, both male and female drivers appear to reduce their yellow light running probabilities in the connected environment across the whole speed range, with the corresponding probability decrease on average being about 23% and 31%, respectively for male and female drivers in the connected environment compared to the baseline driving condition. The connected environment has been found to provide more advantage to male drivers compared to their female counterparts. For instance, the probability reduction for female drivers in the connected environment compared with the baseline condition at 9 m/s is about 32% (Figure 7(a)), while the corresponding reduction for male drivers is about 45% (Figure 7(b)), implying that male drivers appear to better utilize the advance information from the connected environment. This is also supported by the large area between curves for the baseline and connected environment for male drivers (i.e., 5.80) compared to female drivers (i.e., 4.22). Although previous research has also documented a higher propensity of yellow light running of female drivers when they are driving without driving assistance systems (Yang and Najm, 2006), such higher propensity appears to be reduced when female drivers are assisted with advance information in the connected environment.



(a) Male drivers (b) Female drivers

Fig. 7. Probability of yellow light running for gender type; *CE: connected environment*

Furthermore, this study finds that the approaching speed of male drivers in the connected environment is about 0.65 m/s lower (and statistically significant) than that in the baseline condition, while the corresponding reduction in the speed of female drivers is about 0.22 m/s (statistically insignificant, though). This result also substantiates that the connected environment assists drivers in making better decisions, as they tend to reduce their speeds significantly, avoiding being in the dilemma zone, where drivers can neither cross the intersection without causing red light violations nor stop before the stop line without applying hard braking (Haque et al., 2016a, Papaioannou, 2007).

6. Conclusions

This study examined driver stop/go decisions at the onset of yellow lights at signalized intersections when they are assisted with advance information of traffic light change provided by a connected environment. Data related to driver decisions were obtained from the CARRS-

Q Advanced Driving Simulator. A hybrid framework of decision tree and a panel mixed logit model (more specifically, correlated grouped random parameters logit with heterogeneity-in-means approach) leveraged the strengths of both these approaches, as the former approach heuristically provides information about unknown relationships while the latter approach has the ability to test the significance of observed effects by capturing unobserved heterogeneity associated with driver decisions as well as the correlation between random parameters. Modeling results revealed that although the majority of drivers in the connected environment decide to stop at the onset of the yellow signal, there also exists a class of drivers who decide to proceed through the intersection in the connected environment. Results also uncovered that such heterogeneity is associated with gender, as male drivers are less likely to proceed at the onset of the yellow light in the connected environment compared to female drivers. Moreover, by allowing the correlation between random parameters, it was found that with a higher distance to the stop line in the connected environment, the probability of yellow light running may increase. Furthermore, the speed selection behavior of drivers to approach a signalized intersection was found to be significantly influenced by the connected environment. In general, drivers in the connected environment appeared to select relatively lower approaching speeds. More specifically, young and middle-aged drivers selected lower speeds in the connected environment resulting in a lower probability of yellow light running, unlike older drivers who were found to take the most advantage of the advance information compared to other age groups. Meanwhile, both male and female drivers selected lower approaching speeds, and their probabilities of yellow light running also reduced in the connected environment.

As this study analyzed the effects of a connected environment at signalized intersections for different driver demographics, the resulting impact should be viewed with respect to the age groups and gender within the sample of this study. Note that the age groups considered in this study are aligned with Australian guidelines and some past studies (Tränkle et al., 1990, Zhang et al., 1998, Makishita and Matsunaga, 2008, Cheung and McCartt, 2011, .idcommunity, 2016). However, given the discrepancy in the definition of age groups in the literature (Thompson et al., 2012), future studies can examine whether the impact of the connected environment is sensitive to age group definition. Furthermore, we made significant efforts in ensuring the realistic representativeness of the general population in our participant recruitment; however, it can be observed that the cohort of participants is skewed towards young and male drivers. Future studies can try to maintain an equal ratio of different age groups and gender to obtain a full picture of the connected environment's impact on driver decisions at signalized intersection. In particular, the effects of the connected environment on drivers aged more than 65 years needs to be studied

Although this study employed decision tree analysis to systematically obtain higher order interactions to investigate the complex interactions of driver gender, age group, driving conditions, and traffic operational variables, many of these interaction effects were found to be insignificant and thus dropped (except for two interactions) from the parsimonious model. This restricts the current study from analysing the effects on the connected environment for driver characteristics, such as young female drivers versus young male drivers, etc. A possible reason for such insignificance could be the small sample size. As such, it is recommended for future studies to collect data from more participants to gain more insights into higher-order interactions and the probability of yellow light running. Note that although the probabilities of yellow light running are calculated using the adopted approach (i.e., random parameters with heterogeneity-in-means), this study only uses mean values of the random parameters. As future work, a simulation-based approach could be employed to obtain more insights about the probabilistic nature of driver decisions at the onset of yellow light.

Furthermore, as the time of dissemination of advance information about traffic light change was fixed in the connected environment, more research is required to understand the relationship of varying time with the effectiveness of the connected environment. It will be interesting to see whether driver decisions can change with change in the time when the information is provided. As noted in the literature, driver decisions at the onset of the yellow light are a function of driver's position in a traffic stream. To minimize the confounding factors, this study intentionally did not place other traffic in the direction of travel, which would have restricted us to investigate the effect of driver's position in the traffic stream on driver decisions combined with the promise of a connected environment. Investigating such effects will allow us to develop a relationship of the degree of effectiveness of a connected environment with driver's position in the traffic stream. This study considered a fixed threshold of 5 s for changing the traffic light in the driving simulator experiment to avoid confounding factors in the analysis. It would be interesting to examine driver decisions' sensitivity with respect to different time gaps to the stop line. Findings from such an exercise will add new insights into how the behavioral response towards the yellow light dilemma may be adjusted in the highly anticipated connected vehicle environment. In addition, this study is only concerned with an uninterrupted supply of information aids from a connected environment; however, the information supply could be impaired, such as communication delay, and the study of the effects of such impaired communication merits a research pursuit.

Acknowledgments

The authors would like to acknowledge Mr. Andrew Haines for programming the simulator experiment, and the help of Dr. Anshuman Sharma and Dr. Mohammad Saifuzzaman in the experiment design and data collection. This research is partly funded by Australian Research Council grants DE160100449 and LP160101021.

References

.Idcommunity. 2016. *Service age groups | Australia | Community profile* [Online]. [Accessed 19 August 2019].

Ali, Y. 2020. *Investigation of lane-changing behaviour in a connected environment.* Ph.D. Dissertation, The University of Queensland.

Ali, Y., Bliemer, M., Zheng, Z. and Haque, M. M. 2020a. Comparing the usefulness of real-time driving aids in a connected environment during mandatory and discretionary lane-changing manoeuvres. *Transportation Research Part C,* 121**,** 102871.

Ali, Y., Bliemer, M., Zheng, Z. and Haque, M. M. 2020b. Cooperate or not? Exploring drivers' interactions and response times to a lane-changing request in a connected environment. *Transportation Research Part C,* 120**,** 102816.

Ali, Y., Haque, M., Zheng, Z., Washington, S. and Yildirimoglu, M. 2019a. A hazard-based duration model to quantify the impact of connected driving environment on safety during mandatory lane-changing. *Transportation Research Part C,* 106**,** 113-131.

Ali, Y., Sharma, A., Haque, M., Zheng, Z. and Saifuzzaman, M. 2020c. The impact of the connected environment on driving behavior and safety: A driving simulator study. *Accident Analysis and Prevention,* 144**,** 105643.

Ali, Y., Zheng, Z. and Haque, M. 2018. Connectivity's impact on mandatory lane-changing behaviour: evidences from a driving simulator study. *Transportation Research Part C,* 93, 292-309.

Ali, Y., Zheng, Z., Haque, M. and Wang, M. 2019b. A game theory-based approach for modelling mandatory lane-changing behaviour in a connected environment. *Transportation Research Part C,* 106, 220-242.

Ali, Y., Zheng, Z., Haque, M., Yildirimoglu, M. and Washington, S. 2020d. Detecting, analysing, and modelling failed lane-changing attempts in traditional and connected environments. *Analytic Methods in Accident Research,* 28, 100138.

Ali, Y., Zheng, Z., Haque, M., Yildirimoglu, M. and Washington, S. 2020e. Understanding the discretionary lane-changing behaviour in the connected environment. *Accident Analysis and Prevention,* 137, 105463.

Baguley, C. 1988. Running the red'at signals on high-speed roads. *Traffic Engineering and Control,* 29 (7-8), 415-420.

Brooks, J., Goodenough, R., Crisler, M., Klein, N., Alley, R., Koon, B., Logan Jr, W., Ogle, J., Tyrrell, R. and Wills, R. 2010. Simulator sickness during driving simulation studies. *Accident Analysis and Prevention,* 42 (3), 788-796.

Caird, J., Chisholm, S., Edwards, C. and Creaser, J. 2007. The effect of yellow light onset time on older and younger drivers' perception response time (PRT) and intersection behavior. *Transportation Research Part F,* 10 (5), 383-396.

Caird, J. K., Chisholm, S. and Lockhart, J. 2008. Do in-vehicle advanced signs enhance older and younger drivers' intersection performance? Driving simulation and eye movement results. *International journal of human-computer studies,* 66 (3), 132-144.

Chang, S., Lin, C., Hsu, C., Fung, C. and Hwang, J. 2009. The effect of a collision warning system on the driving performance of young drivers at intersections. *Transportation Research Part F,* 12 (5), 371-380.

Cheung, I. and Mccartt, A. 2011. Declines in fatal crashes of older drivers: Changes in crash risk and survivability. *Accident Analysis and Prevention,* 43 (3), 666-674.

Choi, E. 2010. Crash factors in intersection-related crashes: An on-scene perspective. National Highway Traffic Safety Administration, U.S. Department of Transportation.

Choudhary, P. and Velaga, N. 2019. Driver behaviour at the onset of yellow signal: a comparative study of distraction caused by use of a phone and a music player. *Transportation Research Part F,* 62, 135-148.

Dtmr 2019. 2018 Summary Road Crash Report, Queensland Road Fatalities. Customer Services, Safety & Regulation Division, Department of Transport and Main Roads, Brisbane, Australia: Queensland Transport.

Eker, U., Ahmed, S., Fountas, G. and Anastasopoulos, P. 2019. An exploratory investigation of public perceptions towards safety and security from the future use of flying cars in the United States. *Analytic Methods in Accident Research,* 23, 1-20.

Elmitiny, N., Yan, X., Radwan, E., Russo, C. and Nashar, D. 2010. Classification analysis of driver's stop/go decision and red-light running violation. *Accident Analysis and Prevention,* 42 (1), 101-111.

Eluru, N. and Yasmin, S. 2016. Disentangling the influence of cell phone usage in the dilemma zone: An econometric approach. *Accident Analysis and Prevention,* 96, 280-289.

Fountas, G. and Anastasopoulos, P. 2017. A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. *Analytic Methods in Accident Research,* 15, 1-16.

Fountas, G., Anastasopoulos, P. and Abdel-Aty, M. 2018a. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. *Analytic Methods in Accident Research,* 18, 57-68.

Fountas, G., Pantangi, S., Hulme, K. and Anastasopoulos, P. 2019. The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: a correlated grouped random parameters bivariate probit approach. *Analytic Methods in Accident Research,* 22**,** 1-15.

Fountas, G., Sarwar, M., Anastasopoulos, P., Blatt, A. and Majka, K. 2018b. Analysis of stationary and dynamic factors affecting highway accident occurrence: a dynamic correlated grouped random parameters binary logit approach. *Accident Analysis and Prevention,* 113**,** 330-340.

Greene, W. 2012. LIMDEP Version 10/NLOGIT Version 5. *Econometric Modeling Guide.*

Haque, M., Ohlhauser, A., Washington, S. and Boyle, L. 2016a. Decisions and actions of distracted drivers at the onset of yellow lights. *Accident Analysis and Prevention,* 96**,** 290-299.

Haque, M., Oviedo-Trespalacios, O., Debnath, A. and Washington, S. 2016b. Gap acceptance behavior of mobile phone–distracted drivers at roundabouts. *Transportation Research Record,* 2602 (1)**,** 43-51.

Huo, X., Leng, J., Hou, Q. and Yang, H. 2020. A Correlated Random Parameters Model with Heterogeneity in Means to Account for Unobserved Heterogeneity in Crash Frequency Analysis. *Transportation Research Record***,** 0361198120922212.

Iihs. 2020. *Red light running* [Online]. Available: https://www.iihs.org/topics/red-light-running [Accessed 16 September 2020].

Khatoon, M., Tiwari, G. and Chatterjee, N. 2013. Impact of grade separator on pedestrian risk taking behavior. *Accident Analysis & Prevention,* 50**,** 861-870.

Kramer, A. F., Cassavaugh, N., Horrey, W. J., Becic, E. and Mayhugh, J. L. 2007. Influence of age and proximity warning devices on collision avoidance in simulated driving. *Human factors,* 49 (5)**,** 935-949.

Lee, J. and Park, B. 2012. Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment. *IEEE Transactions on Intelligent Transportation Systems,* 13 (1)**,** 81-90.

Leung, S. and Starmer, G. 2005. Gap acceptance and risk-taking by young and mature drivers, both sober and alcohol-intoxicated, in a simulated driving task. *Accident Analysis and Prevention,* 37 (6)**,** 1056-1065.

Lu, G., Liu, M., Wang, Y., Wan, H. and Tian, D. 2015. Logit-based analysis of drivers' crossing behavior at unsignalized intersections in China. *Human Factors,* 57 (7)**,** 1101-1114.

Mahalel, D. and Prashker, J. 1987. A behavioral approach to risk estimation of rear-end collisions at signalized intersections. *Transportation Research Record,* 1114**,** 96-102.

Makishita, H. and Matsunaga, K. 2008. Differences of drivers' reaction times according to age and mental workload. *Accident Analysis and Prevention,* 40 (2)**,** 567-575.

Mannering, F. and Bhat, C. 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research,* 1**,** 1-22.

Mannering, F., Bhat, C., Shankar, V. and Abdel-Aty, M. 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research,* 25**,** 100113.

Mannering, F., Shankar, V. and Bhat, C. 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research,* 11**,** 1-16.

Montgomery, J., Kusano, K. and Gabler, H. 2014. Age and gender differences in time to collision at braking from the 100-car naturalistic driving study. *Traffic Injury Prevention,* 15 (sup1)**,** S15-S20.

Newton, C., Mussa, R., Sadalla, E., Burns, E. and Matthias, J. 1997. Evaluation of an alternative traffic light change anticipation system. *Accident Analysis & Prevention,* 29 (2)**,** 201-209.

Nhtsa 2006. Traffic safety facts. *National Highway Traffic Safety Administration.*

Njobelo, G., Sando, T., Sajjadi, S., Mtoi, E., Ozguven, E. and Sobanjo, J. 2018. Safety evaluation of the advanced stop assist system in connected vehicle environment. *Transportation Research Record,* 2672 (22)**,** 47-57.

Pantangi, S., Fountas, G., Sarwar, M., Anastasopoulos, P., Blatt, A., Majka, K., Pierowicz, J. and Mohan, S. 2019. A preliminary investigation of the effectiveness of high visibility enforcement programs using naturalistic driving study data: a grouped random parameters approach. *Analytic Methods in Accident Research,* 21**,** 1-12.

Papaioannou, P. 2007. Driver behaviour, dilemma zone and safety effects at urban signalised intersections in Greece. *Accident Analysis and Prevention,* 39 (1)**,** 147-158.

Porter, B. and England, K. 2000. Predicting red-light running behavior: a traffic safety study in three urban settings. *Journal of Safety Research,* 31 (1)**,** 1-8.

Preusser, D. F., Williams, A. F., Ferguson, S. A., Ulmer, R. G. and Weinstein, H. B. 1998. Fatal crash risk for older drivers at intersections. *Accident Analysis & Prevention,* 30 (2)**,** 151-159.

Ramotowski, M. and Fitzgerald, R. 2020. Chi-Squared Automatic Inference Detection (CHAID) decision tree. *Apache Software License,* Version 5.3.0.

Retting, R., Chapline, J. and Williams, A. 2002. Changes in crash risk following re-timing of traffic signal change intervals. *Accident Analysis and Prevention,* 34 (2)**,** 215-220.

Revelt, D. and Train, K. 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics,* 80 (4)**,** 647-657.

Sam, D., Evangelin, E. and Raj, V. 2015. Improving road safety for pedestrians in black spots using a hybrid vanet of vehicular sensors and pedestrian body unit. *ARPN Journal of Engeering and Applied Sciences,* 10**,** 4639-4644.

Sharma, A., Ali, Y., Saifuzzaman, M., Zheng, Z. and Haque, M. Human factors in modelling mixed traffic of traditional, connected, and automated vehicles. International Conference on Applied Human Factors and Ergonomics, 2017. Springer, 262-273.

Sharma, A., Zheng, Z., Kim, J., Bhaskar, A. and Haque, M. 2019. Estimating and Comparing Response Times in Traditional and Connected Environments. *Transportation Research Record***,** 0361198119837964.

Sharma, A., Zheng, Z., Kim, J., Bhaskar, A. and Haque, M. 2020a. Is an informed driver a better decision maker? A grouped random parameters with heterogeneity-in-means approach to investigate the impact of the connected environment on driving behaviour in safety-critical situations. *Analytic Methods in Accident Research***,** 1-24.

Sharma, A., Zheng, Z., Kim, J., Bhaskar, A. and Haque, M. 2020b. Is an informed driver a better decision maker? A grouped random parameters with heterogeneity-in-means approach to investigate the impact of the connected environment on driving behaviour in safety-critical situations. *Analytic Methods in Accident Research***,** 100127.

Sheffi, Y. and Mahmassani, H. 1981. A model of driver behavior at high speed signalized intersections. *Transportation Science,* 15 (1)**,** 50-61.

Tfnsw. 2019. *Centre for Road Safety* [Online]. Transport for New South Wales, Australia https://roadsafety.transport.nsw.gov.au/statistics/interactivecrashstats/nsw.html?tabnsw=2 [Accessed 26 November 2019].

Thompson, K., Johnson, A., Emerson, J., Dawson, J., Boer, E. and Rizzo, M. 2012. Distracted driving in elderly and middle-aged drivers. *Accident Analysis and Prevention,* 45**,** 711-717.

Tränkle, U., Gelau, C. and Metker, T. 1990. Risk perception and age-specific accidents of young drivers. *Accident Analysis and Prevention,* 22 (2)**,** 119-125.

Washington, S., Karlaftis, M., Mannering, F. and Anastasopoulos, P. 2020. *Statistical and econometric methods for transportation data analysis,* CRC press.

Xiang, W., Yan, X., Weng, J. and Li, X. 2016. Effect of auditory in-vehicle warning information on drivers' brake response time to red-light running vehicles during collision avoidance. *Transportation Research Part F,* 40**,** 56-67.

Xiong, H., Narayanaswamy, P., Bao, S., Flannagan, C. and Sayer, J. 2016. How do drivers behave during indecision zone maneuvers? *Accident Analysis and Prevention,* 96**,** 274-279.

Yang, C. D. and Najm, W. 2006. *Analysis of red light violation data collected from intersections equipped with red light photo enforcement cameras*, US Department of Transportation, National Highway Traffic Safety Administration.

Zhang, J., Fraser, S., Lindsay, J., Clarke, K. and Mao, Y. 1998. Age-specific patterns of factors related to fatal motor vehicle traffic crashes: focus on young and elderly drivers. *Public health*, 112 (5)**,** 289-295.