Accident Analysis and Prevention 96 (2016) 316–328

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Red-light running violation prediction using observational and simulator data

Arash Jahangiri [a], Hesham Rakha [a,*], Thomas A. Dingus [b]

[a] Center for Sustainable Mobility, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States
[b] Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States

ARTICLE INFO

*Article history:*
Received 18 May 2015
Received in revised form 9 March 2016
Accepted 14 June 2016
Available online 29 June 2016

*Keywords:*
Driver violation
Red-light running
Signalized intersection
Violation prediction
Observational data
Simulator data
Random forest
Machine learning

ABSTRACT

In the United States, 683 people were killed and an estimated 133,000 were injured in crashes due to running red lights in 2012. To help prevent/mitigate crashes caused by running red lights, these violations need to be identified before they occur, so both the road users (i.e., drivers, pedestrians, etc.) in potential danger and the infrastructure can be notified and actions can be taken accordingly. Two different data sets were used to assess the feasibility of developing red-light running (RLR) violation prediction models: (1) observational data and (2) driver simulator data. Both data sets included common factors, such as time to intersection (*TTI*), distance to intersection (*DTI*), and velocity at the onset of the yellow indication. However, the observational data set provided additional factors that the simulator data set did not, and vice versa. The observational data included vehicle information (e.g., speed, acceleration, etc.) for several different time frames. For each vehicle approaching an intersection in the observational data set, required data were extracted from several time frames as the vehicle drew closer to the intersection. However, since the observational data were inherently anonymous, driver factors such as age and gender were unavailable in the observational data set. Conversely, the simulator data set contained age and gender. In addition, the simulator data included a secondary (non-driving) task factor and a treatment factor (i.e., incoming/outgoing calls while driving). The simulator data only included vehicle information for certain time frames (e.g., yellow onset); the data did not provide vehicle information for several different time frames while vehicles were approaching an intersection. In this study, the random forest (RF) machine-learning technique was adopted to develop RLR violation prediction models. Factor importance was obtained for different models and different data sets to show how differently the factors influence the performance of each model. A sensitivity analysis showed that the factor importance to identify RLR violations changed when data from different time frames were used to develop the prediction models. *TTI*, *DTI*, the required deceleration parameter (*RDP*), and velocity at the onset of a yellow indication were among the most important factors identified by both models constructed using observational data and simulator data. Furthermore, in addition to the factors obtained from a point in time (i.e., yellow onset), valuable information suitable for RLR violation prediction was obtained from defined monitoring periods. It was found that period lengths of 2–6 m contributed to the best model performance.



## 1. Introduction

According to National Highway Traffic Safety Administration (NHTSA) report, during 2012, more than 2.5 million intersection-related crashes occurred in the United States, of which 2850 were fatal crashes and 680,000 were injurious crashes (NHTSA, 2014). Specifically, statistics demonstrate that a large number of crashes occur at signalized intersections due to traffic violations, of which running red lights has been reported to be a serious issue. According to the Insurance Institute for Highway Safety (IIHS), 683 people were killed and an estimated 133,000 were injured in crashes in the United States during 2012 due to running red lights (IIHS, 2012). The AAA Foundation for Traffic Safety surveyed 2000 United States residents aged 16 and older. The survey showed that approximately 93% of drivers believe that running through a red light is unacceptable if it is possible to stop safely. However, one-third mentioned they ran through a red light during the past 30 days. This shows that, although drivers are generally aware of the dangers of this type of violation, they are likely to occasionally run a red light (IIHS, 2010).

* Corresponding author.
*E-mail addresses:* ArashJ@vt.edu (A. Jahangiri), HRakha@vtti.vt.edu (H. Rakha), TDingus@vtti.vt.edu (T.A. Dingus).

http://dx.doi.org/10.1016/j.aap.2016.06.009
0001-4575/

### 1.1. Dilemma zone and influential factors

Drivers approaching signalized intersections need to make a decision whether to stop or proceed when the traffic signal changes from green to yellow. If they decide to proceed and the signal turns red before the driver passes the stop bar (or before clearing the intersection), the driver is considered a red-light violator. Violating a red light may lead to crashes with the side-street traffic. In another scenario, if the driver abruptly stops at the onset of a yellow light while a following vehicle makes a conflicting decision (i.e., decision to proceed), rear-end crashes may occur. The area in which drivers decide what action to take when the traffic signal changes from green to yellow is known as the dilemma zone, which can be defined in space or time (Rakha et al., 2007). The dilemma zone is a classic problem first introduced by Gazis et al. (1960), followed by numerous studies (Zegeer and Deen, 1978; Sheffi and Mahmassani, 1981; Chang et al., 1985; Bonneson et al., 2002; Pant et al., 2005; Gates et al., 2007; Liu et al., 2007; Rakha et al., 2007; Wei et al., 2011; Ghanipoor Machiani and Abbas, 2014a,b, 2015a,b).

When predicting red-light running (RLR) violations, factors that influence driver behavior in the dilemma zone need to be taken into account. A number of studies have attempted to investigate factors that affect driver behavior when approaching signalized intersections. These factors influence the driver's decision to stop or proceed when facing a yellow light and, consequently, have an impact on the probability of rear-end and right-angle crashes. A summary of these factors are listed in Table 1. A comprehensive list of influential factors that have been investigated throughout relevant literature can be found in a study conducted by Abbas et al. (2014).

### 1.2. Data collection methods

Data collection methods restrict factors that can be considered during analyses. For example, using driving simulators as applied in many studies (Mussa et al., 1996; Caird et al., 2007; Boyle and Lee, 2010; Peng et al., 2013; Ghanipoor Machiani and Abbas, 2014a,b, 2015a,b; Haque et al., 2015) facilitates the examination of many factors (e.g. age, gender, presence of police, work zone, distraction, cell phone use, etc.). However, the behavior of the drivers in a simulator may not reflect their natural behaviors when driving in real-world conditions. Using observational data collection methods (i.e., through video cameras), such as (Gates et al., 2007; Liu et al., 2007; Wei et al., 2011), the drivers' natural behaviors are captured as the drivers are not aware that their data are being collected. Also, when using a naturalistic data collection method (i.e. having participants drive/ride instrumented vehicles/bicycles) (Dingus et al., 2006), the drivers' natural behaviors can be captured as the drivers become quickly accustomed to the instrumentation and drive as they normally would. Nevertheless, factors such as age and gender cannot be captured in observational (i.e., video recording at infrastructure) studies due to the inherent anonymity of observational data collection methods. Also, other specific environmental factors, such as the presence of bicycles or police, may not be available in naturalistic or observational data, whereas these scenarios can be easily developed using driving simulators. Another data collection method involves using an experimental test track and running scenarios in a controlled environment. This data collection method essentially combines the capabilities of observational (or naturalistic) and simulator data collection methods. That is, the behavior of the drivers in a test-track environment is more natural than their behavior in a driving simulator. However, their behavior may be affected because the participants know they are in an experiment and may adjust their behavior. A test-track environment also allows specific scenarios to be run, whereas observational or naturalistic data are limited to what scenarios occur in real-world driving. Studies in which data were collected in a controlled field environment include (El-Shawarby et al., 2007; Rakha et al., 2007; Amer et al., 2011a,b; Li et al., 2012).

Ideally, naturalistic or observational data (if collected over a sufficient time period) should be used to test and validate an RLR violation prediction model because natural driver behavior occurring in the real world can only be observed in naturalistic or observational data. Nevertheless, demographic information such as age and gender is not available through observational data because usually the distance from the camera and the drivers is too far to obtain such information. Also, naturalistic data collection should be conducted in a long period of time to collect enough information, which makes is expensive compared to other methods. For example, in order to observe a specific scenario (e.g. texting while driving), data should be collected for several weeks or months for an individual to see that specific scenario, and even after a long period of time you may find some drivers who never text while driving. Furthermore, in some scenarios, examination of factors such as distraction and cell phone use that affect driver behavior at the onset of a yellow light may be difficult or impossible. When observational or naturalistic data collection is utilized, the scenario of interest may not occur as frequently as desired for proper statistical analysis. Also, when test track experiments are considered, having such scenarios would place human subjects in danger. Hence, driving simulators can be adopted. Driving simulators are also considered vital for assessing factors related to new vehicle technologies (Boyle and Lee, 2010). Using driving simulators provide more control in general; therefore many different factors (e.g. gender, age, presence of police, presence of other drivers, presence of pedestrians/cyclists at the intersection, cell phone use) can be included in a single scenario.

### 1.3. Study focus and objectives

Some studies concentrate on investigating the characteristics of red-light violators and conditions in which drivers are more or less likely to violate (Retting and Williams, 1996; Porter and England, 2000; Porter and Berry, 2001; Bonneson and Son, 2003; Neale and McGhee, 2006; Retting et al., 2008). For example, drivers who do not use safety belts and non-Caucasian drivers were more likely to violate the red light. Moreover, larger intersections and higher traffic volumes were associated with higher RLR violation rates (Porter and England, 2000). Several studies developed models to estimate the frequency of red-light violators. For instance, Bonneson and Son (2003) developed a regression model using several factors, such as flow rate, cycle length, and yellow duration. Also, a topic closely related to the present study is predicting driver decision (i.e., stop or proceed) when the traffic signal turns from green to yellow; this topic was the focus of such studies as (Elhenawy et al., 2014; Ghanipoor Machiani and Abbas, 2014a,b). However, not all driver decisions to proceed lead to red-light violations (i.e., passing the intersection during yellow time). Statistical and probabilistic approaches have also been applied (Sheffi and Mahmassani, 1981; Zhang et al., 2009). For example, Zhang et al. (2009) developed a probabilistic model to predict RLR violations, taking into consideration minimizing both false alarm rates and missing errors.

The underlying assumption is that vehicle data such as speed and acceleration can be obtained as the vehicle is approaching an intersection. These kinds of data are obtainable through video processing (camera recordings at intersections) or through connected vehicle technology. Hence, the model developed in this study can be applied in an infrastructure-based safety system to predict red light running behavior. Consequently, the system would take appropriate action to minimize crashes due to RLR violations.

The present study focuses on developing prediction models aiming at identifying RLR violations using two different data sets (i.e.,

**Table 1**
Influential factors affecting driver behavior when approaching signalized intersections.

| Factor name | Studies |
|---|---|
| Perception-reaction time | Caird et al. (2007), Gates et al. (2007), Liu et al. (2007), Rakha et al. (2007), Amer et al. (2011a,b), Wei et al. (2011), Li et al. (2012) |
| Acceleration/deceleration rate | Caird et al. (2007), Gates et al. (2007), Liu et al. (2007), Wei et al. (2011), Aoude et al. (2012), Li et al. (2012), Jahangiri et al. (2015a,b) |
| Age | Caird et al. (2007), El-Shawarby et al. (2007), Rakha et al. (2007), Amer et al. (2011a,b), Liu et al. (2011), Li et al. (2012) |
| Gender | Caird et al. (2007), El-Shawarby et al. (2007), Rakha et al. (2007), Amer et al. (2011a,b), Liu et al. (2011), Li et al. (2012) |
| Time to intersection (*TTI*) at the onset of yellow | Caird et al. (2007), Rakha et al. (2007), Aoude et al. (2012), Li et al. (2012), Ghanipoor Machiani and Abbas (2015a,b), Jahangiri et al. (2015a,b) |
| Distance to intersection (*DTI*) at the onset of yellow | Gates et al. (2007), Aoude et al. (2012), Li et al. (2012), Haque et al. (2015), Jahangiri et al. (2015a,b) |
| Approach speed | Gates et al. (2007), Liu et al. (2007), Amer et al. (2011a,b), Liu et al. (2011), Wei et al. (2011), Aoude et al. (2012), Haque et al. (2015), Jahangiri et al. (2015a,b) |
| Vehicle type | Gates et al. (2007), Liu et al. (2011), Elhenawy et al. (2015) |
| Presence of side-street vehicles, pedestrians, bicycles, or opposing vehicles waiting to turn left | Gates et al. (2007), Ghanipoor Machiani and Abbas (2014a,b) |
| Flow rate | Gates et al. (2007), Liu et al. (2011) |
| Length of yellow interval | Gates et al. (2007), Li et al. (2012) |
| Cycle length | Gates et al. (2007), Liu et al. (2011) |
| Presence of police | Ghanipoor Machiani and Abbas (2014a,b, 2015a,b) |
| Pavement and weather conditions (wet, rainy) | Li et al. (2012), Elhenawy et al. (2015) |
| Cell phone use | Liu et al. (2011), Haque et al. (2015) |
| Speed limit | Gates et al. (2007), Amer et al. (2011a,b) |
| Roadway grade | Amer et al. (2011a,b), Li et al. (2012) |
| Driver aggressiveness | Elhenawy et al. (2015) |
| Driver's learning through different signal settings | Ghanipoor Machiani and Abbas (2014a,b) |
| Required deceleration parameter (RDP) at onset of yellow | Doerzaph et al. (2010), Aoude et al. (2012), Jahangiri et al. (2015a,b) |
| Time lost/gained | Ghanipoor Machiani and Abbas (2015a,b) |

observational data set and driving simulator data set). The goal is to infer driver behavior as individuals approach an intersection. Random forest (RF) as an artificial intelligent (AI) technique was employed to construct RLR violation prediction models. AI techniques have been adopted to solve many problems in the transportation domain, such as real-time detection of driver cognitive distraction (Liang et al., 2007), lane detection and tracking (Kim, 2008), transportation mode recognition (Jahangiri and Rakha, 2015), traffic sign detection (Balali and Golparvar-Fard, 2014), and incident detection (Yuan and Cheu, 2003). These studies illustrate that applying AI methods can lead to promising results. However, studies that apply AI techniques to develop RLR violation prediction models are limited, thus more research is needed to provide insights about modelling driver behavior at intersections using such algorithms. In particular, the RF method has not been applied to predict RLR behavior to the best of our knowledge. Since RF has shown in the realm of AI to perform greatly, it was selected in the present research to investigate the benefits in predicting RLR violations. When using AI methods, different terms may be used to refer to the factors that are employed to build models, such as "factors," "features," "variables," "attributes," and "predictors." These terms may be used interchangeably. However, the word "factor" is used throughout this paper for consistency and clarity.

The objectives of the present study are as follows: (1) Create several factors in model development and use a selection method to determine the most useful factors. (2) Conduct a sensitivity analysis to determine an appropriate monitoring period corresponding to the onset of a yellow light to capture the information that reflects drivers' decisions. (3) Investigate how the RF method can predict RLR violations using different monitoring periods while providing enough time for endangered drivers and/or the infrastructure to respond. (4) Use observational data and driving simulator data to develop prediction models. (5) Identify important factors in predicting RLR violations.

Relevant work is reviewed in the second section of this paper, while the third section describes the observational data and the simulator data used. The fourth section explains the model development, which includes the RF method that was employed, the selected monitoring period, and the adopted factor selection method. The results are presented in the fifth section. Finally, the conclusion is provided in the sixth section.

## 2. Relevant work

Elmitiny et al. (2010) developed a classification tree model to predict RLR violations and found the most important factors to be the vehicle distance at the onset of a yellow light, the operating speed at the onset of a yellow light, and position in traffic flow. However, only a limited number of factors were examined, thus Elmitiny et al. (2010) did not use any factor selection method to determine the most useful factor in predicting RLR violations. To examine several factors, a proper selection becomes critical towards using the most relevant factors. Moreover, the factors examined and found to be important by Elmitiny et al. (2010) were obtained from an instant in time (i.e., the onset of a yellow light). Although the yellow onset is an important instant (i.e., because it is the moment the drivers encounter the yellow and, consequently, must decide whether to stop or proceed), the drivers' decisions are not made at that instant but, rather, are made during a short time period. Therefore, other factors that can explain drivers' behaviors in a time period immediately following the yellow onset should also be examined when predicting RLR violations. The current study makes

an effort to first construct such factors and then use a selection method to identify the most useful factors.

Aoude et al. (2012) applied the support vector machine (SVM) and hidden Markov model (HMM) to build RLR violation prediction models. They showed how their models outperform traditional methods of prediction. Similarly to Elmitiny et al. (2010), they did not use any factor selection method. Using SVM, Aoude et al. (2012) found by experimenting on different combinations that three factors, namely *DTI*, speed, and acceleration, led to the best result (i.e., lowest error). Similarly, when applying the HMM method, they tested different combinations of factors to find the ones that led to the lowest error, namely *DTI*, speed, acceleration, *TTI*, and the *RDP*.

Aoude et al. (2012) considered a point in time after which the prediction becomes invalid because not enough time is available for a driver in a potential collision to react. Their logic was that, if a safety system employs any prediction model, the prediction task needs to be conducted before a certain period to provide sufficient time for endangered drivers to respond. Furthermore, they considered a monitoring period during which the required factors were extracted for model development. This period was chosen right before the aforementioned critical point in time. They determined this critical point based on two criteria, whichever was the first to occur: (1) Minimum time threshold. Three different values were selected based on a human response time distribution as discussed by McLaughlin et al. (2008): 1, 1.6, and 2 s, which are sufficient for 45%, 80%, and 90% of the population to respond, respectively; and (2) Minimum distance threshold. If the vehicles' approaching speeds were very low and vehicles drew too close to the intersection, it is unlikely that the minimum time threshold criterion meets, thus a minimum *DTI* was assumed. However, the minimum time threshold is not enough to avoid a possible collision since it only corresponds to the driver response time without considering the vehicle response time. Consequently, to avoid a potential collision, two time periods were taken into account in the present research: (1) The driver response time and (2) The vehicle response time.

Aoude et al. (2012) did not use any factors that reflect the interaction between the drivers and the signal setting. For example, it appears that factors incorporated into the Aoude et al. (2012) SVM model, such as *DTI* and speed, were extracted from the monitoring period that they defined without knowing the yellow light onset (i.e., the yellow light may start before or after their defined monitoring period). In the present paper, the yellow onset and the monitoring period immediately after onset were both accounted for as such time periods contain the information reflecting the drivers' decisions when encountering a yellow light.

Zhang et al. (2009) and Zhang et al. (2012) proposed a probabilistic framework to identify RLR violations, which was adopted in dynamic all-red extension (DARE) as an intersection collision avoidance method. Their prediction model was based on vehicle speed measurements from a minimum set of two point sensors and their corresponding event time stamps. They applied Neyman–Pearson (NP) criterion to maximize the probability of prediction while keeping the false-alarm rate equal to or lower than a given level.

## 3. Data description

### 3.1. Observational data

The observational data used in this research were derived from the Cooperative Intersection Collision Avoidance Systems for Violations (CICAS-V) project. As part of the CICAS-V project, different equipment, such as radars, video cameras, and signal phase sniffers, were included in the data acquisition systems (DASs) designed and developed by the Center for Technology Development (CTD) at the Virginia Tech Transportation Institute (VTTI). The DAS units were installed at six stop-controlled intersections and three signalized intersections in the New River Valley area of Southwest Virginia. A data sample from one of the signalized intersections (i.e., the intersection of Franklin Street and Depot Street) was used in the present study. Doerzaph and Neale (2010) provide additional details about the data collection of CICAS-V. The sample used in this study includes approximately 500 observations for the RLR violation behavior and 500 observations reflecting compliant behavior. RLR violations are defined based on the location of the vehicles when the traffic light changes to red. According to National Cooperative Highway Research Program (NCHRP) report, there are two definitions for RLR violations: *"under 'permissive' yellow law, drivers may enter the intersection during the entire duration of the yellow change interval and legally be in the intersection while the red signal indication is displayed, so long as entrance occurred before or during the yellow signal indication. Under the "restrictive" yellow law, (1) drivers may not enter the intersection during the yellow signal indication unless it can be entirely cleared prior to the onset of the red signal indication, or (2) drivers may not enter the intersection unless it is impossible or unsafe to stop."* (NCHRP, 2012). In the present research, it was assumed that the "permissive" yellow law is followed. That is, drivers were identified as RLR violators if the light was red the moment they crossed the stop line. For each individual vehicle approaching an intersection, data such as speed, acceleration,*DTI*, and signal setting information were collected at high resolution. Table 2 presents the factors used in this study.

**Table 2**
Factors used from the observational data (CICAS-V).

| No. | Factor |
|---|---|
| 1 | *DTI* |
| 2 | *TTI* |
| 3 | *DTI* at onset of yellow |
| 4 | *TTI* at onset of yellow |
| 5 | *RDP* at onset of yellow |
| 6 | Velocity at onset of yellow |
| 7 | Acceleration at onset of yellow |
| 8 | Vehicle speed |
| 9 | Vehicle acceleration |

### 3.2. Driving simulator data

The data set from a driving simulator study provided for a data contest at the 93rd Annual Meeting of the Transportation Research Board was obtained through the *Journal of Accident Analysis & Prevention* website. The data contain several factors, of which a subset was used in this study, as shown in Table 3.

According to the provided data description, the practice runs denoted as "FAMILIAR" were omitted from the data set. The RLR violations were identified as follows: Looking at the "Distance from stop line when the vehicle comes to stop" (factor 6), drivers who did not stop and those who stopped beyond the stop line were extracted first. Subsequently, considering only this subset of observations, those with a frame number at the stop line (factor 11) greater than the frame number at yellow to red (factor 5) represent violators. To summarize, violators in the simulator data are those who passed the stop line when the light was red, no matter if they passed the intersection completely or stopped just beyond the stop line.

**Table 3**
Factors used from the simulator data.

| No. | Factor |
|---|---|
| 1 | Age group: young (18–25), middle-aged (30–45), and older (50–60) |
| 2 | Gender |
| 3 | The secondary (or non-driving) task condition:<br>• Using handheld wireless for dialing and conversing.<br>• Using hands-free wireless for voice dialing using digits, and hands-free using external speaker kit for conversing.<br>• Using the phone for voice dialing using digits, and the headset for hands-free conversing. |
| 4 | Treatment: Baseline (no call), Outgoing call, and Incoming call |
| 5 | Frame number at which the traffic light changes from Yellow to Red |
| 6 | Distance from stop line when the vehicle comes to stop. Notes: Drivers who stopped beyond the stop bar as well as the drivers who did not stop were also coded. |
| 11 | The frame number for when the participant had an accelerator pedal change of greater than 10% percent. |
| 12 | Acceleration Pedal Change Direction: −1 = released, 1 = depressed |
| 7 | Max deceleration between the 10% increase in Acceleration Pedal and when driver goes past intersection. |
| 8 | Max acceleration between the 10% increase in Acceleration Pedal and when driver goes past intersection. |
| 9 | Velocity at onset of yellow. |
| 10 | Distance from stop line at onset of yellow. |
| 11 | Frame number when participant reached stop line. |

## 4. Model development

### 4.1. RF method

A machine-learning method, namely RF, was employed in the present research to predict RLR violations. RF was used as it offers several advantages (Breiman 2001; Liaw and Wiener, 2002). Namely, the model performance is as good as (and sometimes better than) other powerful methods, such as Adaboost, discriminant analysis, SVMs, and neural networks. RF is robust against overfitting and is relatively robust to outliers and noise. It is faster than bagging or boosting, it provides useful internal error estimates known as out-of-bag (OOB) errors, and it provides factor importance. In addition, RF is easy to tune and requires only two parameters, thus it can be simply optimized. The RF method, as proposed by Breiman (2001), is an ensemble learning approach based on predictions of a number of decision trees. Instead of a single decision tree, RF uses a group of decision trees from which a majority vote makes the predictions. In a similar method, called Bagging, each tree uses all the available variables. Each tree in the RF method, as mentioned above, only uses a few variables to minimize the correlations amongst different trees. Two model parameters, namely the number of decision trees and the number of factors to use in each tree, should be determined to apply the RF method. To construct each tree, the recursive binary splitting method was adopted in which factors are selected to divide the data into different parts. Different criteria may be used to determine how to separate the data. The Gini index criterion was used in this study, as presented in Eq. (1). It should be noted that the Gini index and cross-entropy criteria act similarly and are both recommended approaches (Hastie et al., 2009).

$$G = \sum_{k=1}^{K} P_k^m \left(1 - P_k^m\right) \quad (1)$$

Where,

$$P_k^m = \frac{1}{N^m} \sum_{x_i^m} I\left(y_i^m = k\right)$$

$P_k^m$ Proportion of class $k$ observations in node $m$

$N^m$ Number of observations received at node $m$

$y_i^m$ The response value corresponding to the observation $i$ at node $m$

$x_i^m$ The feature vector corresponding to the observation $i$ at node $m$

$k$ class.

### 4.2. Monitoring period

Fig. 1 illustrates two vehicles approaching a signalized intersection: (1) A violator vehicle denoted by $v$, which is presumably going to violate the red light; and (2) An endangered vehicle denoted by $e$, which is going to be at risk of a right-angle crash due to the RLR violation of vehicle $v$. In previous work (Jahangiri et al., 2015a,b), fixed monitoring periods were determined to extract the information reflecting driver behavior when approaching a signalized intersection. This fixed period was defined in both time (i.e., based on *TTI*; Jahangiri et al., 2015a,b) and in space (i.e., based on *DTI*; Jahangiri et al., 2015a,b). Consequently, prediction models were developed based on the information obtained from those periods. In the present study, the monitoring period was defined in space similar to (Jahangiri et al., 2015a,b), but different period lengths were examined instead of a single period. *DTI* was used here instead of *TTI* due to convenience. That is, identifying the monitoring period based only on the *TTI* becomes difficult, especially when speeds are very low that lead to very high *TTI* values. Moreover, using *DTI* makes it easier to conduct a sensitivity analysis with different monitoring periods. The monitoring period for each observation was defined depending on the *DTI* of each driver at the yellow onset instead of defining a fixed period for all drivers, as was performed in a previous study (Jahangiri et al., 2015a,b). It should be noted that the monitoring period was only applied when using the observational data. Although simulators in general can record data at 60 Hz, the particular simulator data used in this study did not include factors (e.g., speed, acceleration) for all data frames. Hence, it was not possible to define a monitoring period in the simulator data.

The monitoring period ($t_{mon}^v$) was defined by its start and end points based on *DTI* values, as illustrated in Fig. 1. For example, selecting *DTI* for start and end points as 40 and 25 m, respectively, lead to a monitoring period of 15 m. The start point of the monitoring period should not be selected too early to exclude unnecessary information (i.e., when the drivers are very far from the intersection, their behavior may not reflect their decision to violate the red light). The behavior of the drivers before the yellow onset may not be related to their decision to stop or proceed. Therefore, *DTI* at the yellow onset was selected as the start point for each observation. The end point was restricted in previous work to a point (i.e., $t_{min}^v$ as shown in Fig. 1) that provides sufficient time for the endangered vehicle to respond if a possible collision is predicted; $t_{min}^v$ is equivalent to the sum of two terms, namely the time required for the endangered driver to respond ($t_{driver}^e$) and the time required for the endangered vehicle to stop ($t_{vehicle}^e$). As a result, the sum of the two terms ($t_{driver}^e$ and $t_{vehicle}^e$) dictated the end point of the $t_{mon}^v$ (Jahangiri et al., 2015a,b). In the present study, however, a sensitivity analysis was conducted to investigate the effects of different monitoring periods on predicting RLR violations. The sensitivity analysis was performed to show how accurately the prediction models can perform with different monitoring periods.

Crashes resulting from an RLR violation can occur in two scenarios: (1) When no vehicle is waiting at the intersection, the

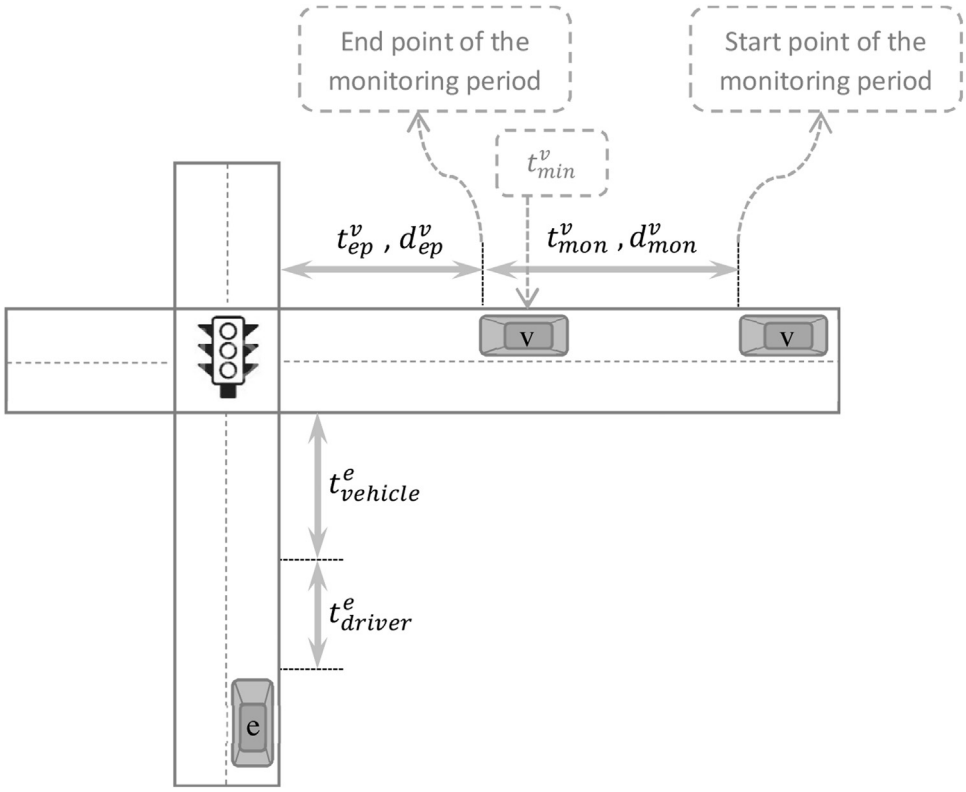


**Fig. 1.** Determination of the monitoring period; vehicle v as the RLR violator and vehicle e as the endangered vehicle.

endangered vehicle is approaching the intersection, and the traffic light turns green before the endangered vehicle needs to stop. This scenario is illustrated in Fig. 1 and was the focus of previous work with a single fixed monitoring period (Jahangiri et al., 2015a,b); and (2) When the endangered vehicle is already at the intersection waiting for the light to turn green. This vehicle is in front of a possible waiting queue and is the first vehicle to enter the intersection as soon as the light turns green. The present study focuses on both scenarios with consideration of different monitoring periods. For the first scenario, if enough time is available, the endangered driver can be notified so he or she can avoid a possible collision. However, if there is not enough time for the driver to react, or if the second scenario occurs, the infrastructure can be notified so appropriate action can be taken (e.g., extending a red clearance interval such as what Zhang et al. (2012) proposed). Therefore, even at the end of the yellow time, if the violation is predicted, the algorithm can extend the red clearance interval to avoid potential a collision.

**Table 4**
Observational data – list of examined factors.

| No. | Factor |
|---|---|
| 1 | *DTI* at onset of yellow |
| 2 | Velocity at onset of yellow |
| 3 | Acceleration at onset of yellow |
| 4 | *TTI* at onset of yellow |
| 5 | *RDP* at onset of yellow |
| 6 | *mean* (*Velocity*) over the $t^v_{mon}$ |
| 7 | *range* (*Velocity*) over the $t^v_{mon}$ |
| 8 | *max* (*Velocity*) over the $t^v_{mon}$ |
| 9 | *min* (*Velocity*) over the $t^v_{mon}$ |
| 10 | *std* (*Velocity*) over the $t^v_{mon}$ |
| 11 | *mean* (*Acceleraiton*) over the $t^v_{mon}$ |
| 12 | *range* (*Acceleraiton*) over the $t^v_{mon}$ |
| 13 | *max* (*Acceleraiton*) over the $t^v_{mon}$ |
| 14 | *min* (*Acceleraiton*) over the $t^v_{mon}$ |
| 15 | *std* (*Acceleraiton*) over the $t^v_{mon}$ |
| 16 | *mean* (*DTI*) over the $t^v_{mon}$ |
| 17 | *mean* (*TTI*) over the $t^v_{mon}$ |

### 4.3. Factor creation

In addition to the available factors in both observational and simulator data sets, others were created as follows:

#### 4.3.1. Using the observational data

Using the factors from observational data, as presented in Table 2, additional factors were created for examination. Table 4 lists all of the factors that were used for model development. The vehicle information for different time frames included velocity, acceleration, TTI, and DTI. As discussed in the previous section, a monitoring period to be examined for each vehicle contains the information between the start and end points of the period. To describe changes in the defined monitoring period, the additional factors were created using mostly statistical measures of dispersion (e.g., maximum, minimum, and standard deviation). The idea was that any change in driver behavior due to the drivers' decisions to stop or proceed can directly affect the kinetic information within the monitoring period. It should be noted that the RDP is the deceleration value required for a vehicle to be able to stop at the stop bar. The RDP was obtained through Eq. (2) as follows (Doerzaph et al., 2010).

$$RDP = \frac{V^2}{2.DTI \times g} \quad (2)$$

Where,

$V$ Vehicle's instantaneous velocity
$DTI$ Distance to intersection
$g$ Gravitational constant.

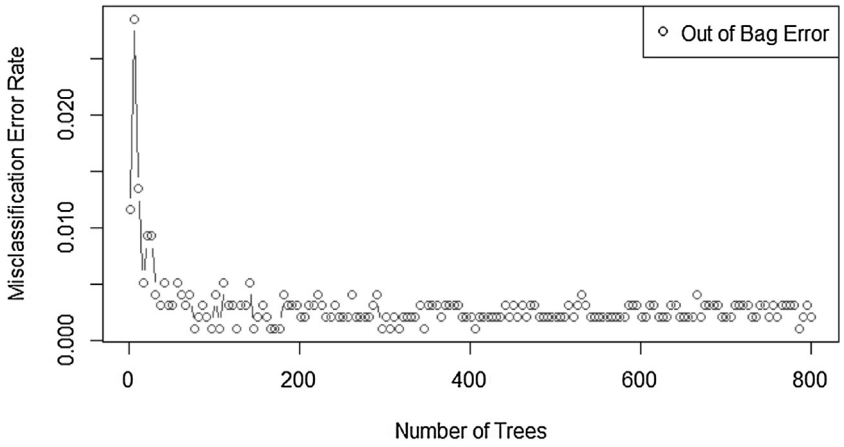


**Fig. 2.** Observational data – selecting the required number of trees.

**Table 5**
Simulator data – list of the examined factors.

| No. | Factor |
|---|---|
| 1 | Gender |
| 2 | Age |
| 3 | Acceleration Pedal Changed 10% |
| 4 | Acceleration Pedal Change Direction |
| 5 | Max deceleration |
| 6 | Max Acceleration |
| 7 | Velocity at onset of yellow |
| 8 | DTI at onset of yellow |
| 9 | The secondary task condition |
| 10 | Treatment |
| 11 | TTI at onset of yellow |
| 12 | RDP at onset of yellow |

#### 4.3.2. Using the simulator data

The additional factors created in the simulator data included *TTI* at the onset of a yellow light and the *RDP* at the onset of a yellow light. Total factors examined from this data set are shown in Table 5. As mentioned earlier, the data points within a defined monitoring period were not available for this particular data set. Therefore, unlike the observational data, no factors were created based on the monitoring period.

### 4.4. Factor selection

In the model development process, factors that can provide useful information for prediction need to be identified. Advantages of an appropriate factor selection method include reducing the dimensionality of the problem, reducing the noise, and identifying more important and more interpretable factors. In a previous study (Jahangiri et al., 2015a,b), the minimum redundancy maximum relevance (mRMR) approach was adopted to select the five most representative factors in predicting RLR violations. This approach attempts to find factors that have both the highest level of relevance and the lowest level of redundancies between factors (Ding and Peng, 2005). In the present study, a different approach was used to identify the most useful factors. Although the RF method is considered as a black box machine learning algorithm, an invaluable benefit of using RF is that it produces insights about factor importance. While using the RF method for developing prediction models, the individual contribution of each factor, called the factor importance, was obtained. Thus, the factors were ranked based on that measure. In other words, all of the factors were taken into account to develop prediction models; as a result, the importance of individual factors was obtained. Subsequently, a desirable subset of only important factors (e.g., top 5, or top 10) can be identified based on the factor importance measure. The factor selection based on the factor importance metric leads to selecting the most representative factors in a more accurate way when compared to other techniques, such as mRMR. This is because the actual contribution of each factor is assessed while developing the prediction models. However, the factor selection using factor importance is most valuable when using the RF method as it internally calculates the factor importance.

## 5. Results

For each data set, prediction models were developed, and the individual factor importance was obtained. To assess the model performances, the OOB error was used. When evaluating tree-based models such as RF, the unbiased estimation of the error, namely the OOB error, is obtained internally and is nearly identical to the cross-validation accuracy (Hastie et al., 2009).

### 5.1. Observational data results

To implement the RF method, the R software and RandomForest package were used (Liaw and Wiener, 2002; R Core Team, 2014). First, all of the factors that were obtained or created as listed in Table 4 were taken into account to develop a prediction model. When developing the RF model, different numbers of trees was examined, as shown in Fig. 2. The error rate became stable by increasing the number of trees beyond 400. However, since increasing the number of trees does not lead to over-fitting, the value of 800 was selected to ensure that a sufficient number of trees were applied. Another parameter that needed to be determined was the number of factors that each tree requires to grow. As Fig. 3 illustrates, the effect of the number of factors was negligible, but the lowest error rate was achieved when six factors were used to grow each tree.

Furthermore, to conduct the sensitivity analysis, different monitoring periods were evaluated. As mentioned earlier, a monitoring period is defined by two parameters: the start and end points. The start point is always the *DTI* at the yellow onset. Therefore, different monitoring periods were obtained by changing the end point, as shown in Fig. 4. As a result, monitoring periods with different lengths (i.e., from 2 to 30 m) were assessed.

Using all 17 factors in the observational data, the models contributed to low error rates (0.11–0.53%) for all monitoring periods, as shown in Fig. 5. The factor importance was obtained to rank all factors and identify the most useful ones. Factor importance can be assessed based on two measures: (1) Mean decrease accuracy, which shows how the detection accuracy is decreased if a factor

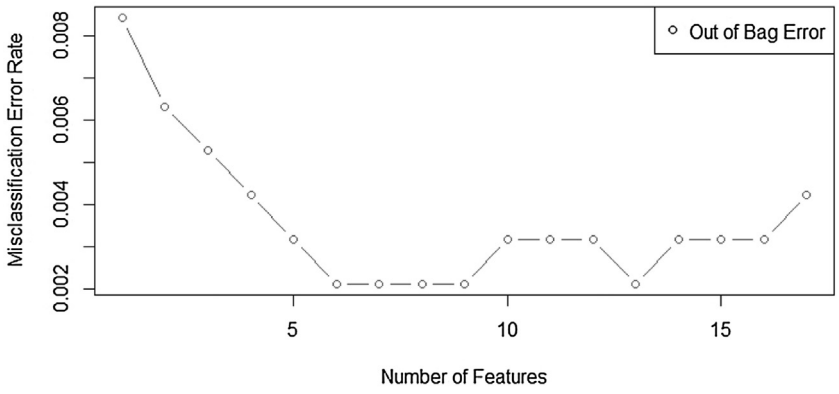


**Fig. 3.** Observational data – selecting the number of factors for each tree.

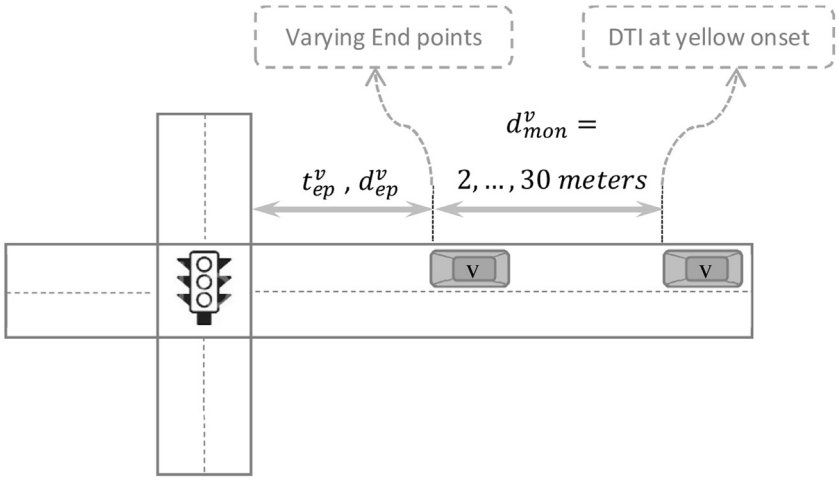


**Fig. 4.** Parameters to select monitoring periods.

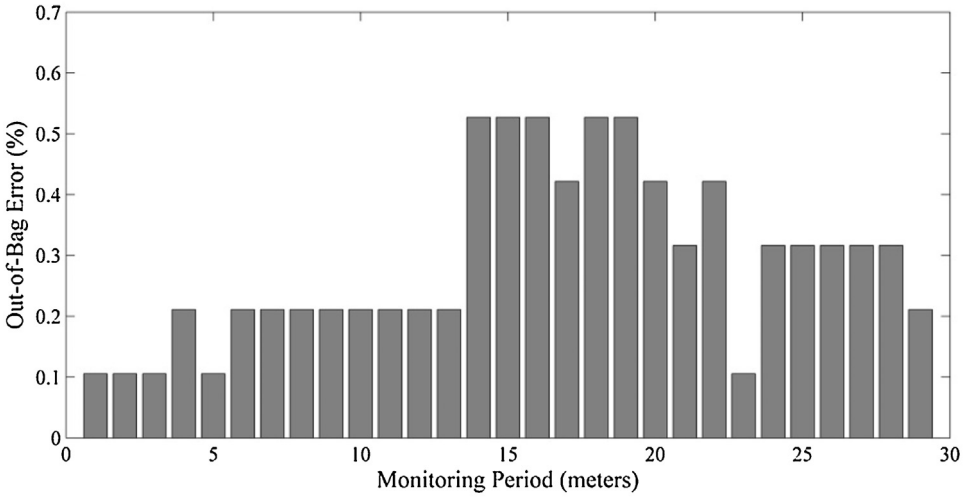


**Fig. 5.** Observational data – OOB error rate in percentage using all 17 observational data factors.

is excluded, averaged across all trees, and normalized by the standard deviation of the differences in accuracy; and (2) Mean decrease Gini, which shows how a single factor contributes to decrease the Gini index across all of the trees. The factors identified by these two measures were found to be the same. Thus, only the mean decrease Gini was used for evaluations. The order of importance for the factors was found to be different when different monitoring periods were assessed. For example, Fig. 6 illustrates how the importance of the third factor identified in Table 4 (i.e., acceleration at the onset of a yellow light) changes for different monitoring periods. The figure shows that this factor was recognized as more important when the monitoring period of 4–7 m was used compared to when longer monitoring periods were applied (e.g., >10 m). Therefore, the factor importance change was taken into account when a different

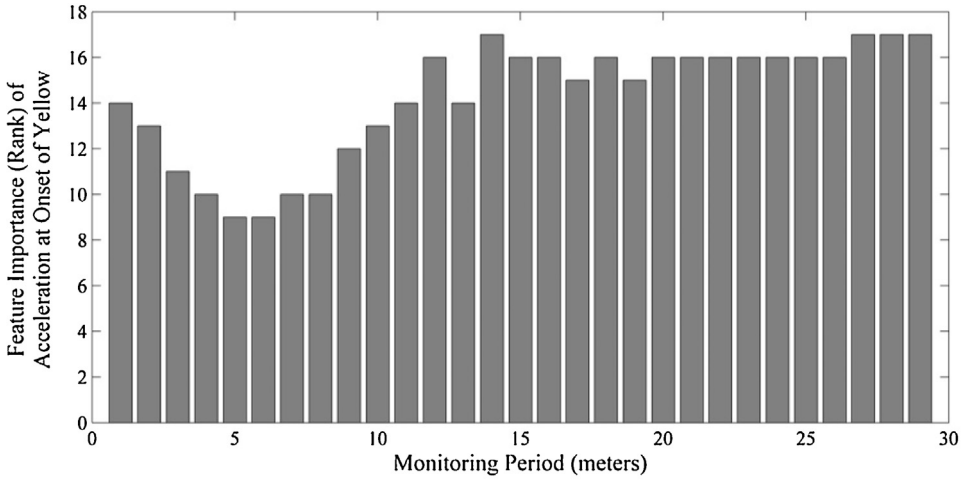


**Fig. 6.** Observational data – factor importance change – acceleration at onset of yellow.

**Table 6**
Factor importance change.

| | | Factor Rank | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| Monitoring Period (meters) | 2 | 4 | 17 | 1 | 16 | 5 | 2 | 7 | 10 | 6 | 9 | 8 | 13 | 12 | 3 | 15 | 11 | 14 |
| | 3 | 4 | 17 | 1 | 16 | 5 | 7 | 2 | 9 | 10 | 6 | 8 | 13 | 3 | 15 | 14 | 11 | 12 |
| | 4 | 4 | 17 | 1 | 16 | 5 | 2 | 9 | 6 | 10 | 7 | 3 | 13 | 14 | 15 | 8 | 12 | 11 |
| | 5 | 4 | 17 | 1 | 16 | 5 | 2 | 9 | 6 | 10 | 3 | 7 | 15 | 8 | 13 | 11 | 14 | 12 |
| | 6 | 4 | 17 | 1 | 16 | 7 | 8 | 5 | 2 | 3 | 6 | 15 | 9 | 10 | 11 | 13 | 12 | 14 |
| | 7 | 4 | 17 | 1 | 16 | 7 | 8 | 5 | 2 | 3 | 6 | 9 | 15 | 10 | 11 | 13 | 12 | 14 |
| | 8 | 4 | 17 | 1 | 16 | 7 | 8 | 5 | 2 | 9 | 3 | 6 | 15 | 10 | 11 | 13 | 12 | 14 |
| | 9 | 4 | 17 | 1 | 16 | 7 | 8 | 5 | 2 | 11 | 3 | 15 | 6 | 9 | 10 | 13 | 12 | 14 |
| | 10 | 4 | 17 | 1 | 16 | 7 | 8 | 5 | 2 | 15 | 13 | 9 | 3 | 6 | 11 | 10 | 12 | 14 |
| | 11 | 4 | 17 | 1 | 16 | 7 | 5 | 8 | 2 | 13 | 11 | 9 | 6 | 3 | 10 | 15 | 12 | 14 |
| | 12 | 4 | 17 | 1 | 16 | 5 | 7 | 2 | 13 | 8 | 9 | 6 | 15 | 11 | 3 | 10 | 12 | 14 |
| | 13 | 4 | 17 | 1 | 16 | 5 | 7 | 2 | 13 | 15 | 9 | 8 | 6 | 10 | 11 | 12 | 3 | 14 |
| | 14 | 4 | 17 | 1 | 16 | 7 | 5 | 2 | 15 | 8 | 13 | 9 | 6 | 11 | 3 | 12 | 10 | 14 |
| | 15 | 4 | 17 | 1 | 16 | 5 | 7 | 2 | 15 | 8 | 9 | 6 | 13 | 10 | 11 | 12 | 14 | 3 |
| | 16 | 4 | 17 | 1 | 16 | 7 | 5 | 15 | 2 | 12 | 9 | 8 | 11 | 6 | 10 | 13 | 3 | 14 |
| | 17 | 4 | 17 | 1 | 16 | 5 | 7 | 15 | 2 | 12 | 9 | 6 | 11 | 13 | 10 | 8 | 3 | 14 |
| | 18 | 4 | 17 | 1 | 16 | 5 | 7 | 15 | 2 | 12 | 9 | 6 | 10 | 13 | 11 | 3 | 14 | 8 |
| | 19 | 4 | 17 | 1 | 16 | 5 | 15 | 7 | 2 | 12 | 9 | 6 | 11 | 13 | 10 | 8 | 3 | 14 |
| | 20 | 4 | 17 | 1 | 16 | 5 | 15 | 7 | 2 | 12 | 9 | 6 | 11 | 10 | 13 | 3 | 8 | 14 |
| | 21 | 4 | 17 | 1 | 16 | 15 | 5 | 2 | 12 | 13 | 7 | 9 | 6 | 11 | 10 | 8 | 3 | 14 |
| | 22 | 4 | 17 | 1 | 16 | 15 | 5 | 12 | 13 | 2 | 9 | 11 | 6 | 7 | 8 | 10 | 3 | 14 |
| | 23 | 4 | 17 | 1 | 16 | 15 | 5 | 12 | 2 | 11 | 13 | 9 | 8 | 6 | 7 | 10 | 3 | 14 |
| | 24 | 4 | 17 | 1 | 16 | 5 | 15 | 13 | 12 | 2 | 11 | 6 | 9 | 8 | 10 | 7 | 3 | 14 |
| | 25 | 4 | 17 | 1 | 16 | 15 | 5 | 11 | 12 | 2 | 9 | 8 | 13 | 6 | 10 | 7 | 3 | 14 |
| | 26 | 4 | 17 | 1 | 16 | 15 | 5 | 12 | 11 | 13 | 2 | 9 | 6 | 8 | 10 | 7 | 3 | 14 |
| | 27 | 4 | 17 | 1 | 16 | 15 | 5 | 11 | 13 | 12 | 2 | 9 | 6 | 10 | 7 | 8 | 3 | 14 |
| | 28 | 4 | 17 | 1 | 16 | 15 | 5 | 11 | 13 | 12 | 2 | 9 | 6 | 8 | 10 | 7 | 14 | 3 |
| | 29 | 4 | 17 | 1 | 16 | 15 | 5 | 11 | 13 | 2 | 9 | 8 | 6 | 7 | 10 | 12 | 14 | 3 |
| | 30 | 4 | 17 | 1 | 15 | 16 | 11 | 5 | 13 | 2 | 9 | 8 | 6 | 10 | 7 | 12 | 14 | 3 |

number of factors were applied, as shown in Table 6. For example, when using the top four factors, factor numbers 4, 17, and 1 were used for all monitoring periods, factor number 16 was used for monitoring periods of 2–29 m, and factor number 15 was used for the monitoring period of 30 m. As a result, RF models were developed for different combinations of monitoring periods and the number of factors used, as illustrated in Fig. 7.

According to Table 6, factor numbers 4 (*TTI* at the onset of a yellow light), 17 (*mean* (*TTI*) over $t^{v}_{mon}$), 1 (*DTI* at the onset of a yellow light), and 16 (*mean* (*DTI*) over $t^{v}_{mon}$) were found to be the four most important factors for all monitoring periods except one case; there was one situation during which factor 15 was identified as the fourth important factor when using a monitoring period of 30 m. Factor numbers 5, 7, 15, and 16 were identified as the fifth most important factor, depending on which monitoring period was used.

It was expected that the *TTI* at the onset of a yellow indication would be among the most important factors in predicting RLR violations as this has been found in the past studies to be a significant factor. Other factors found to be important were measured during the monitoring period suggesting that tracking data over the monitoring period is also an important consideration. In other words, the monitoring period can be seen as a period in which the driver decision (i.e. to proceed or stop) can be identified by observing changes in factor measurements such as *mean* (*TTI*) over $t^{v}_{mon}$. For example, if a driver decides to stop, he/she would decrease the vehicle speed and as a result *TTI* would increase. Consequently, *mean* (*TTI*) over $t^{v}_{mon}$ would increase, which can be seen as an indicator of driver attempting to stop.

Based on Fig. 7, low error rates were obtained using different monitoring periods and number of top factors (i.e., from 0.1 ~ 1.6%),

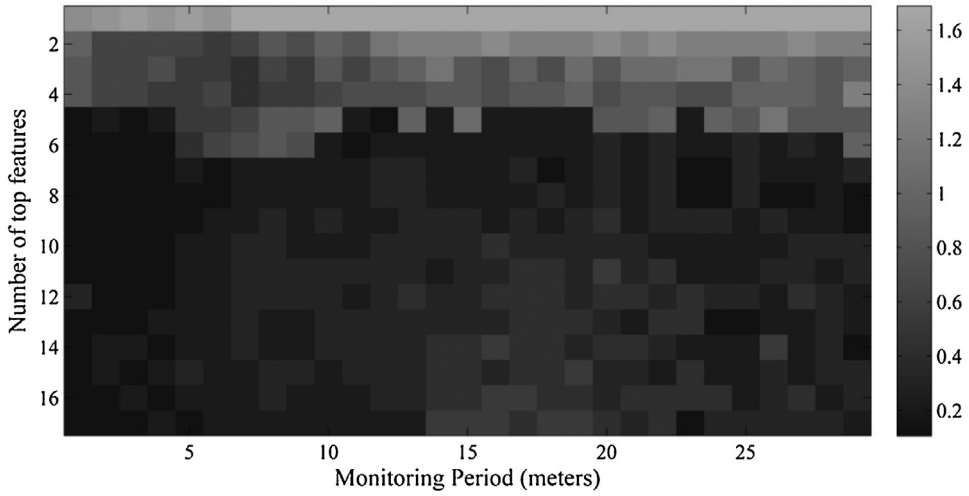


**Fig. 7.** OOB error for different monitoring periods and number of top factors.

with darker regions representing lower error rates. Therefore, the prediction models achieved the lowest error rates when using more than five factors and the monitoring period of 2–6 m. As the monitoring period increases with a specific number of factors (e.g., top 10), the error rate generally increases. As mentioned earlier, the start point of the monitoring period is the DTI at the yellow onset. Hence, it can be inferred that a short monitoring period (i.e., 2–6 m) immediately following the yellow onset is the most appropriate period that should be monitored to predict RLR violations. Moreover, when using more factors with a particular monitoring period (e.g., 5 m), the error rate decreases. However, minimal benefit was gained when employing more than six factors. Comparing to the existing RLR violation prediction models, the RF model resulted in a higher detection accuracy. However, in an ideal condition, comparing model performance between studies should be considered when models are applied to the same data set with the similar parameter settings such as the monitoring period. Comparison of model performance of some traditional RLR violation prediction models as well as two modern techniques can be found in (Aoude et al., 2012).

### 5.2. Simulator data results

As mentioned earlier, factors listed in Table 5 were obtained from the simulator data. To develop RF models, a procedure was used similar to that of the model development using observational data. The number of trees and number of factors used for each tree in the simulator data were determined to be 700 and 3, respectively, as shown in Figs. 8 and 9. Since it was not possible to define a monitoring period for the simulator data set, no sensitivity analysis was conducted. The importance of different factors was obtained and is illustrated in Fig. 10; the associated factor ranking is presented in Table 7. *TTI* at the onset of a yellow light was found to be the most important factor, followed by the *RDP* at the onset of a yellow light and the *DTI* at the onset of a yellow light. Driver factors (i.e., age and gender), the treatment factor, and the secondary task factor were among the least important factors.

Similarly to the observational data model, factor rankings of the simulator data were obtained using both mean decrease Gini and mean decrease accuracy criteria, as shown in Fig. 10. According to Fig. 10 and Table 7, factor numbers 11 (*TTI* at the onset of a yellow light), 12 (*RDP* at the onset of a yellow light), and 8 (*DTI* at the onset of a yellow light) were found by both criteria to be the three most important factors. Different models were developed using a different number of factors with respect to importance.

**Table 7**
Simulator data – factor ranking.

| No. | Factor | Rank |
|---|---|---|
| 1 | Gender | 10, 11 |
| 2 | Age | ***7*** |
| 3 | Acceleration Pedal Changed 10% | ***8, 11*** |
| 4 | Acceleration Pedal Change Direction | ***10, 12*** |
| 5 | Max deceleration | ***4, 6*** |
| 6 | Max Acceleration | ***5, 6*** |
| 7 | Velocity at onset of yellow | ***4, 5*** |
| 8 | DTI at onset of yellow | ***3*** |
| 9 | The secondary task condition | ***8, 9*** |
| 10 | Treatment | ***9, 12*** |
| 11 | TTI at onset of yellow | ***1*** |
| 12 | RDP at onset of yellow | 2 |

For example, a model was developed using only the top five factors, and so forth. After using the top three factors, minimal benefit was gained, as illustrated in Fig. 11 There is even a slight increase in the error rate after using more than three factors, suggesting that adding more factors does not necessarily result in lower error rates. This shows the significance of factor selection, for which obtaining factor importance was shown to be a useful method when developing RF models.

### 5.3. Model comparison: observational data vs. simulator data

A direct comparison between the models developed using the observational data set and the simulator data set is not meaningful. This is because the models were constructed using different data sets, and some factors were available in the observational data set that were unavailable in the simulator data set, and vice versa. However, two points can be made: (1) Important factors identified by models that were developed using both data sets were similar. *TTI* at the onset of a yellow light, *DTI* at the onset of a yellow light, *RDP* at the onset of a yellow light, and velocity at the onset of a yellow light were among the most important factors identified by models constructed using both data sets; and (2) Models developed using the observational data achieved lower error rates compared to those constructed by the simulator data. It initially appears that the observational data models are more accurate (i.e., lower error rates) because the observational data provided information for several frame numbers, thus enabling the use of a monitoring period. However, even when using one factor (i.e., *TTI* at yellow onset) that was not based on the monitoring period in the observational data, the error rate was significantly lower than

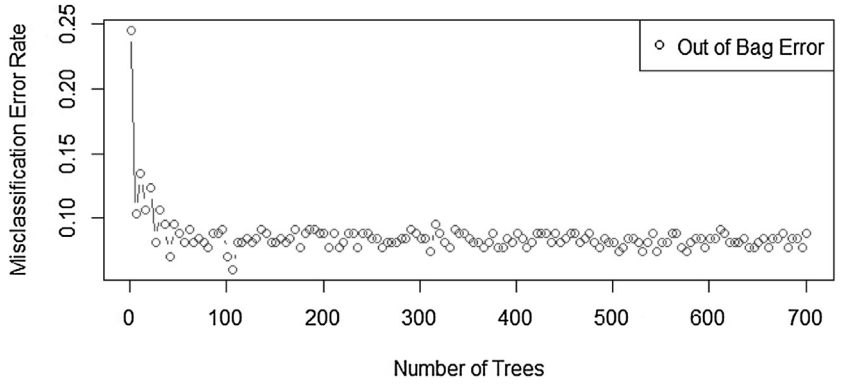


**Fig. 8.** Simulator data – selecting the required number of trees.

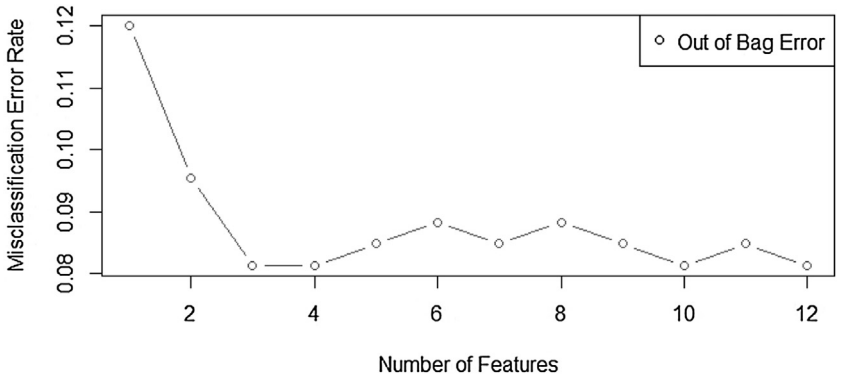


**Fig. 9.** Simulator data – selecting the number of factors for each tree.

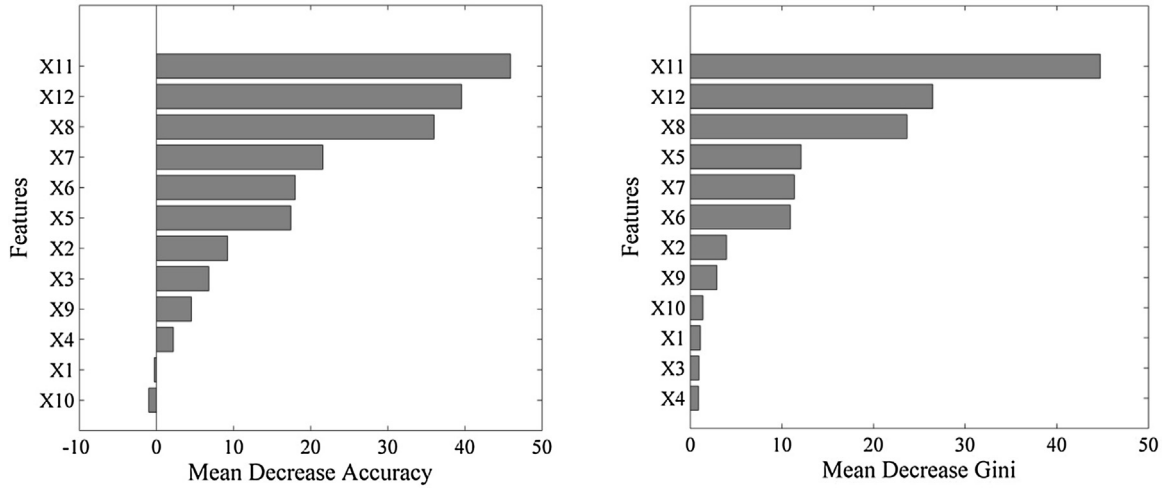


**Fig. 10.** Simulator data – factor importance.

that of the simulator data models (i.e., 1.6%). This suggests that the monitoring period may not be the reason the observational data models show higher performance rates. In other words, using *TTI* at yellow onset, RLR violations were identified with a high accuracy (i.e., 1.6% error rate) with the observational data, while using the same factor in the simulator data model led to a poor performance (i.e., error rate of 17.95%). Thus, it seems that the driver behavior in an observational situation may be significantly different than driver behavior in a simulator condition. This suggests that a direct comparison between simulator data and observational data is necessary in future research endeavors. For this to be a meaningful comparison, the same participants would be required to collect both observational and simulator data.

### 5.4. Implementation considerations

In real-world RLR predictions the factors such as speed and acceleration can be obtained at a high frequency using Vehicle-to-vehicle (V2 V) and vehicle-to-infrastructure (V2I) technology. Thus, significant factors such as RDP can be calculated at any time and consequently the RLR prediction algorithm can constantly monitor individual vehicles as they approach an intersection. In situations where the endangered driver has sufficient time, a warning can be

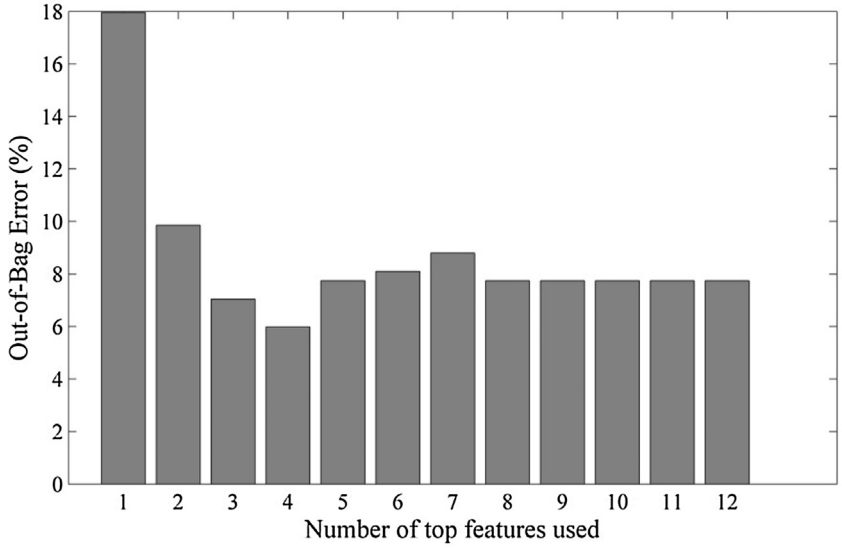


**Fig. 11.** OOB error for different number of top factors used.

issued to the driver to respond. In cases where insufficient time is available, the infrastructure can take appropriate actions (e.g., extending a red clearance interval). Connected-vehicle technology uses dedicated short-range communications (DSRC) that provide a reliable and fast communication (latency of less than 100 ms) and a range of less than 1000 m. Once the prediction model was developed, the prediction time for a single observation was extremely fast (between 1 to 10 ms on an i5–3230 M CPU at 2.6 Hz and RAM of 8 GB). It may take a significant amount of time to develop a model depending on data size and frequency. However, the model is constructed using pre-collected (i.e., historical) data, thus the development time is excluded in real-time applications. Therefore, it appears that the future intersection safety systems will be capable of performing violation predictions and notifying endangered drivers and/or the infrastructure in a fraction of a second.

## 6. Conclusions

The study focused on developing prediction models using observational and simulator data to identify RLR violations before they occur, so the endangered driver and/or the infrastructure can be notified. This is considered vital in advanced intersection safety systems in which vehicles can be tracked (e.g. through connected vehicle technology) as they approach the intersection. The RF machine-learning technique, although a well-known method in the field of Artificial Intelligence has not been applied in RLR violation prediction, was adopted in this study. This method was found to be a powerful technique in predicting RLR violations using observational data (i.e., error rates of 0.1% to 1.6%). Moreover, the RF method produces additional insight into factor importance, which contributes to identifying the most significant factors in predicting RLR behavior. Two types of factors were evaluated: (1) Factors that describe quantities at a point in time (e.g., *TTI* at yellow onset) and (2) Factors that describe quantities over a time period that occurs during the defined monitoring period (e.g., *max* (*Velocity*) over $t^{v}_{mon}$). The importance of the first type has been widely stated throughout the literature and was also reaffirmed in the present study. The second type was also found critical; this shows that, in addition to the yellow onset, an appropriate monitoring period can provide useful information reflecting driver behavior (i.e., decision to stop or proceed) when approaching a signalized intersection. A sensitivity analysis showed that the monitoring period lengths of 2–6 m led to the best prediction accuracy. Applying simulator data resulted in fair prediction accuracies (i.e., error rates of 5.9% to 17.9%). Simulator data had the advantage of accounting for driver factors (i.e., age and gender) and specific hypothetical scenarios, as indicated by the treatment and the secondary task condition factors. However, these factors were found to be among the least important in predicting RLR violations.

Comparing the models developed using observational and simulator data, the *TTI* at the onset of a yellow indication, *DTI* at the onset of a yellow indication, *RDP*at the onset of a yellow indication, and velocity at the onset of a yellow indication were among the most important factors identified by models constructed using both data sets. Moreover, models developed using observational data contributed to higher prediction accuracies. Using only one common factor (i.e. *TTI* at the yellow onset) in model development, the observational data model resulted in a 1.6% error, whereas the simulator data model led to a high error rate of 17.95%. However, a direct comparison between models developed using two different data sets may not be appropriate. To have a meaningful comparison, the same human subjects would be required to collect both observational and simulator data.

## Acknowledgement

This research effort was funded by the Tier 1 U.S. Department of Transportation Connected Vehicle/Infrastructure University Transportation Center (CVI-UTC).

## References

Abbas, M., Machiani, S.G., Garvey, P.M., Farkas, A., Lord-Attivor, R., 2014. Modeling the Dynamics of Driver's Dilemma Zone Perception Using Machine Learning Methods for Safer Intersection Control.

Amer, A., Rakha, H., El-Shawarby, I., 2011a. Agent-based stochastic modeling of driver decision at onset of yellow light at signalized intersections. Transp. Res. Rec.: J. Transp. Res. Board 2241 (1), 68–77.

Amer, A., Rakha, H., El-Shawarby, I., 2011b. Novel stochastic procedure for designing yellow intervals at signalized intersections. J. Transp. Eng. 138 (6), 751–759.

Aoude, G.S., Desaraju, V.R., Stephens, L.H., How, J.P., 2012. Driver behavior classification at intersections and validation on large naturalistic data set. Intell. Transp. Syst. IEEE Trans. 13 (2), 724–736.

Balali, V., Golparvar-Fard, M., 2014. Video-based detection and classification of US traffic signs and mile markers using color candidate extraction and feature-based recognition. Computing in Civil and Building Engineering, ASCE.

Bonneson, J.A., Son, H.J., 2003. Prediction of expected red-light-running frequency at urban intersections? Transp. Res. Rec.: J. Transp. Res. Board 1830 (1), 38–47.

Bonneson, J.A., Middleton, D., Zimmerman, K., Charara, H., Abbas, M., 2002. Intelligent Detection-control System for Rural Signalized Intersections. Texas Transportation Institute, Texas A&M University System.
Boyle, L.N., Lee, J.D., 2010. Using driving simulators to assess driving safety. Accid. Anal. Prev. 42 (3), 785–787.
Breiman, L., 2001. Random forests? Mach. Learn. 45 (1), 5–32.
Caird, J.K., Chisholm, S., Edwards, C.J., Creaser, J.I., 2007. The effect of yellow light onset time on older and younger drivers' perception response time (PRT) and intersection behavior. Transp. Res. F: Traffic Psychol. Behav. 10 (5), 383–396.
Chang, Myung-Soon, Carroll J. Messer, Alberto J. Santiago, 1985. Timing Traffic Signal Change Intervals Based on Driver Behavior. No. HS-040 068.
Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinf. Comput. Biol. 3 (2), 185–205.
Dingus, T.A., Klauer, S., Neale, V., Petersen, A., Lee, S., Sudweeks, J., Perez, M., Hankey, J., Ramsey, D., Gupta, S., 2006. The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment (No. HS-810 593).
Doerzaph, Z.R., Neale, V., 2010. Data acquisition method for developing crash avoidance algorithms through innovative roadside data collection. Transportation Research Board 89th Annual Meeting.
Doerzaph, Z.R., Neale, V., Kiefer, R., 2010. Cooperative intersection collision avoidance for violations: threat assessment algorithm development and evaluation method. Transportation Research Board 89th Annual Meeting.
El-Shawarby, I., Rakha, H.A., Inman, V.W., Davis, G.W., 2007. Age and gender impact on driver behavior at the onset of a yellow phase on high-speed signalized intersection approaches. Transportation Research Board 86th Annual Meeting.
Elhenawy, M., Rakha, H., El-Shawarby, I., 2014. Enhancing driver stop/run modeling at the onset of a yellow indication using historical behavior and machine learning techniques. Transportation Research Board 93rd Annual Meeting.
Elhenawy, M., Jahangiri, A., Rakha, H.A., El-Shawarby, I., 2015. Classification of driver stop/run behavior at the onset of a yellow indication for different vehicles and roadway surface conditions using historical behavior. In: 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015, Las Vegas, Nevada, USA.
Elmitiny, N., Yan, X., Radwan, E., Russo, C., Nashar, D., 2010. Classification analysis of driver's stop/go decision and red-light running violation. Accid. Anal. Prev. 42 (1), 101–111.
Gates, T.J., Noyce, D.A., Laracuente, L., Nordheim, E.V., 2007. Analysis of driver behavior in dilemma zones at signalized intersections. Transp. Res. Rec.: J. Transp. Res. Board 2030 (1), 29–39.
Gazis, D., Herman, R., Maradudin, A., 1960. The problem of the amber signal light in traffic flow. Oper. Res. 8 (1), 112–132.
Ghanipoor Machiani, S., Abbas, M., 2014a. Dynamic driver's perception of dilemma zone: experimental design and analysis of driver's learning in a simulator study. In: The 93nd Annual Meeting of the Transportation Research Board, Washington, DC.
Ghanipoor Machiani, S., Abbas, M., 2014b. Predicting drivers decision in dilemma zone in a driving simulator environment using canonical discriminant analysis. In: The 93nd Annual Meeting of the Transportation Research Board, Washington, DC.
Ghanipoor Machiani, S., Abbas, M., 2015a. Assessment of driver stopping prediction models before and after the onset of yellow using two driving simulator datasets. Accid. Anal. Prev. http://www.sciencedirect.com/science/article/pii/S0001457515001785.
Ghanipoor Machiani, S., Abbas, M., 2015b. Safety surrogate histograms (SSH): a novel real-time safety assessment of dilemma zone related conflicts at signalized intersections. Accid. Anal. Prev. http://www.sciencedirect.com/science/article/pii/S000145751500161X.
Haque, M.M., Ohlhauser, A.D., Washington, S., Boyle, L.N., 2015. Decisions and actions of distracted drivers at the onset of yellow lights. Accid. Anal. Prev. http://www.sciencedirect.com/science/article/pii/S0001457515001293.
Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R., 2009. The Elements of Statistical Learning. Springer.
IIHS, 2010. Status Report: Public Seeks Safer Roads but Still Takes Risks vol. 45, http://www.iihs.org/iihs/sr/statusreport/article/45/12/2.
IIHS, 2012. Red Light Running. IIHS http://www.iihs.org/iihs/topics/t/red-light-running/topicoverview.
Jahangiri, A., Rakha, H., 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE transactions on intelligent transportation systems 16, no. 5, 2406–2417.
Jahangiri, A., Rakha, H., Dingus, T.A., 2015a. Predicting red-light running violations at signalized intersections using machine learning techniques. Transportation Research Board 93rd Annual Meeting.
Jahangiri, A., Rakha, H.A., Dingus, T.A., 2015b. Adopting machine learning methods to predict red-light running violations. 18th International IEEE Conference on Intelligent Transportation Systems (ITSC).
Kim, Z., 2008. Robust lane detection and tracking in challenging scenarios. Intell. Transp. Syst. IEEE Trans. 9 (1), 16–26.
Li, H., Rakha, H., El-Shawarby, I., 2012. Designing yellow intervals for rainy and wet roadway conditions. Int. J. Transp. Sci. Technol. 1 (2), 171–190.
Liang, Y., Reyes, M.L., Lee, J.D., 2007. Real-time detection of driver cognitive distraction using support vector machines. Intell. Transp. Syst. IEEE Trans. 8 (2), 340–350.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2 (3), 18–22.
Liu, Y., Chang, G.-L., Tao, R., Hicks, T., Tabacek, E., 2007. Empirical observations of dynamic dilemma zones at signalized intersections? Transp. Res. Rec.: J. Transp. Res. Board 2035 (1), 122–133.
Liu, Y., Chang, G.-L., Yu, J., 2011. Empirical study of driver responses during the yellow signal phase at six maryland intersections? J. Transp. Eng. 138 (1), 31–42.
McLaughlin, S.B., Hankey, J.M., Dingus, T.A., 2008. A method for evaluating collision avoidance systems using naturalistic driving data. Accid. Anal. Prev. 40 (1), 8–16.
Mussa, R.N., Newton, C.J., Matthias, J.S., Sadalla, E.K., Burns, E.K., 1996. Simulator evaluation of green and flashing amber signal phasing? Transp. Res. Rec.: J. Transp. Res. Board 1550 (1), 23–29.
McGee, H., Moriarty, K., Eccles, K., Liu, M., Gates, T., Retting, R., 2012. NCHRP, Guidelines for Timing Yellow and All-Red Intervals at Signalized Intersections (No. 731). Transportation Research Board, Washington, DC.
NHTSA, 2014. Traffic Safety Facts 2012, National Center for Statistics and Analysis. US Department of Transportation, Washington, DC, pp. 812032.
Neale, V.L., McGhee, C.C., 2006. Intersection Decision Support: Evaluation of a Violation Warning System to Mitigate Straight Crossing Path Collisions. Virginia Transportation Research Council.
Pant, P.D., Cheng, Y., Rajagopal, A., Kashayi, N., 2005. Field Testing and Implementation of Dilemma Zone Protection and Signal Coordination at Closely-spaced High-speed Intersections. University of Cincinnati.
Peng, Y., Boyle, L.N., Ghazizadeh, M., Lee, J.D., 2013. Factors affecting glance behavior when interacting with in-vehicle devices: implications from a simulator study. In: 7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Bolton Landing, NY.
Porter, B.E., Berry, T.D., 2001. A nationwide survey of self-reported red light running: measuring prevalence, predictors, and perceived consequences. Accid. Anal. Prev. 33 (6), 735–741.
Porter, B.E., England, K.J., 2000. Predicting red-light running behavior: a traffic safety study in three urban settings. J. Safety Res. 31 (1), 1–8.
R Core Team, 2014. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URLhttp://www.R-project.org/.
Rakha, H., El-Shawarby, I., Setti, J.R., 2007. Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger. Intell. Transp. Syst. IEEE Trans. 8 (4), 630–640.
Retting, R.A., Williams, A.F., 1996. Characteristics of red light violators: results of a field investigation. J. Saf. Res. 27 (1), 9–15.
Retting, R.A., Ferguson, S.A., Farmer, C.M., 2008. Reducing red light running through longer yellow signal timing and red light camera enforcement: results of a field investigation? Accid. Anal. Prev. 40 (1), 327–333.
Sheffi, Y., Mahmassani, H., 1981. A model of driver behavior at high speed signalized intersections. Transp. Sci. 15 (1), 50–61.
Wei, H., Li, Z., Yi, P., Duemmel, K.R., 2011. Quantifying dynamic factors contributing to dilemma zone at high-Speed signalized intersections? Transp. Res. Rec.: J. Transp. Res. Board 2259 (1), 202–212.
Yuan, F., Cheu, R.L., 2003. Incident detection using support vector machines. Transp. Res. C: Emerg. Technol. 11 (3), 309–328.
Zegeer, C., R. Deen, 1978. Green-extension systems at high-speed intersections. ITE J. 48, no. 11, 19–24.
Zhang, L., Zhou, K., Zhang, W.-b., Misener, J.A., 2009. Prediction of red light running based on statistics of discrete point sensors. Transp. Res. Rec.: J. Transp. Res. Board 2128 (1), 132–142.
Zhang, L., Wang, L., Zhou, K., Zhang, W.-b., 2012. Dynamic all-red extension at a signalized intersection: a framework of probabilistic modeling and performance evaluation. Intell. Transp. Syst. IEEE Trans. 13 (1), 166–179.