Transportation Research Part C 72 (2016) 225–238

Contents lists available at ScienceDirect

Transportation Research Part C

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

A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data

Chuan Ding, Xinkai Wu *, Guizhen Yu, Yunpeng Wang

School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 16 June 2016
Received in revised form 18 September 2016
Accepted 28 September 2016
Available online 7 October 2016

Keywords:
Driving behavior
Gradient boosting logit
Variable importance
High-resolution traffic even data

ABSTRACT

Driver's stop-or-run behavior at signalized intersection has become a major concern for the intersection safety. While many studies were undertaken to model and predict drivers' stop-or-run (SoR) behaviors including Yellow-Light-Running (YLR) and Red-Light-Running (RLR) using traditional statistical regression models, a critical problem for these models is that the relative influences of predictor variables on driver's SoR behavior could not be evaluated. To address this challenge, this research proposes a new approach which applies a recently developed data mining approach called gradient boosting logit model to handle different types of predictor variables, fit complex nonlinear relationships among variables, and automatically disentangle interaction effects between influential factors using high-resolution traffic and signal event data collected from loop detectors. Particularly, this research will first identify a series of related influential factors including signal timing information, surrounding traffic information, and surrounding drivers' behaviors using thousands drivers' decision events including YLR, RLR, and first-to-stop (FSTP) extracted from high-resolution loop detector data from three intersections. Then the research applies the proposed data mining approach to search for the optimal prediction model for each intersection. Furthermore, a comparison was conducted to compare the proposed new method with the traditional statistical regression model. The results show that the gradient boosting logit model has superior performance in terms of prediction accuracy. In contrast to other machine learning methods which usually apply 'black-box' procedures, the gradient boosting logit model can identify and rank the relative importance of influential factors on driver's stop-or-run behavior prediction. This study brings great potential for future practical applications since loops have been widely implemented in many intersections and can collect data in real time. This research is expected to contribute to the improvement of intersection safety significantly.

1. Introduction

In recent years, concerns over intersection safety issue have prompted a growing body of study into the driver's Stop-or-Run (SoR) behavior at signalized intersections. There is a significant number of intersection crashes caused by the driving

* Corresponding author.
E-mail addresses: cding@buaa.edu.cn (C. Ding), xinkaiwu@buaa.edu.cn (X. Wu), yugz@buaa.edu.cn (G. Yu), ypwang@buaa.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.trc.2016.09.016
0968-090X/© 2016 Elsevier Ltd. All rights reserved.

behavior every year. According to the National Highway Traffic Safety Administration's (NHTSA) report (NHTSA, 2012), more than 2.3 million crashes occurred at the intersection, and a great number of them were related to the drivers' SoR behavior such as yellow-light running (YLR) and red-light running (RLR). Therefore, it is critical to model and understand the drivers' SoR behavior at signalized intersections, and thereby predict it accurately to improve the intersection safety (Wu et al., 2013).

Many studies were undertaken to model and predict the YLR and RLR using the statistical regression models (e.g. Lu et al., 2015; Wu et al., 2013; Zhang et al., 2009; Bonneson and Son, 2003; Sheffi and Mahmassani, 1981; etc.). However, a critical problem for these models is that the relative influences of predictor variables on driver's SoR behavior were not evaluated. Understanding the relative importance of the influential factors on driver's SoR behavior would significantly help SoR prediction therefore contributes to the future improvement of intersection safety. Furthermore, understanding the relative importance of the influential factors would help to identify influential factors for which data should be collected and maintained. This is beneficial for data collection and maintenance as the process of collecting and maintaining data is cost-prohibitive. Although the sensitivity analysis can be conducted based on the statistical regression model (Saha et al., 2015), only one variable is evaluated at one time under the assumption that other variables remain the unchanged values. Therefore, the important relationships among the influential factors have been ignored.

To address the above mentioned challenge, this research proposes a brand new approach which applies a recently developed data mining approach called gradient boosting logit model (Zhang and Haghani, 2015; Saha et al., 2015). The proposed method is based on traffic data collected from loop detectors. Much existing research on YLR or RLR were using the high quality video data (e.g. Bonneson and Son, 2003; Bonneson and Zimmerman, 2004; Gates et al., 2007; Yang and Najm, 2007; Zhang et al., 2009; etc.). Video data is a reliable source, but such data is relatively rare since the data quality is constrained by the types of video cameras. Furthermore, most of the video data is off-line, and real-time video data analysis is time-consuming and costly.

The proposed research is based on loop detector data, which can be easily and automatically obtained in real time with low cost since most of signalized intersections have been equipped with loop detectors. Particularly, this research utilizes high-resolution traffic event data collected by loop detections. Traditional loop detector data which are usually aggregated into 30 s, 5 min or even 15 min are too coarse to describe individual driver's SoR behavior. With recent improvement of data collection methods (Lu et al., 2015; Liu et al., 2009; Smaglik et al., 2007), high-resolution traffic data (event-based or second-by-second data), which provide detailed vehicle arrivals and departures from loop detectors, become more and more popular. Such data, combined with the signal phase changes provided by signal control system, could be better used to analyze and predict the driver's SoR behavior (Wu et al., 2013).

The proposed data mining approach is designated to address challenging problems, such as drivers' behavior study, which has mixed types of predictor variables and complex nonlinear relationships. The proposed method can also automatically disentangle interaction effects between influential factors. To implement this approach, this research will first identify a series of related influential factors including signal timing information (e.g. time to yellow start and used yellow time), surrounding traffic information (e.g. occupancy time and time gaps of surrounding vehicles), and surrounding drivers' behaviors (i.e. drivers' RLR, YLR, and green-light running (GLR) decisions) using thousands drivers' decision events including YLR, RLR, and first-to-stop (FSTP) extracted from high-resolution loop detector data from three intersections. Then the research applies the proposed data mining approach to search for the optimal prediction model for each intersection. Furthermore, a comparison was conducted to compare the proposed new method with the traditional statistical regression model. In contrast to sensitivity analysis and other machine learning methods as 'black-box' procedures (Ding et al., 2015; Ma et al., 2015; Yu et al., 2014), the proposed method can identify and rank the relative importance of influential factors on driver's SoR behavior prediction. This study is expected to contribute to the improvement of intersection safety significantly.

The remaining of the paper is organized as follows. The second section provides a brief description of data collection, followed by the model specification. Model results and discussion are demonstrated in the forth section. Conclusion and future research directions are outlined at the end.

2. Data collection

Data collection has three parts: first is to collect high-resolution traffic and signal event data; second is to identify YLR, RLR and FSTP during yellow using stop-bar detectors, and third is to match events between stop-bar and advance detectors since the information collected from advance detectors will be used for investigation. Here stop-bar detectors are the detectors located right behind the stop-line and advance detectors are the detectors located 400 feet upstream from the stop-line (see Fig. 1).

2.1. High-resolution traffic and signal event data collection

The High-resolution traffic event data was collected by the SMART-SIGNAL (Systematic Monitoring of Arterial Road Traffic and SIGNAL) system developed at the University of Minnesota (Liu et al., 2009).The SMART-SIGNAL is capable of continuously collecting and archiving high-resolution event-based vehicle-detector actuation and signal phase change data. The

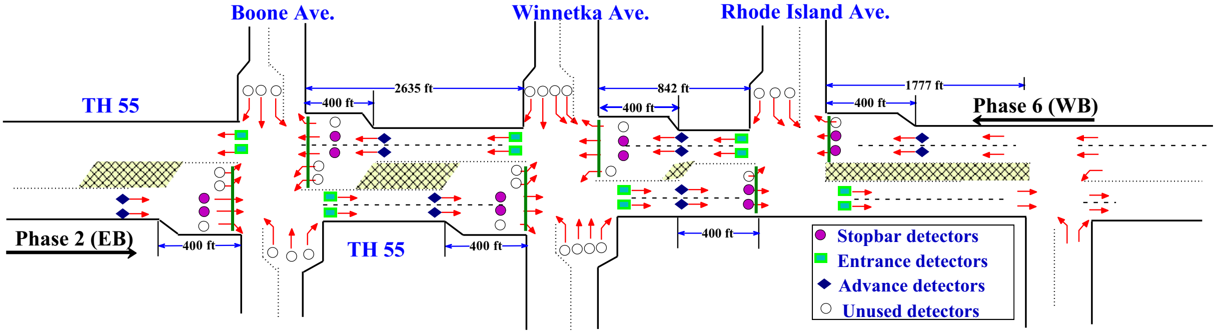


Fig. 1. Layout of trunk highway (TH) 55 test site (EB = eastbound; WB = westbound).

SMART-SIGNAL system has been installed on a major arterial (Trunk Highway 55) with six intersections in the Twin Cities area since July 2008. All intersections are equipped with vehicle-actuated signals, with advance detectors typically located 400 feet upstream from the stop-line for green extension on the major approach and stop-bar detectors located 10–50 feet behind the stop-line for presence detection on the minor approach. Because three intersections are coordinated with each other, most of the signal timings such as cycle length and yellow time are the same. But different cycle lengths are adopted at separate analysis periods: from 9 PM to 5 AM, a constant cycle length can hardly be determined because of actuated signal control; during peak periods, e.g., 7 AM–9 AM and 4 PM–6 PM, a longer cycle length of 180 s is adopted; during the period of flat peak, two cycle lengths are selected, i.e., 110 s or 150 s; and during 5 AM–6 AM, a shorter cycle length of 60 s is set due to lower flow rates. Based on the data collected in July 2009 from Intersection Winnetka, the correlations between hourly traffic volume and number of YLR during weekday and weekend are presented in Fig. 2. It can be seen that the hourly number of YLR cases proportionally changes with hourly traffic volume.

In addition, for research purposes, we have also installed stop-bar and link entrance detectors on major approaches. In this research, we use the event data collected from three intersections (Boone Ave., Winnetka Ave., and Rhode Island Ave., see Fig. 1 for the detector configuration). Three different months' data are randomly chosen in this research (for Boone Ave., the data from November, 2008, May, 2009, and June, 2009 are used; for Winnetka Ave., the data from November, 2008, January, 2009, and June, 2009 are used; and for Rhode Island Ave., the data from November, 2008, January, 2009, and February, 2009 are used.) Note we only use the stop-bar and advance detectors on the major approach, i.e. TH55. The data collected from entrance detectors are used for verification purposes only.

2.2. First-to-Stop (FSTP), Yellow-Light-Running (YLR), and Red-Light-Running (RLR) identification using stop-bar detectors

The ultimate goal of this research is to develop a model, which can predict drivers' decisions of SoR using the real-time information collected from loop detectors located several hundred feet upstream from stop line, i.e., advance detectors. This

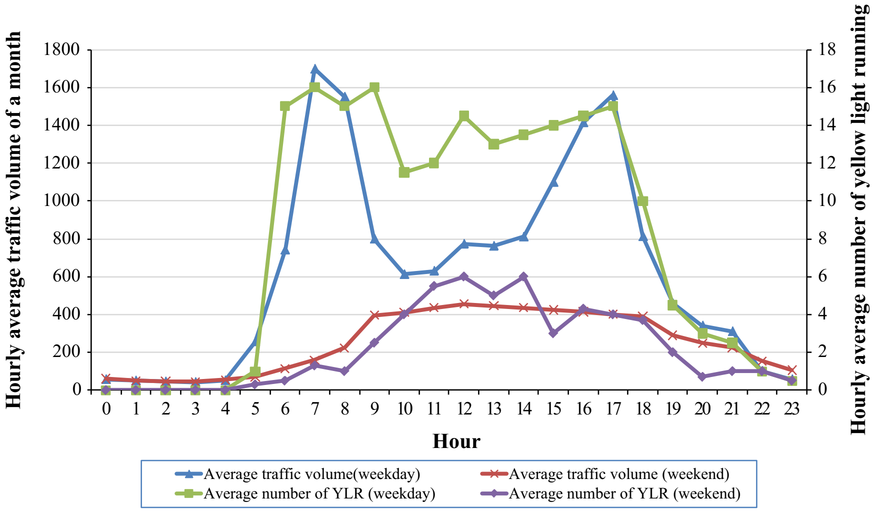


Fig. 2. Average hourly volume and average number of YLR for Intersection Winnetka.

prediction model can then be used to help avoid any potential collisions by adjusting signal timing. In particular, if the proposed model predicts that a driver's decision is "GO", and if we further estimate that there is no enough time for the driver to pass through the stop-line before the signal turns to red based on the information of the yellow time left and the vehicle's current driving status (i.e. speed and acceleration), a potential RLR could be identified. In order to avoid any potential collisions caused by this potential RLR, several methods, such as variable message warning signs, autonomous vehicle technologies, and signal timing adjustment, could be applied. One example for signal timing adjustment to avoid the potential collisions caused by RLR is to extend all red. By extending all red, the potential RLR and any potential collisions caused by the RLR could be avoided.

This paper mainly addresses the prediction of driver's SoR. The first crucial step is to explore whether there is a strong connection between a driver's decision of SoR and her driving speed and signal timing information as well as the information of driving status of preceding and adjacent vehicles. Since a driver's decision at the end of a cycle when the signal changes to yellow could be first-to-stop (i.e. FSTP; note if there is already a stopped vehicle, the driver's decision is simply stop), or running through the intersections either during yellow (YLR) or red (RLR) (see Fig. 3), we need to identify FSTP, YLR, and RLR cases respectively.

However, merely using an advance detector cannot accurately detect RLR, YLR and FSTP. Instead, we use stop-bar detectors to identify these cases. As mentioned before, stop-bar detectors are located 10–50 feet upstream to the stop line. The exact time when a vehicle passes through intersection stop line cannot be recorded. Thus, it is necessary to estimate such time for RLR, YLR and FSTP identification. In real, a driver's decision at the end of a cycle when presented with yellow indication can be simply to decelerate and stop before the stop line (FSTP if no vehicle stops ahead), or run through the intersection at a rather high speed either during yellow (YLR) or during red (RLR). Based on the above analysis, an identification method is developed as follows.

First, stopping and running cases after the onset of yellow light can be initially classified by checking the vehicle's passing speed at stop-bar detectors. The basic idea is intuitive: if the vehicle passes through the stop-bar detector with a relatively high speed (higher than a threshold value), it is concluded that the driver decides to run through the intersection; otherwise the first stopped vehicle is identified as FSTP. For the running cases, if the signal indication is yellow when the vehicle passes through the stop-bar detector, it is an YLR; and it is a RLR if the signal is red.

As the information of individual vehicle's arrival time at stop-bar detector (T_i^{onS} for ith vehicle) and occupancy time (t_i^{onS} for ith vehicle) are recorded, given an effective vehicle length (l_{eff}), which is defined as the sum of vehicle length and detector length (l_d), the speed of ith vehicle (v_i^S) can be calculated by the first equation in Eq. (1). Then, considering the distance between the stop-bar detector and stop line (D_S), the vehicle passing time through the stop line (T_i^{pass}) can be estimated by the second equation in Eq. (1). 25 feet is used as the effective vehicle length since that value was calibrated in the authors' previous research (Liu et al., 2009).

$$\begin{cases} v_i^S = l_{eff}/t_i^{onS} \\ T_i^{pass} = (D_S + l_{eff} - l_d)/v_i^S + T_i^{onS} \end{cases} \tag{1}$$

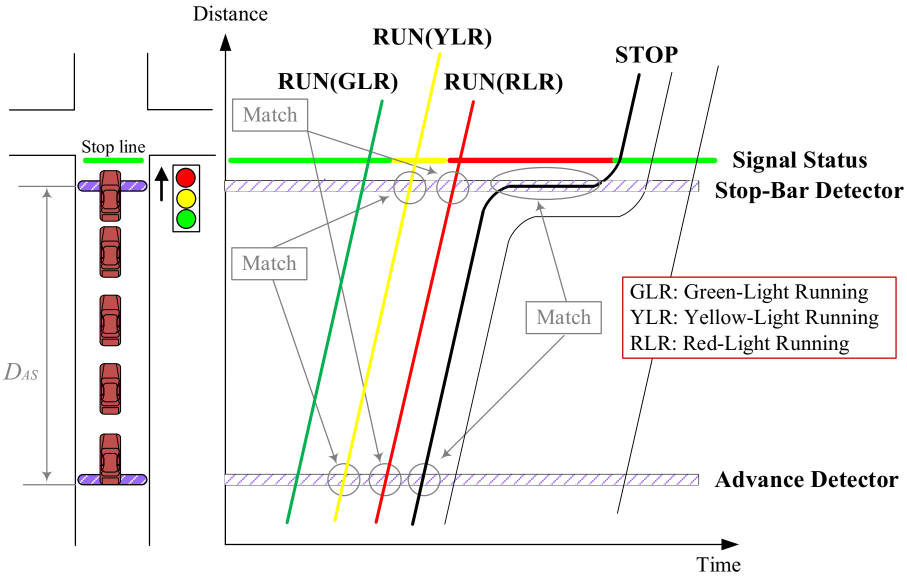


Fig. 3. GLR, YLR, RLR, and FSTP.

Note if the stop-bar detector is installed directly behind the stop-bar, e.g., a short distance of 10 feet as at Intersection Rhode Island shown in Fig. 1, the threshold value is set as 10 mph. But for some intersections like Intersection Boone and Intersection Winnetka, the detectors are located about 50 feet behind stop-line. Therefore, a different threshold value of 20 mph is applied. For the cases with the speed lower than the threshold value (i.e., potential stop), a simple equation of motion is then applied to estimate the vehicle's stopping distance by assuming a deceleration rate of 10 ft/s^2, and confirm that the distance between the stop-bar detector and stop line is long enough for the vehicle to fully stop (Wu et al., 2013). In addition, the speed difference between the target vehicle and three preceding vehicles are also calculated, and used to determine if the vehicle is indeed slowing down or passing through the intersection. Note the method presented here cannot identify all cases of FSTP, YLR, and RLR. Some cases could be missing since the current method focuses more on "accurately" identifying FSTP, YLR, and RLR cases.

2.3. Event data matching between stop bar and advance detectors

The last step is to match RLR events identified by a stop-bar detector with the vehicle events recorded by the corresponding advance detector located on the same lane but a few hundred feet upstream (Fig. 3). As mentioned before, this step is important since only the information collected by advance detectors will later be used to estimate SoR in our models. The matching method introduced here assumes no lane changing takes place between advance and stop-bar detectors. This is an appropriate assumption considering the short distance between two detectors.

A simple "window-searching" method (Lu et al., 2015) is applied to match the events recorded by advance and stop-bar detectors (see Fig. 4). This method first identifies a "time window" for each event recorded by the advance detector based on a possible maximum and minimum travel time required for a vehicle traveling from advance detector to stop-bar detector. Based on the vehicle speed measured by the advance detector v_i^A (as in Eq. (2)), the maximum travel time $TT_i^{\max}$ is estimated base on the assumption that the vehicle will fully stop at the stop-bar, as in Eq. (3); and the minimum travel time $TT_i^{\min}$ is estimated by assuming a maximum acceleration rate of 6 ft/s^2 suggested by Long (2000), as in Eq. (4). The correct match can be detected if one event recorded by the corresponding stop-bar detector has been identified within this window. Note if multiple events within the "time window" have been found, one needs to first estimate the ideal time TI_i required for a vehicle traveling from advance detector to stop bar detector (as in Eq. (5)), and then identify the most likely event to match in light of the percentage error E_i between the ideal travel time and the possible ones as in Eq. (6).

$$v_i^A = \frac{l_{eff}}{t_i^{ocA}} \tag{2}$$

where t_i^{ocA} is the occupancy time of ith vehicle when it passes advance detector.

$$TT_i^{\max} = \frac{2D_{AS}}{v_i^A} \tag{3}$$

$$TT_i^{\min} = \frac{2D_{AS}}{v_i^A + \sqrt{(v_i^A)^2 + 2a_{\max}D_{AS}}} \tag{4}$$

$$TI_i = \frac{2D_{AS}}{v_i^A + v_i^S} \tag{5}$$

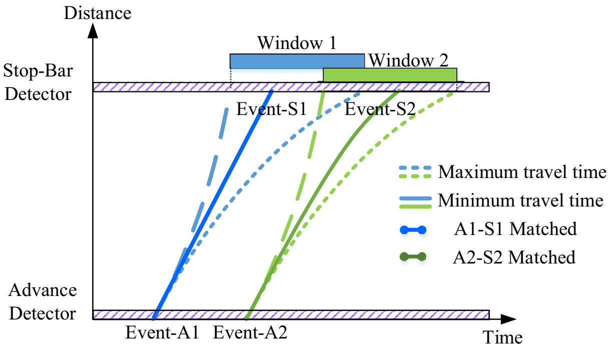


Fig. 4. "Window-searching" method.

$$E_i = |1 - \frac{TI_i}{T_i^{onS} - T_i^{onA}}| \times 100\% \tag{6}$$

3. Methodology

The goal here is to model driver's "stop" (i.e. FSTP) or "run" (i.e. YLR and RLR) behavior and, more importantly, to rank the relative influences of these predictor variables in order to finally improve the prediction accuracy of a driver's SoR decisions. The statistical modeling approach and artificial intelligent method cannot be used to evaluate the relative importance of the influential factors. Although previous literature has highlighted the need to prioritize influential factor in traffic prediction (Saha et al., 2015), recognition of the relative importance of predictor variables in drivers' SoR behavior at signalized intersection modeling is still quite limited, which might be relevant to the complicated modeling process for discrete response, relative to continuous response (e.g. Zhang and Haghani, 2015; Saha et al., 2015; Guelman, 2012). To address this issue, this research proposes a new approach which applies a recently developed data mining approach called gradient boosting logit model. This section provides the mathematical details of the proposed model.

3.1. Gradient boosting logit model

Assuming that $F(x)$ is an approximation function of the discrete response y (-1 = stop, 1 = run) based on a set of predictor variables x, a negative binomial log-likelihood can be formulated as the loss function to estimate the approximation function, as follows (Friedman, 2001, 2002; Saha et al., 2015):

$$L(y, F(x)) = \log(1 + \exp(-2yF(x))), y \in \{-1, 1\} \tag{7}$$

where the values of -1 and 1 indicate "stop" and "run" situations, and

$$F(x) = \frac{1}{2}\log\left[\frac{Pr(y = 1|x)}{Pr(y = -1|x)}\right] \tag{8}$$

Assuming that the number of splits is J for each regression tree, therefore each tree partitions the input space into J disjoint regions $R_{1m}, \ldots, R_{jm}$ and predicts a constant value b_{jm} for region R_{jm}. In this case, each regression tree itself has the additive form as follows (De'ath, 2007; Chung, 2013):

$$h_m(x) = \sum_{j=1}^{J} b_{jm} I(x \in R_{jm}), \ \textit{where} \ I = 1 \ \textit{if} \ x \in R_{jm}; \ I = 0 \ \textit{otherwise} \tag{9}$$

Using the training data $\{y_i, x_i\}_1^N$, the gradient boosting logit regression tree iteratively constructs M different individual regression trees $h_1(x), \ldots, h_M(x)$. The updating approximation function $F_m(x)$ and gradient descent step size ρ_m can be described as follows (Hastie et al., 2009; Zhang and Haghani, 2015):

$$F_m(x) = F_{m-1}(x) + \rho_m \sum_{j=1}^{J} b_{jm} I(x \in R_{jm}) \tag{10}$$

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L\left(y_i, F_{m-1}(x_i) + \rho \sum_{j=1}^{J} b_{jm} I(x \in R_{jm})\right) \tag{11}$$

For the gradient boosting logit regression tree, the gradient descent step size ρ_m becomes

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} \log\left(1 + \exp\left(-2y_i\left(F_{m-1}(x_i) + \rho \sum_{j=1}^{J} b_{jm} I(x \in R_{jm})\right)\right)\right) \tag{12}$$

Using a separate optimal γ_{jm} for each region R_{jm}, b_{jm} can be discarded (Friedman, 2001; Zhang and Haghani, 2015), the Eq. (10) can be alternatively expressed as follows:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}) \tag{13}$$

and the optimal γ_{jm} can be obtained as follows:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} \log(1 + \exp(-2y_i(F_{m-1}(x_i) + \gamma))) \tag{14}$$

There is no closed-form solution to Eq. (14). Following negative binomial log-likelihood, we approximate it by a single Newton-Raphson step (Friedman, 2001). The optimal γ_{jm} becomes

$$\gamma_{jm} = \frac{\sum_{x_i \in R_{jm}} \tilde{y}_i}{\sum_{x_i \in R_{jm}} |\tilde{y}_i|(2 - |\tilde{y}_i|)} \tag{15}$$

where

$$\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F_m(x)=F_{m-1}(x)} = \frac{2y_i}{1 + \exp(2y_i F_{m-1}(x_i))} \tag{16}$$

The gradient boosting logit regression tree builds the model in a stagewise fashion and updates the model by minimizing the expected value of certain loss function. According to the algorithm of the gradient boosting regression tree for continuous response (Friedman, 2002; Hastie et al., 2009; Zhang and Haghani, 2015), the algorithm of the gradient boosting logit model for discrete response can be summarized as follows (see Fig. 5):

To prevent over-fitting and improve prediction accuracy, the gradient boosting logit regression tree applies a shrinkage strategy (Friedman, 2001; Schonlau, 2005). Shrinkage, also called learning rate, is used to scale the contribution of each base tree model by introducing a factor of ξ ($0 < \xi \leqslant 1$) as shown in Eq. (17):

$$F_m(x) = F_{m-1}(x) + \xi \cdot \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}), \ \textit{where}\ 0 < \xi \leqslant 1 \tag{17}$$

where the smaller ξ, the greater the shrinkage is. Through applying the shrinkage strategy, the over-fitting problem can be avoided by reducing or shrinking the impact of each additional tree. Smaller shrinkage values better minimizes loss function. However, it requires larger number of trees to be added to the model. Another parameter, tree complexity C refers to the number of splits (or the number of nodes) that is used for fitting each decision tree. It represents the depth of variable interaction in a tree. Increasing the tree complexity can capture more complex interactions among variables and utilize the strength of gradient boosting logit regression tree. Depending on the value of shrinkage and tree complexity, the corresponding optimal number of trees can be obtained through checking how well the model fits on the dataset (Schonlau, 2005). Optimal performance of the gradient boosting logit regression tree depends on selecting the combination of learning rate and tree complexity. In this study, the pseudo-R^2 is used as the measure of the model performance. The pseudo-R^2 is defined as $R^2 = 1 - L1/L0$, where $L1$ and $L0$ are the log likelihood of the full model and intercept-only model, respectively.

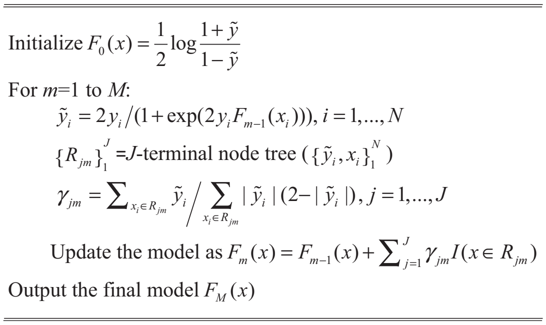
3.2. Relative importance of influential factors

Generally, the influences of the predictor variable on response are different. It is often useful to learn the relative importance or contribution of each independent variable in predicting the response. However, accuracy and interpretability, which are two fundamental objectives of predictive learning, do not always coincide (Guelman, 2012). In contrast to the statistical modeling approach such as autoregressive integrated moving average (ARIMA) type model, support vector machines (SVM), and neural networks, gradient boosting logit regression tree method can identify and rank the influences of predictor variables on response predictions.

For a single decision tree T, Breiman et al. (1984) proposed the following measure as an approximation of relative importance of the predictor x_κ in predicting the response:

$$I_\kappa^2(T) = \sum_{t=1}^{J-1} \hat{\tau}_t^2 I(v(t) = \kappa) \tag{18}$$

where the summation is over the non-terminal nodes t of J-terminal node tree T, x_κ is the splitting variable associated with node t, and $\hat{\tau}_t^2$ is the corresponding empirical improvement in squared error as a result of using predictor x_κ as a splitting



```
Initialize $F_0(x) = \frac{1}{2}\log\frac{1+\bar{y}}{1-\bar{y}}$
For $m$=1 to $M$:
    $\tilde{y}_i = 2y_i/(1+\exp(2y_i F_{m-1}(x_i))),\ i = 1,\dots,N$
    $\{R_{jm}\}_1^J$ =$J$-terminal node tree ($\{\tilde{y}_i, x_i\}_1^N$)
    $\gamma_{jm} = \sum_{x_i \in R_{jm}} \tilde{y}_i \Big/ \sum_{x_i \in R_{jm}} |\tilde{y}_i|(2-|\tilde{y}_i|),\ j = 1,\dots,J$
    Update the model as $F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$
Output the final model $F_M(x)$
```

Fig. 5. Algorithm for the gradient boosting logit regression tree.

variable as the non-terminal node t. For a collection of decision trees $\{T_m\}_1^M$, obtained through gradient boosting approach, can be generalized by its average over all of the additive trees:

$$I_\kappa^2 = \frac{1}{M}\sum_{m=1}^{M} I_\kappa^2(T_m) \tag{19}$$

In this study, we divided the sample data into two subsets: training dataset and test dataset. 80% of the sample data is used as the training data to determine the optimal combination of learning rate, tree complexity, and number of trees for best model performance. The remaining 20% subsets are used as the test data to assess the predictive accuracy of the selected gradient boosting logit model.

4. Model results

4.1. Data preparation

Using the event data collected from three months for each intersection, the program identified a total of 49,531 cases for Intersection Boone, in which 27,035 belong to FSTP, 22,794 are YLR, and 128 are RLR; a total of 15,238 cases for Intersection Winnetka, in which 2804 belong to FSTP, 12009 are YLR, and 425 are RLR; and a total of 33,572 cases for Intersection Rhode Island, in which 10,609 belong to FSTP, 22,582 are YLR, and 381 are RLR. Interestingly, the percent of RLR incidents vary largely at the three intersections: 0.26% vs. 2.79% vs. 1.14% for Boone, Winnetka and Rhode Island respectively. The possible reason, after we carefully explored the data, could be the signal progression design, which intends to make vehicles stop at Boone to form a "gate" to avoid over-congestion at downstream intersections. This "gating" strategy makes many vehicles have to stop at Boone, as indicated by much higher FSTP cases (i.e. 54.58% for Boone vs. 18.29% and 31.60% for Winnetka and Rhode Island, respectively); and also forms a "green wave" for a large portion of traffic driving through Intersections Winnetka and Rhode Island without stops, as indicated by lower FSTP ratios for these two intersections. But this non-stop design at Winnetka and Rhode Island also leads to much higher YLR cases (i.e. 46.02% for Boone vs. 78.35% and 67.26% for Winnetka and Rhode Island, respectively) and RLR cases (i.e. 0.26% for Boone vs. 2.77% and 1.13% for Winnetka and Rhode Island, respectively). The overall low RLR ratios could also be due to the actuation signal control design along the corridor with long amber times of 5.5 s. The more explorations of the impact of signal control on the total numbers of FSTP, RLR and YLR on macroscopic level will be left for future research, as this paper mainly focuses on the predictions of drivers' SoR on microscopic behavior level.

As we know, drivers usually make their decisions at the start of the yellow phase and may adjust their behaviors within a range of area called dilemma zone as indicated in much previous research (Gates et al., 2007; Bonneson and Son, 2003; Sheffi and Mahmassani, 1981). Previous studies have identified some factors that influence driver's SoR decision (Wu et al., 2013; Ren et al., 2016), including approaching speed, yellow time, traffic flow, vehicles on adjacent lanes. According to these studies, here in this research, we consider all the potential factors extracted from the high-resolution traffic and signal event data which could impact a driver's SoR behavior, as described in Table 1.

Using the information collected from detectors, five categorical groups of influential factors were used to predict drivers' SoR behavior at the signalized intersections. For the signal timing related information, time to yellow start (i.e. time left until signal changes to yellow) and used yellow time (i.e. portion of yellow time that elapsed before vehicle arrives at advance detector) were chosen. Using the advance loops, the occupancy time for the target vehicle and preceding vehicle that reflect the vehicle's velocity were collected. The time gaps that reflect the distance between the target vehicle and the leading vehicles at the time at which the target vehicle passes the detector were included. Considering the effects from the preceding vehicle's decision, the vehicles' behaviors for three preceding vehicles were collected. In addition, the information on whether there is a vehicle driving on the adjacent lane was also included. In this study, 20% of sample data is used as the test data while the remaining subsets are used to train the model.

4.2. Model optimization

To test the model performance of different combination of regularization parameters, a series of gradient boosting logit models are built with various learning rate ($\xi = 0.10–0.001$), tree complexity ($C = 2, 4, 6, 8, 10$) by fitting a maximum of $M = 20{,}000$ trees. Given the learning rate and tree complexity, it is critical to determine the corresponding optimal number of trees to achieve a lower prediction error, and meanwhile prevent the over-fitting problem. In the gradient boosting process, we specified the maximum number of trees in the model. For the three signalized intersections, we obtained the corresponding optimal number of trees which the minimum error is achieved for each combination of learning rate and tree complexity. In this case, if the number of trees continues to increase, then the over-fitting issues will arise. The model performance based on different combination of regularization parameters are described in the following tables.

For the three intersections, the relationship among regularization parameters is shown in Figs. 6–8. It can be seen that increasing the value of shrinkage parameter with a given tree complexity will need fewer trees to achieve its minimum error. This is due to the fact that a higher value of shrinkage parameter can increase the contribution of each tree in the model, thereby needing fewer trees to be added. For a given shrinkage parameter, the corresponding optimal number of trees will

Table 1
Description of candidate predictor variables for drivers' SoR behavior.

Categories	Variables name	Variable description	Value set
Signal timing information	YS_A	Time left until signal changes to yellow	Continuous variable: **R+**
	YT_A	Portion of yellow time that elapsed before vehicle arrives at advance detector	Continuous variable: **R+**
Occupancy time	Occ_A	Occupancy time for the target vehicle when passing advance detector	Continuous variable: **R+**
	Occ_A1	Occupancy times for the nearest preceding vehicles	Continuous variable: **R+**
	Occ_A2	Occupancy times for the second preceding vehicles	Continuous variable: **R+**
	Occ_A3	Occupancy times for the third preceding vehicles	Continuous variable: **R+**
Time gaps	Gap_A	Time gap between the target vehicle and the nearest preceding vehicle	Continuous variable: **R+**
	Gap_A1	Time gaps for the nearest preceding vehicles	Categorical variable
	Gap_A2	Time gaps for the second preceding vehicles	Categorical variable
	Gap_A3	Time gaps for the third preceding vehicles	Categorical variable
Preceding vehicles' decision	Dec_A1	Vehicles' behaviors for the nearest preceding vehicles (1 = GLR; 2 = YLR; 3 = RLR)	Categorical variable
	Dec_A2	Vehicles' behaviors for the second preceding vehicles (1 = GLR; 2 = YLR; 3 = RLR)	Categorical variable
	Dec_A3	Vehicles' behaviors for the third preceding vehicles (1 = GLR; 2 = YLR; 3 = RLR)	Categorical variable
Adjacent lane	Adj_AA	Presence of running vehicles on the adjacent lane (1 = yes; 0 = no)	Categorical variable

Notes: GLR = green-light running, YLR = yellow-light running, RLR = red-light running.

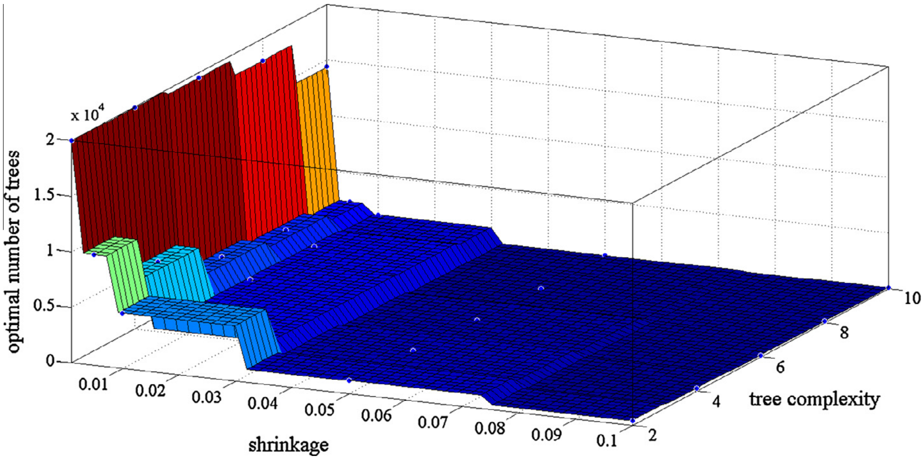


Fig. 6. Relationship among regularization parameters for Boone intersection.

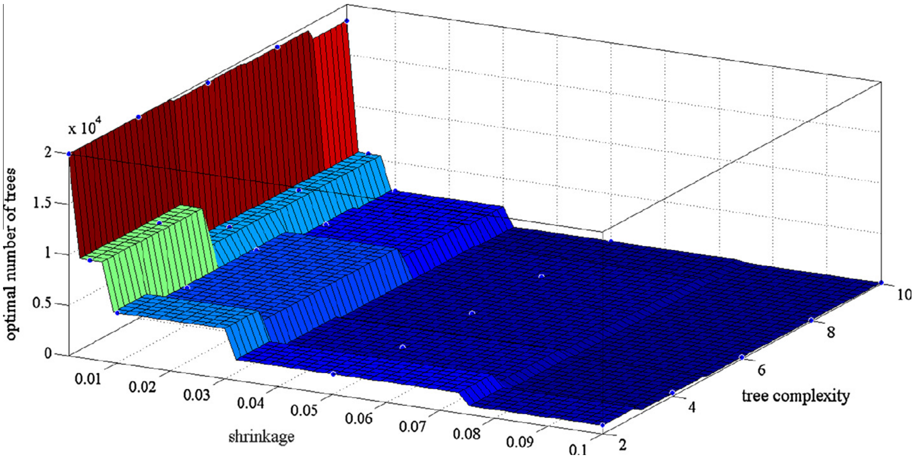


Fig. 7. Relationship among regularization parameters for Winnetka intersection.

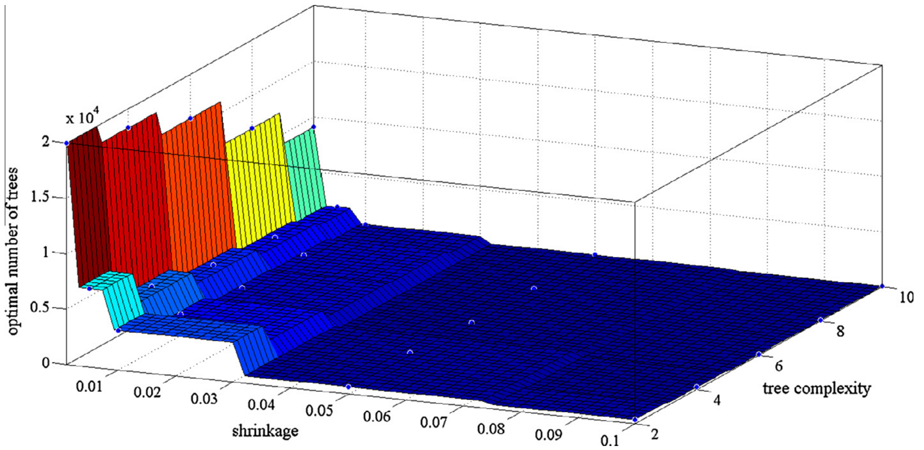


Fig. 8. Relationship among regularization parameters for Rhode Island intersection.

decrease as the value of tree complexity increases. This is due to the fact that increasing the value of tree complexity will generally capture more detailed information from the dataset and leads to a more complex model, thus requires fewer trees to achieve a minimum error. Furthermore, as shown in Tables 2–4, the results indicate that different combinations of shrinkage parameter and tree complexity with different optimal number of tress lead to different model performance. The computational time for the model depends on the three regularization parameters. Generally, a lower value of shrinkage parameter, and a higher value of tree complexity and number of trees, results in longer computational time. Therefore, there is a trade-off between the model performance and the computational time.

According to the combination of regularization parameters, the model performances for the Boone intersection, Winnetka intersection, and Rhode Island intersection are presented in Tables 2–4, respectively. By comparing the model results and

Table 2
Model performance for Boone intersection.

Shrinkage	R^2 and corresponding optimal number of trees (on training datasets)									
	Tree complexity = 2		Tree complexity = 4		Tree complexity = 6		Tree complexity = 8		Tree complexity = 10	
	R^2	Trees	R^2	Trees	R^2	Trees	R^2	Trees	R^2	Trees
0.10	0.6503	414	0.6611	245	0.6626	153	0.6827	117	0.6792	84
0.05	0.6506	1180	0.6926	754	0.6690	457	0.7072	219	0.6886	167
0.01	0.6510	4992	0.6532	1851	0.6641	1843	0.6729	1761	0.6811	1543
0.005	0.6511	9997	0.6947	6260	**0.7229**	**3576**	0.6795	2973	0.6789	2466
0.001	0.6395	20000[a]	0.6533	19919	0.6642	19527	0.6731	18013	0.6809	14362

Notes: Training sample size = 39,625.
[a] Indicates the optimal number of trees is larger than the given maximum value, and R^2 did not reach to its best value; numbers in bold are the best model performance with optimal number of trees.

Table 3
Model performance for Winnetka intersection.

Shrinkage	R^2 and corresponding optimal number of trees (on training datasets)									
	Tree complexity = 2		Tree complexity = 4		Tree complexity = 6		Tree complexity = 8		Tree complexity = 10	
	R^2	Trees	R^2	Trees	R^2	Trees	R^2	Trees	R^2	Trees
0.10	0.7100	827	0.7347	293	0.7440	156	0.7501	127	0.7407	86
0.05	0.7297	2014	0.7654	876	0.7541	564	0.7894	488	0.7441	298
0.01	0.6414	4952	0.7700	3700	0.7510	3698	0.7892	2512	**0.8193**	**2169**
0.005	0.6418	9804	0.7708	9743	0.7525	5621	0.7901	5528	0.7625	5475
0.001	0.6131	20000[a]	0.6480	19999[a]	0.6758	19673	0.7003	19463	0.7219	18376

Notes: Training sample size = 12,190.
[a] Indicates the optimal number of trees is larger than the given maximum value, and R^2 did not reach to its best value; numbers in bold are the best model performance with optimal number of trees.

Table 4
Model performance for Rhode Island intersection.

Shrinkage	R^2 and corresponding optimal number of trees (on training datasets)									
	Tree complexity = 2		Tree complexity = 4		Tree complexity = 6		Tree complexity = 8		Tree complexity = 10	
	R^2	Trees	R^2	Trees	R^2	Trees	R^2	Trees	R^2	Trees
0.10	0.7254	405	0.7240	242	0.7245	117	0.7247	84	0.7231	69
0.05	0.7283	651	**0.7312**	**611**	0.7278	244	0.7290	206	0.7291	184
0.01	0.7260	3565	0.7293	1953	0.7296	1214	0.7278	1079	0.7272	777
0.005	0.7275	7068	0.7301	4155	0.7290	2973	0.7300	2377	0.7295	2059
0.001	0.7234	20000[a]	0.7295	18301	0.7288	16047	0.7284	12043	0.7302	9070

Notes: Training sample size = 26,858.
[a] Indicates the optimal number of trees is larger than the given maximum value, and R^2 did not reach to its best value; numbers in bold are the best model performance with optimal number of trees.

computational time, the best model performance for the three intersections can be obtained. For the Boone intersection, the best performance is obtained at the shrinkage parameter of 0.005 and tree complexity of 6 with an optimal ensemble of 3676 trees. As to the Winnetka intersection, the best model performance occurs at the shrinkage parameter of 0.01 and tree complexity of 10 with an optimal ensemble of 2169 trees. With regards to the Rhode Island intersection, the model reaches its best performance with the shrinkage parameter of 0.05, tree complexity of 4, and optimal ensemble of 611 trees. The final R^2 for the three models with best performance are 0.7229, 0.8193, and 0.7312, indicating a good model fit. Theoretically, the gradient boosting model can handle different types of predictor variables, capture interactions among the predictor variables and fit complex nonlinear relationship (Elith et al., 2008). Hence, in this study the proposed gradient boosting logit model can handle the nonlinear relationship between drivers' SoR behavior at signalized intersection and its influential factors, thereby leading to superior prediction accuracy. The growing application of gradient boosting regression model in the field of travel time prediction (Zhang and Haghani, 2015) and highway crash prediction (Saha et al., 2015) has confirmed the advantage of gradient boosting method.

4.3. Stop-or-Run prediction

One important motivation of this study is to examine whether the loop detector information can be applied to predict the drivers' SoR behavior when the signal switches to the state of yellow. As to the gradient boosting logit model, the final approximation $F_M(x)$ is related to log-odds through Eq. (8), which can be inverted to calculate the probability of stop (p_{stop}) and run (p_{run}), as follows:

$$p_{stop}(x) = \hat{Pr}(y = -1|x) = \frac{1}{1 + e^{2F_M(x)}} \tag{20}$$

$$p_{run}(x) = \hat{Pr}(y = 1|x) = \frac{1}{1 + e^{-2F_M(x)}} \tag{21}$$

Traditionally, statistical logit regression approach can be used to predict the drivers' SoR behavior with the significant factors. To investigate the effectiveness of gradient boosting logit model employed for drivers' SoR behavior prediction, a comparison was conducted with the traditional statistical logit regression model. The estimated coefficients for the three intersections are shown in Table 5. As shown in Table 5, time to yellow start (YS_A), used yellow time (YT_A), occupancy time (Occ_A), time gaps (Gap_A, Gap_A1), preceding vehicle's decision (Dec _A2), and adjacent lane (Adj_AA) are all significant for the three intersection. The R^2 values for the three models are all lower than that of gradient boosting logit model. In the testing process, there are 9906, 3048, and 6714 test cases for the intersection of Boone, Winnetka, and Rhode Island, respectively. Based on the estimated model, the probability of stop (p_{stop}) and run (p_{run}) was calculated using the test data for the three intersections. Finally, a total of 7431, 2202, and 4691events were correctly estimated, respectively. The corresponding accuracy rates are 75.02%, 72.24%, and 69.87%. With regards to the gradient boosting logit model, it can correctly predict 8803, 2901, and 6051 events for the three intersections. The corresponding accuracy rates are 88.87%, 95.18%, and 90.13%. Tables 6 and 7 compare the two models' prediction performance based on the accuracy rate.

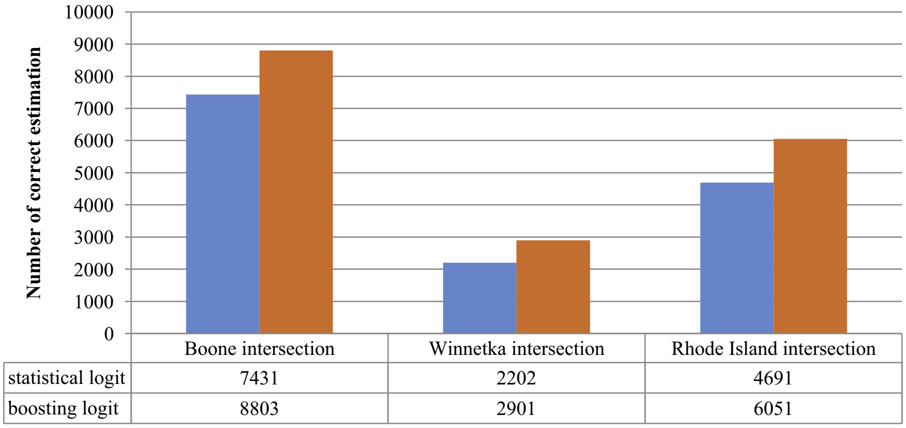
By comparing the results of traditional and proposed prediction techniques, we can see that the gradient boosting logit mode received the better predictive accuracy for all the three intersections. Specifically, for the Boone intersection, the predictive accuracy rate is 0.1385 higher than that from the traditional model. For the Winnetka intersection, the gradient boosting logit exhibits the best performance to predict drivers' SoR behavior with the highest predictive accuracy rate of 95.18%, which is 0.2294 greater than that from the traditional model. For the Rhode Island intersection, the proposed model also obtains a better predictive accuracy rate with the value of 90.13%. In overall, the gradient boosting logit model outperforms the traditional logit model. This finding confirms the advantage of the gradient boosting logit model in modeling complex relationship between drivers' SoR behavior and its determinants.

Table 5
Statistical binary logit regression model results.

Categories	Variables	Boone intersection		Winnetka intersection		Rhode Island intersection	
		Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value
Constant	Constant	−1.34	<0.001	3.49	<0.001	2.23	<0.001
Signal timing information	YS_A	0.79	<0.001	0.23	<0.001	0.04	<0.001
	YT_A	−1.78	<0.001	−1.38	<0.001	−2.18	<0.001
Occupancy time	Occ_A	−7.97	<0.001	−8.58	<0.001	−5.81	<0.001
	Occ_A1	−0.87	<0.001	0.08	0.71	0.07	0.24
	Occ_A2	0.06	0.56	−0.17	<0.05	0.15	0.38
	Occ_A3	0.04	0.51	0.54	<0.05	0.30	<0.05
Time gaps	Gap_A	−0.05	<0.001	−0.08	<0.001	−0.06	<0.001
	Gap_A1	−0.01	<0.001	−0.02	<0.001	−0.01	<0.001
	Gap_A2	−0.01	<0.001	−0.01	<0.05	−0.01	<0.05
	Gap_A3	−0.01	<0.001	−0.01	0.92	−0.01	0.95
Preceding vehicles' decision	Dec_A1	0.76	<0.001	0.75	<0.001	0.09	0.16
	Dec_A2	1.54	<0.001	0.60	<0.001	0.65	<0.001
	Dec_A3	1.23	<0.001	0.28	0.46	1.49	<0.05
Adjacent lane	Adj_AA	0.29	<0.001	0.67	<0.001	−0.08	<0.001
Model fit	R^2	0.6144		0.5243		0.4309	

Notes: "stop" is the base alternative; Dec_A1, Dec _A2, and Dec _A3 are treated as ordered category variables.

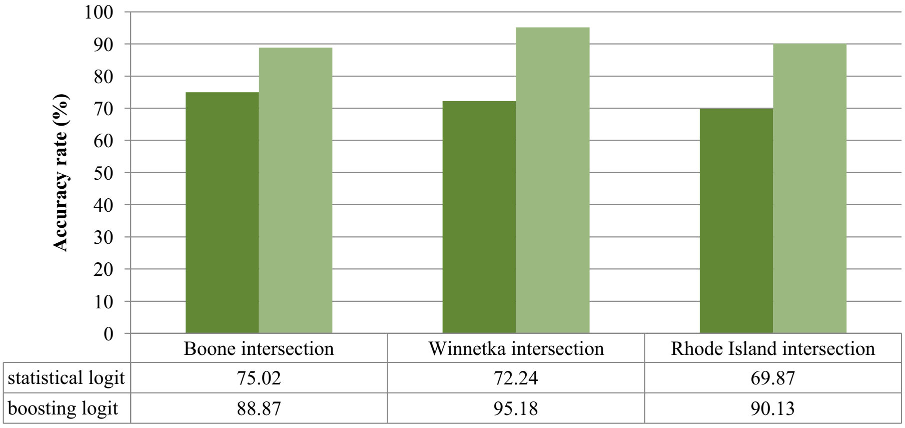
Table 6
Comparison of number of correct estimation for statistical and boosting logit models.



	Boone intersection	Winnetka intersection	Rhode Island intersection
statistical logit	7431	2202	4691
boosting logit	8803	2901	6051

Notes: For the intersections of Boone, Winnetka, and Rhode Island, the testing sample size is 9906, 3048, and 6714, respectively.

Table 7
Comparison of accuracy rate for statistical and boosting logit models.



	Boone intersection	Winnetka intersection	Rhode Island intersection
statistical logit	75.02	72.24	69.87
boosting logit	88.87	95.18	90.13

Table 8
Relative influence of predictor variables on drivers' SoR behavior.

Categories	Variables	Boone intersection		Winnetka intersection		Rhode Island intersection	
		Rank	Relative importance (%)	Rank	Relative importance (%)	Rank	Relative importance(%)
Signal timing information	YS_A	3	8.91	3	7.44	3	10.86
	YT_ A	1	72.85	1	61.69	1	57.42
Occupancy time	Occ_A	4	1.51	4	6.90	4	3.31
	Occ_A1	5	1.41	6	1.59	6	2.03
	Occ_A2	8	0.21	10	0.61	7	0.97
	Occ_A3	10	0.16	9	0.91	9	0.53
Time gaps	Gap_A	2	14.06	2	15.97	2	21.12
	Gap_A1	7	0.25	5	1.68	5	2.46
	Gap_A2	6	0.28	7	1.27	8	0.63
	Gap_A3	9	0.19	8	1.21	10	0.42
Preceding vehicles' decision	Dec_A1	11	0.05	12	0.12	12	0.04
	Dec_A2	12	0.05	14	0.01	13	0.03
	Dec_A3	14	0.01	13	0.04	14	0.02
Adjacent lane	Adj_AA	13	0.06	11	0.56	11	0.16

4.4. Influence ranking

Using the high-resolution traffic data, to explore the different influences of predictor variables on drivers' SoR behavior is another critical motivation of this study. The relative contributions of predictor variables for the three intersections were calculated based on the optimal models as shown in Table 8, respectively. A higher value of relative importance indicates stronger influences of predictor variables in predicting drivers' SoR behavior at signalized intersections. It should be noted that different with the typical method used in sensitivity analysis, which is to alter the value of one predictor variable and then estimate the change in output at one time (Saha et al., 2015), the gradient boosting logit model measures the influence of each factor on drivers' SoR behavior simultaneously, meanwhile accounting for the possible association between the factors (Zhang and Haghani, 2015; Saha et al., 2015; Chung, 2013).

As shown in Table 8, each predictor variable has different impact on drivers' SoR behavior. For all three intersections, the factor of used yellow time (YT_A) contributes most in predicting drivers' SoR behavior with a relative importance of 72.85%, 61.69%, and 57.42%, respectively. This finding is consistent with our expectation that the drivers' immediate decision on SoR is closely related with the yellow time that elapsed before vehicle arrives at advance detector. The factor of time gap (Gap_A) with a contribution of 14.06%, 15.97%, and 21.12% to the drivers' SoR behavior prediction, respectively, ranks the second most influential predictor variable for the Boone, Winnetka, and Rhode Island intersections. This result indicates that the time gap between the target vehicle and the nearest preceding vehicle has an important effect on drivers' SoR behavior prediction. The factor of time left until signal changes to yellow (YS_A), with a relative contribution of 8.91%, 7.44%, and 10.86%for the three intersections, respectively, is the third most influential variable. Followed by the factor of occupancy time (Occ_A), it contributes to drivers' SoR behavior prediction with a relative importance of 1.51%, 6.90%, and 3.31% for the three intersections, respectively. It can be seen that the target vehicle's velocity is highly related to the driver's decision on SoR. Another factor of occupancy time (Occ_A1) also contributes more than 1% to the model for all the three intersections (1.41%, 1.59%, and 2.03%, respectively), indicating that the nearest preceding vehicle's velocity also has played an importance role in drivers' SoR behavior. In addition, there are also some important factors that have more than 1% relative influence, but not for all intersections. Specially, as to the Winnetka intersection, the influences of the factors of time gap (Gap_A1, Gap_A2, and Gap_A2) contribute 1.68%, 1.27%, and 1.21% to the model output. We can see that the (time) distance between the target and the leading vehicles is associated with the drivers' SoR decision prediction. For the Rhode Island intersection, only the factor of time gap (Gap_A1) is closely related with the drivers' SoR decision with a relative contribution of 2.46%.

Another interesting finding relates to the influences of the factors of preceding vehicles' behavior (Dec_A1, Dec_A2, and Dec_A3) on the target vehicle's SoR decision. For all the three intersections, after considering the influences of signal timing information, occupancy time, time gaps, and adjacent lane, the preceding vehicles' decisions have weak effects on drivers' SoR decision prediction with contributions of less than 1%. The results show that the influence of the factor of presence of running vehicles on the adjacent lane (Adj_AA) is similar to that of preceding vehicles' decision, whose contributions are also less than 1% for all the three intersections. This finding indicates that drivers' SoR behavior is influenced by the vehicles in adjacent lane and preceding vehicles similarly.

5. Conclusions

This study contributes to improve drivers' stop-or-run behavior prediction at signalized intersection using high-resolution traffic and signal event data collected from loop detectors. The cases of first-to-stop, yellow-light running and red-light running are identified based on the information from both stop-bar and advance detectors. A related series of influential factors including signal timing information, occupancy time, time gaps, preceding vehicles' decision, and adjacent lane are measured in this study. Then, the gradient boosting logit model is proposed to handle different types of predictor variables, fit complex nonlinear relationships, and automatically disentangle interaction effects between influential factors. Three intersections are selected as the study cases. For each intersection, the models were built with various learning rates and tree complexities by fitting a maximum of trees. According to the different model performances under various combinations of regularization parameters, the optimal gradient boosting logit models were found by balancing the algorithm effectiveness, efficiency, and computational time.

By comparing with the traditional statistical logit regression model, this study shows that the gradient boosting logit model has superior performance in terms of prediction accuracy. The testing experiment indicates that the predictive accuracy rates of the models are as high as more than 88%. Meanwhile, the proposed models also have interpretation power, which is different from the traditional computational intelligence algorithms (e.g. SVM, neural networks, and random forest) as 'black-box' procedures. By applying the proposed method, the relative influences of predictor variables on drivers' stop-or-run behavior prediction can be identified based on the optimal model. It is greatly helpful to better understand the contribution of each related factor on drivers' stop-or-run behavior and its prediction.

The research shows the possibility of using loop detector data to predict drivers' stop-or-run decisions in real time. This work would be tremendously beneficial for signalized intersection safety improvement. A direct application would be to apply this method to avoid red-light running by adjusting all-red time when a vehicle is predicted as "run" and there is not enough yellow time for the vehicle to cross the intersection. It should be noted that there are also several issues need to be further investigated. Firstly, the data about driver individual information (e.g. gender, age, and habits) which may also have potential effects on driving behavior is not included in this study. Secondly, due to the complexity of red-light running behavior (e.g. rare event), more efforts should be made to investigate the red-light running behavior in future studies. Thirdly, it would be interesting to further test our model for intersections which are spread out.

Acknowledgement

This work is partly supported by the National Natural Science Foundation of China (71503018, and U1564212), the China Postdoctoral Science Foundation (2016M591045), and the Fundamental Research Funds for the Central Universities. The authors would like to thank the editors and anonymous reviewers for their valuable support and comments.

References

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, 38–47.

Bonneson, J.A., Zimmerman, K.H., 2004. Effect of yellow-interval timing on the frequency of red-light violations at urban intersections. Transp. Res. Rec.: J. Transp. Res. Board 1865, 20–27.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Chung, Y.S., 2013. Factor complexity of crash occurrence: an empirical demonstration using boosted regression trees. Accid. Anal. Prev. 61, 107–118.

De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88 (1), 243–251.

Ding, C., Ma, X., Wang, Y., Wang, Y., 2015. Exploring the influential factors in incident clearance time: disentangling causation from self-selection bias. Accid. Anal. Prev. 85, 58–65.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38 (4), 367–378.

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, 29–39.

Guelman, L., 2012. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst. Appl. 39 (3), 3659–3667.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Liu, H.X., Wu, X., Ma, W., Hu, H., 2009. Real-time queue length estimation for congested signalized intersections. Transp. Res. Part C 17 (4), 412–427.

Long, G., 2000. Acceleration characteristics of starting vehicles. Transp. Res. Rec.: J. Transp. Res. Board 1737, 58–70.

Lu, G., Wang, Y., Wu, X., Liu, H.X., 2015. Analysis of yellow-light running at signalized intersections using high-resolution traffic data. Transp. Res. Part A 73, 39–52.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C 54, 187–197.

NHTSA, 2012. NHTSA's Fatality Analysis Reporting System (FARS) Reports. National Highway Traffic Safety Administration, Washington DC.

Ren, Y., Wang, Y., Wu, X., Yu, G., Ding, C., 2016. Influential factors of red-light running at signalized intersection and prediction using a rare events logistic regression model. Accid. Anal. Prev. 95, 266–273.

Saha, D., Alluri, P., Gan, A., 2015. Prioritizing highway safety manual's crash prediction variables using boosted regression trees. Accid. Anal. Prev. 79, 133–144.

Schonlau, M., 2005. Boosted regression (boosting): an introductory tutorial and a Stata plugin. Stata J. 5 (3), 330–354.

Sheffi, Y., Mahmassani, H., 1981. A model of driver behavior at high speed signalized intersections. Transp. Sci. 15 (1), 50–61.

Smaglik, E., Sharma, A., Bullock, D., Sturdevant, J., Duncan, G., 2007. Event-based data collection for generating actuated controller performance measures. Transp. Res. Rec.: J. Transp. Res. Board 2035, 97–106.

Wu, X., Vall, N.D., Liu, H.X., Cheng, W., Jia, X., 2013. Analysis of drivers' stop-or-run behavior at signalized intersections with high-resolution traffic and signal event data. Transp. Res. Rec.: J. Transp. Res. Board 2365, 99–108.

Yang, C.D., Najm, W.G., 2007. Examining driver behavior using data gathered from red light photo enforcement cameras. J. Safe. Res. 38, 311–321.

Yu, R., Lao, Y., Ma, X., Wang, Y., 2014. Short-term traffic flow forecasting for freeway incident induced delays. J. Intell. Transp. Syst. 18 (3), 254–263.

Zhang, L., Zhou, K., Zhang, W., Misener, J.A., 2009. Prediction of red light running based on statistics of discrete point sensors. Transp. Res. Rec.: J. Transp. Res. Board 2128, 132–142.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transp. Res. Part C 58, 308–324.