STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION
TECHNICAL REPORT DOCUMENTATION PAGE
TR0003 (REV 10/98)

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA17-2646		

4. TITLE AND SUBTITLE
Evaluation of Signalized Intersection Safety Using Centracs System

5. REPORT DATE
11/01/2016

6. PERFORMING ORGANIZATION CODE

7. AUTHOR
Xinkai Wu, Sawanpreet Dhaliwal, Will Recker, Wenlong Jin, Xudong Jia

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
California State Polytechnic University, Pomona
3801 W Temple Ave,
Pomona, CA 91768

10. WORK UNIT NUMBER

11. CONTRACT OR GRANT NUMBER
65A0528 TO 003

12. SPONSORING AGENCY AND ADDRESS
California Department of Transportation
Division of Research, Innovation, and System Information
P.O. Box 942873, MS-83
Sacramento, CA 94273-0001

13. TYPE OF REPORT AND PERIOD COVERED
Final 10/2013-4/2017

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT
This research aims to explore the possibility of using high-resolution data such as sec-by-sec traffic signal data provided by the Centracs system or event-based data provided by SMART-SIGNAL (Systematic Monitoring of Arterial Road Traffic and SIGNAL) system, to evaluate intersection safety. Traditional methods, either using historical crash data collected from infrequently happened collisions, or potential conflicts estimated from microscopic simulation which assumes “accident-free”, cannot provide accurate and timely evaluation of intersection safety. In contrast, this research proposes a comprehensive intersection safety evaluation system, which is able to quantify the safety performance of signalized intersections by using high-resolution traffic signal data collected from existing loop detection systems.
This research proposes an innovative method to identify the emerging and impending hazardous situations including red-light running (RLR) violations and potential traffic conflicts, which essentially indicate the safety level of an intersection. The proposed method first estimates the drivers’ decision to stop-or-run (SoR) by developing a simple method to identify first-to-stop (FSTP), yellow-light running (YLR), and red-light running (RLR) cases using high-resolution data. By applying a binary logistical regression model, this research found out that occupancy time, time gap, used yellow time, time left to yellow start, time gap between the first two preceding vehicles, and decisions of preceding and surrounding vehicles show significant impacts on drivers’ decisions. Furthermore, due to the rare events nature of RLR, a modified rare events logistic regression model was developed for RLR prediction. The results showed that the rare events logistic regression model performed significantly better than the standard logistic regression model.

17. KEY WORDS
Red light running, stop-or-run, traffic conflicts, rare events, logistic regression, high-resolution traffic data, intersection safety.

18. DISTRIBUTION STATEMENT
No restrictions.

19. SECURITY CLASSIFICATION *(of this report)*	20. NUMBER OF PAGES	21. COST OF REPORT CHARGED
Unclassified	66	

DISCLAIMER STATEMENT

University of California Transportation Center

Evaluation of Signalized Intersection Safety Using Centracs System

Xinkai Wu, Ph.D.
Civil Engineering
California State Polytechnic University, Pomona

Sawanpreet Singh Dhaliwal, Graduate Student
Civil Engineering
California State Polytechnic University, Pomona

Will Recker, Ph.D.
Civil and Environmental Engineering
University of California, Irvine

Wenlong Jin, Ph.D.
Civil and Environmental Engineering
University of California, Irvine

Xudong Jia, Ph.D.
Civil Engineering
California State Polytechnic University, Pomona

November 2016

Caltrans report number CA17-2646

University of California
Transportation Center

ABSTRACT

This research aims to explore the possibility of using high-resolution data such as sec-by-sec traffic signal data provided by the Centracs system or event-based data provided by SMART-SIGNAL (Systematic Monitoring of Arterial Road Traffic and SIGNAL) system, to evaluate intersection safety. Traditional methods, either using historical crash data collected from infrequently happened collisions, or potential conflicts estimated from microscopic simulation which assumes "accident-free", cannot provide accurate and timely evaluation of intersection safety. In contrast, this research proposes a comprehensive intersection safety evaluation system, which is able to quantify the safety performance of signalized intersections by using high-resolution traffic signal data collected from existing loop detection systems.

This research proposes an innovative method to identify the emerging and impending hazardous situations including red-light running (RLR) violations and potential traffic conflicts, which essentially indicate the safety level of an intersection. The proposed method first estimates the drivers' decision to stop-or-run (SoR) by developing a simple method to identify first-to-stop (FSTP), yellow-light running (YLR), and red-light running (RLR) cases using high-resolution data. By applying a binary logistical regression model, this research found out that occupancy time, time gap, used yellow time, time left to yellow start, time gap between the first two preceding vehicles, and decisions of preceding and surrounding vehicles show significant impacts on drivers' decisions. Furthermore, due to the rare events nature of RLR, a modified rare events logistic regression model was developed for RLR prediction. The results showed that the rare events logistic regression model performed significantly better than the standard logistic regression model.

This research further proposes an intersection safety evaluation system which evaluates the overall intersection safety based on a combination of the potential traffic conflicts and red-light-running cases. The proposed system has been applied to an intersection in Anaheim, CA operated using Econolite Centracs System and a corridor with 3 intersections located at Minneapolis, MN using SMART-SIGNAL System. The proposed system can be applied to build a real-time intersection collision avoidance system. Also, the intersection safety information collected from a network-wide system can be used to rank the safety severity for all intersections in the network, therefore helps agencies identify the intersections which need most improvement. This research is expected to contribute significantly to the improvement of intersection safety, and build a foundation for future dynamic systems that could alert drivers of emerging and impending hazardous situations.

ACKNOWLEDGEMENTS

The authors are very grateful to Mr. John Thai, the Principal Transportation Engineer at City of Anaheim, CA for providing Centracs System data and offering excellent advices throughout this project.

This work was supported by Caltrans contract 65A0528 Task Order 003.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

An intersection is a planned point of conflict in the roadway system. With different crossing and entering movements by both drivers and pedestrians, an intersection is one of the most complex traffic situations that motorists encounter. Crashes often occur at intersections because of the potential conflicts between left-turning, crossing, and right-turning movements. This makes intersection safety a national, state, and local priority [1]. Based on the Fatality Analysis Reporting System (FARS) and National Automotive Sampling System-General Estimates System (NASS-GES) data, approximately 40 percent of the estimated 5,811,000 crashes that occurred in the United States in 2008 were intersection-related crashes [2]. A majority of intersection-related crashes occurred due to careless and reckless driving behavior such as Red Light Running (RLR) that has resulted in a substantial number of severe injuries and significant property damage. According to the National Highway Traffic Safety Administration's (NHTSA) Traffic Safety Facts 2008 Report [3], there were more than 2.3 million reported intersection-related crashes, resulting in more than 7,770 fatalities. Of these, 762 were caused by red-light running, and red-light runners as reported by FARS injured an estimated 165,000 people annually [2]. As further amplified by the National Survey of Speeding and Other Unsafe Driver Actions [4], 97% of drivers feel that other drivers running red-lights are a significant safety threat. This number is increasing at more than three times the rate of growth for all other fatal crashes [5, 6]. In China, a study shows that RLR had caused over 4227 severe injury crashes and 789 fatalities based on the data collected from Jan. 2012 to Oct. 2012 [7]. Despite improved intersection designs and more sophisticated applications of traffic engineering measures, the annual toll of human loss due to motor vehicle crashes has not substantially changed in the past ten years [8].

Much research has been conducted to study intersection safety, and a significant portion is related to drivers' stop-or-run (SoR) behavior such as red-light running (RLR) and yellow-light running (YLR). For RLR prevention, many methods, such as variable message warning signs, signal timing adjustment, or even autonomous vehicle technologies [9, 10, 11, etc.] have been developed. But to apply these methods, a critical first step is to predict potential SoR behavior of the driver. Notice, RLR in most of the situations is not an intentional decision, but an erratic behavior due to uncertain surrounding conditions such as traffic light switching to yellow, or following a platoon [12]. Many organizations such as the Federal Highway Administration (FHWA), the Institute of Transportation Engineers (ITE), American Association of State Highway and Transportation Officials (AASHTO), AAA, etc. are devoting substantial resources to help reduce intersection crashes. A significant amount of effort has been focused on the assessment of intersection safety. The most common method of assessing safety at an intersection is analyzing its collision history. For example, all eight intersection safety assessment methods suggested by FHWA (collision frequency, collision rate, combined collision frequency and rate, equivalent property damage only method, critical collision rate, risk analysis methods, safety performance functions, and empirical Bayes method) use historical collision data [13]. Also, the Highway Safety Manual (HSM) safety predictive method proposed by AASHTO [14] also relies on historical crash data. However, given the infrequent and random nature of crashes, crash-data-based methods are slow to reveal the need for remediation of either the roadway design or the flow-control strategy, not to mention the real-time safety assessment.

1.1 Datasets

Different data sets have been used to study RLR crashes and red-light runner behaviors. For example, Retting et al. [15] used the data collected from the Fatality Analysis Reporting System (FARS) and General Estimates System (GES) to analyze the vehicle and driver characteristics of red-light runners. Wissinger et al. [16] used focus groups to investigate the attitudes, beliefs, and perceptions of the public toward RLR and red light cameras; and Porter & Berry [17] conducted a national telephone survey to understand those who had run the red light.

To fully understand SoR behavior, data collection plays a vital role. Most of the current research uses the off-line data collected from video cameras. For example, in Bonneson et al.'s before-after study [18], they used videotape records along with other methods like laser speed guns to collect RLR data to analyze the impact of factors such as yellow interval timing, on the frequency of red-light violations, [19]. Gates et al. [20] temporarily installed consumer-grade video cameras at four high-speed and two low-speed intersections to collect driver behavior in dilemma zones. David & Najm [21] examined red light violation behavior using about 47,000 violation records that were captured by photo enforcement cameras from 11 signalized intersections in the city of Sacramento, California, over a four-year period. Some research identified that safety belt use, driving records, ethnicity, etc. were critical for RLR [6, 22]. But apparently, these factors are difficult to determine in real-time and cannot be used for real-time RLR prediction.

Some research analyzed high-quality video data and found out that the vehicle characteristics (such as vehicle speed) and traffic operations (such as flow and signal timing) had a significant impact on RLR [18, 20, 23, 24, 25, 26, 27]. However, using video data for RLR prediction have limitations since the quality of the data is constrained by the quality of cameras and, more importantly, real-time video data analysis is time-consuming and costly. This limitation causes an obstruction for implementation of dynamic systems that can inform or alert drivers of emerging and impending hazardous situations like RLR in real time. Building such dynamic systems has become one of the primary focuses of many federal and local agencies recently. For example, the U.S. Department of Transportation (DOT) and state DOTs have formed several programs to investigate collision avoidance systems such as Intersection Decision Support

Program (IDSP) [9, 10] and the Cooperative Intersection Collision Avoidance System (CICAS) [11].

In contrast to video data collection, loop detector data can be quickly and automatically obtained in real time with low cost. Since majority of signalized intersections have been equipped with loop detectors for signal operations (Figure 1.1), using loop detector data to help measure and improve intersection safety becomes very attractive, as pointed out by Zhang et al. [26]. In their research, Zhang et al. used multiple discrete point sensors to collect data and then predict RLR events. However, the general aggregate data (30-sec, 5-min or even 15-min) ignores details and is too coarse to describe individual drivers' behaviors in depth at signalized intersections. Due to this reason, little research has been conducted using loop detector data to analyze SoR behavior. But with recent improvements in data aggregation methods [28, 29], high-resolution traffic and signal event data can be easily collected at signalized intersections, either from a traffic controller [28] or directly measured from the back panel in a traffic cabinet [29]. High-resolution traffic data (event-based or second-by-second data) provides detailed vehicle arrivals and departures from loop detectors. This data, combined with signal phase changes, could be used to derive vehicle trajectories, which can serve as the foundation for traffic conflict analysis. Moreover, since loop detector data can be easily and automatically obtained in real time with low cost, this could significantly contribute to the implementation of dynamic systems that could inform or alert drivers of emerging and impending hazardous situations. The safety evaluation system developed in this research could be a critical first step in the implementation of IDSP and CICAS for identifying safety hotspot for treatment.

High-resolution traffic and signal event data has great potential to improve intersection safety. But so far, only very preliminary research has been done by Chatterjee & Davis [30], who used event data to identify incidents and help reconstruct car crashes. Therefore, it is necessary to investigate the potentials of applying event data for intersection safety study.

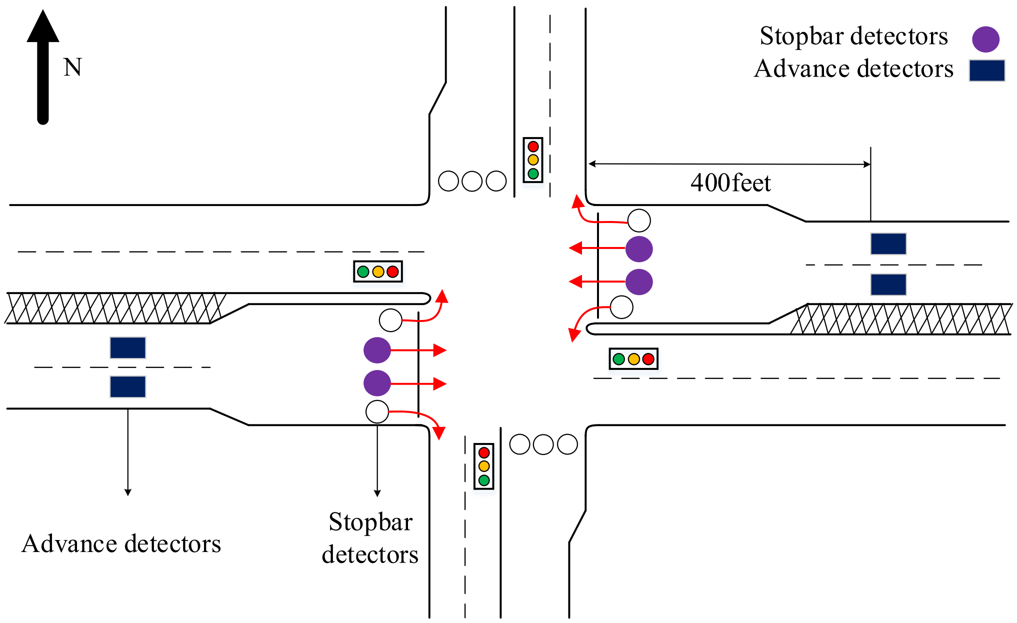


Figure 1.1 A typical detector layout (as shown in the figure, the advance detector is the loop detector typically located 400 feet upstream from the stop-line; and stop-bar detector is located right behind the stop-line).

1.2 Models

Different models have been developed to describe the probability of occurrence of SoR event. For example, Sheffi & Mahmassani [31] developed a probit stopping model by assuming that a driver's decision to stop is normally distributed, while Bonneson et al. [32] used the logistic regression (logit) by assuming that a driver's decision to stop follows a logistic distribution.

For RLR prediction, a general way is to first analyze a large amount of traffic data (video data or loop detector data) to statistically identify the factors that may significantly impact the RLR behavior. From a statistical point of view, a driver's current driving conditions, together with surrounding traffic conditions and signal timing situations, will directly or indirectly lead to the driver's later behavior of RLR. Therefore, by statistically analyzing a large amount of traffic data, the inner correlation between all these impact factors and drivers' RLR behaviors can be derived. Then, based on the obtained relationship between the impact factors and drivers' RLR behaviors, either a Probit [31, 32] or Logit [33, 34, 35] model can be applied to predict RLR.

A critical problem for RLR prediction is that traditional statistical methods like Probit or Logit have difficulties in predicting RLR accurately. These traditional methods work well for Yellow-Light-Running (YLR) prediction [36]. However, when applying these methods for RLR prediction, the results are not satisfactory. A traffic collision is a random event, and the major factor of an accident is drivers' mistake. According to the crash factor report done by the National Highway Traffic Safety Administration [37], out of 787,236 intersection-related crashes, about 96 percent (756,570 crashes) had critical reasons attributed to drivers. These include inadequate surveillance, false assumption of other's action, turn with obstructed view, illegal maneuver, internal distraction, and misjudgment of the gap or other's speed. Such problem has been reported as an imbalanced class problem [38, 39]. While conducting this research, it was revealed that out of 42277 observations (including Red Light Running, Yellow Light Running, and First-to-Stop (FSTP, defined as the first vehicle which stops before the stop line when green ends)) collected from three months' data at one signalized intersection, only 289 cases (0.7%) were RLR. With such small number of RLR, applying standard classifiers, such as logistic regression, will sharply underestimate the probability for RLR [40].

Therefore, the traditional crash-data-based safety assessment method, which aims to investigate the relationship between crashes and other factors such as average annual daily traffic (AADT), intersection geometry design, etc. may ignore the most critical factor, i.e. the driving behavior, which can be described by the trajectories of vehicles. The traffic conflict technique has been developed since the 1960s to relate intersection safety to drivers' actions. This method is based on the understanding that conflict frequency is correlated with the risk of actual collision [41, 42]. The traffic conflict technique analyzes the frequency and character of narrowly averted vehicle-to-vehicle collisions using the information of vehicles' trajectories and therefore does not need historical crash data. Traditional conflict studies utilize trained personnel to identify and record conflicts observed at an intersection, so this method is time-consuming and expensive. To overcome the shortcoming, Gettman et al. developed the Surrogate Safety Assessment Model (SSAM), which combines microsimulation and automated conflict analysis, to assess the safety of traffic facilities without waiting for a statistically above-normal number of crashes and injuries to occur [43]. However, traffic conflicts estimated from SSAM are questionable because vehicle trajectories in this method are derived from micro-simulation which is assumed "accident-free." Proposed system in this research, on the other hand, is based on loop detector data, instead of historical crash data or data obtained by traffic simulation.

1.3 Objectives

This research aims to address above issues related to RLR by first exploring the influential factors which have significant impacts on drivers' RLR behaviors using loop detector data. Mainly, RLR events extracted from high-resolution traffic data collected by loop detectors from three signalized intersections are utilized to identify the factors that significantly affect RLR behaviors. Then based on the data analysis results, and also considering the rare events nature of RLR, this research proposes a modified rare events logistic regression model originally developed by King and Zeng to predict RLR. King and Zeng's rare events logistic regression model has been applied in many fields and shows impressive performance; but so far according to our limited knowledge, none of the previous research has used this method to predict RLR.

The proposed model is further evaluated using high-resolution traffic data collected by loop detectors. The results demonstrate that the new method outperforms the standard logistical regression method with a significant improvement of RLR prediction rate.

The research also aims to quantify the safety performance of signalized intersections and identify the emerging and impending hazardous situations. Such information is much needed for the implementation of a real-time intersection collision avoidance system. Also, the intersection safety information collected from a network-wide system can be used to rank the safety severity for all intersections in the network, therefore helps agencies identify the intersections which need the most improvement. To achieve this goal, this research proposes a comprehensive intersection safety evaluation system which fulfills the following two primary functions:

1) **Predict red-light violations.** The primary function of the proposed system is to identify possible red-light violations, which is a major factor that leads to traffic conflicts. To fulfill this function, we first use high-resolution data collected from advanced detectors, which are located several hundred feet behind stop-line, to predict drivers' decision of STOP-or-RUN (SoR) at the onset of amber phase. A prediction model will be applied to estimate drivers' SoR decisions. After we predict drivers' SoR decision, we can determine red-light violations by combining signal phasing information. For example, if a vehicle is predicted as "RUN" and the remaining amber time is not enough for the vehicle to cross the intersection, this car will end up "red-light violation.". The number of red-light-running violations will provide some indication of intersection safety. Moreover, this model can be applied to predict red-light violation in real time if using real-time information collected from advanced detectors. The real-time red-light violation information, combined with appropriate control strategies like the all-red extension, could be used to reduce potential severe crashes caused by red-light running.

2) **Estimate potential traffic conflicts based on real traffic conditions.** This feature will focus on determining both rear-end and crossing (i.e. right-angle) traffic conflicts. The rear-end conflicts include the same direction conflicts from all approaches, and the crossing traffic conflicts mainly are the conflicts between through movement and conflicting left-turns. To estimate the potential conflicts, vehicles' trajectories will be determined using high-resolution data. Then, the traffic conflict technique will be applied to calculate both types of traffic conflicts at intersections. The overall intersection safety will be evaluated based on a combination of the probability of two types of potential conflicts.

Since the proposed system can be easily applied to most of the signalized intersections equipped with loop detectors, it can provide spatial information of intersection safety for a large area; therefore, provides agencies with a risk assessment tool to prioritize the intersections which need the most attention. More importantly, this research will build a foundation for the future development of dynamic systems that could inform or alert drivers of emerging and impending hazardous situations.

Chapter 2

High Resolution Data

A conflict or Stop-or-Run event at an intersection, as discussed in the previous chapter, can be estimated from the data collected by loop detectors. The proposed regression models are based on the correlation between drivers' behaviors (i.e. RLR, YLR, or FSTP) and impact factors including velocity, time gaps (the time difference between the arrival time of the following vehicle and the departure time of the leading vehicle), etc. To study the driver's behavior for traffic conflict estimation and to develop the prediction model for red-light-running, three steps are involved in the data preparation:

1) Collect high-resolution traffic and signal event data;
2) Identify a driver’s decisions including red-light running (RLR), yellow-light running (YLR), or the first-to-stop (FSTP) during yellow using stop-bar detectors; and
3) Match events between stop bar and advance detectors since the information collected from advance detectors will be needed for the investigation.

We know that data collection is the most crucial step involved in any research as quality of data determines the quality of analysis and hence, the results. As discussed earlier, this research requires high-resolution data for estimating the trajectories of vehicles which in turn are able to explain the driving behavior. Two types of high-resolution data that can be used in this research are:

1. SMART-SIGNAL Data (Event data)
2. Econolite Centracs Data (Sec-by-Sec data)

Both types of data are perfect to estimate drivers’ behavior but this research has been done using SMART-SIGNAL data. Both datasets are valid for this research as both of these data are inter-convertible.

2.1 High-Resolution Traffic and Signal Event Data from SMART-SIGNAL

High-resolution traffic event data were gathered by the SMART-SIGNAL (Systematic Monitoring of Arterial Road Traffic and SIGNAL) system developed at the University of Minnesota [29]. The SMART-SIGNAL is capable of continuously collecting and archiving high-resolution event-based vehicle-detector actuation and signal phase change data. The SMART-SIGNAL system has been installed on a major arterial (Trunk Highway 55) at six intersections in the Twin Cities area since Jul. 2008. All intersections are equipped with vehicle-actuated signals, with advance detectors typically located 400 feet upstream from the stop-line for the green extension on the major approach and stop-bar detectors located right behind the stop line for presence detection on the minor approach. For research purposes, we have also installed stop-bar and link entrance detectors on the major approaches. A total of 21 months of data were used for analysis, amounting to well over 50 million recorded events. In this research, we used the event data collected from three intersections (Boone Ave., Winnetka Ave., and Rhode Island Ave., see Fig. 2.1 for the detector configuration.)

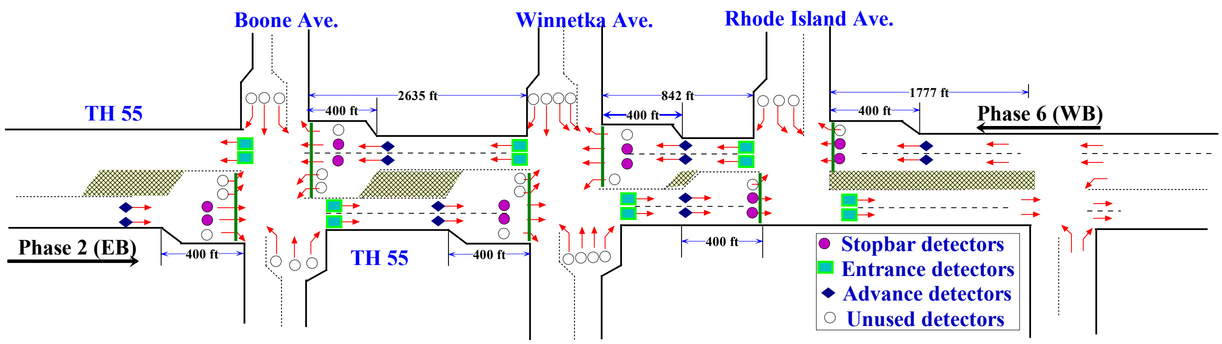


Figure 2.1 Study area and intersections (SMART-SIGNAL).

Data collected by SMART-SIGNAL is a combination of already existing loop detector data and signal status data. The data is collected from individual intersections by the traffic cabinets at the intersections and instantly sent to a database for further processing in a lab. This processed data made available to general users can explain some traffic behaviors that were difficult to explain by using only the loop detector data. A basic architecture of the SMART-SIGNAL system is shown in figure 2.2.

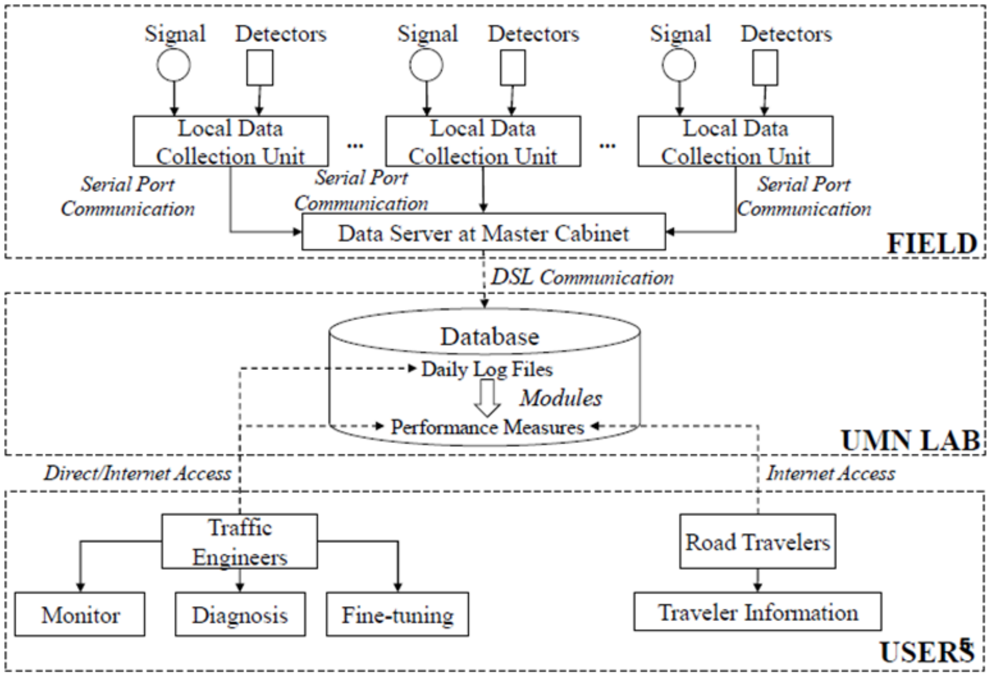


Figure 2.2 SMART-SIGNAL architecture.

The detector and signal data can be collected and joined using some additional hardware such as Traffic Controller Interface Devices (CID), and Traffic Event Recorder, along with existing traffic signal system. These devices can be installed easily in the existing cabinets as shown in figure 2.3.

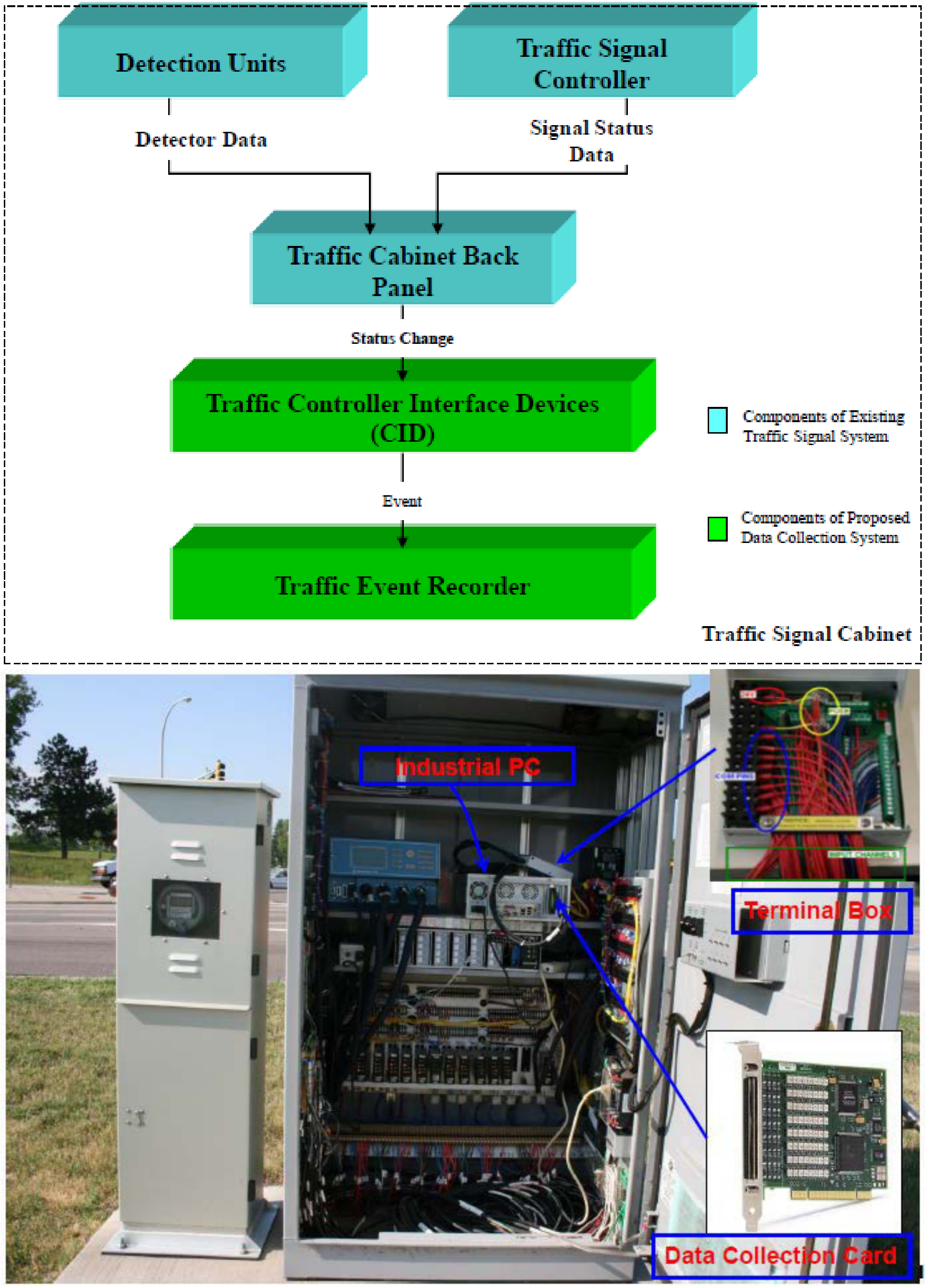


Figure 2.3 Changes made to existing Traffic Signal Cabinets

After the processing of raw data, high-resolution event-based data is generated by the system that gives detailed information about the actuation of detectors and signal phase change with exact timing associated with it. The data provide details for individual events such as the time at which a particular detector, say D1, was turned on or off by a vehicle or time at which a particular signal phase, say G3, was turned on or off and the duration associated with it. The sample format of data can be seen in figure 2.4.

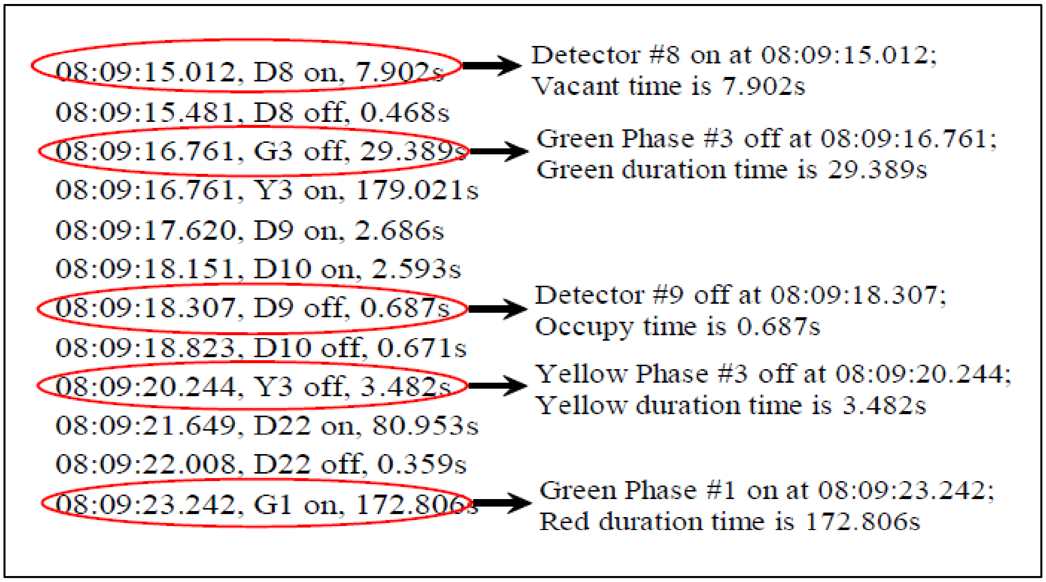


Figure 2.4 Event-based Data Format.

2.2 Sec-by-Sec Centracs data from Econolite Centracs system

Recently, the Centracs system, introduced by Econolite, has the capability to automatically collect and archive sec-by-sec traffic and signal data, which can be used in this research to evaluate intersection safety. More importantly, since Centracs is a robust and cost effective system for improving intersection efficiency, it is very possible that Centracs will be installed in many intersections in the near future. Indeed, Centracs has been deployed by many cities in CA. For example, 9 Orange County cities have adopted the Centracs system since 2011. Data from one such site in City of Anaheim has been used for this project as shown in figure 2.5.

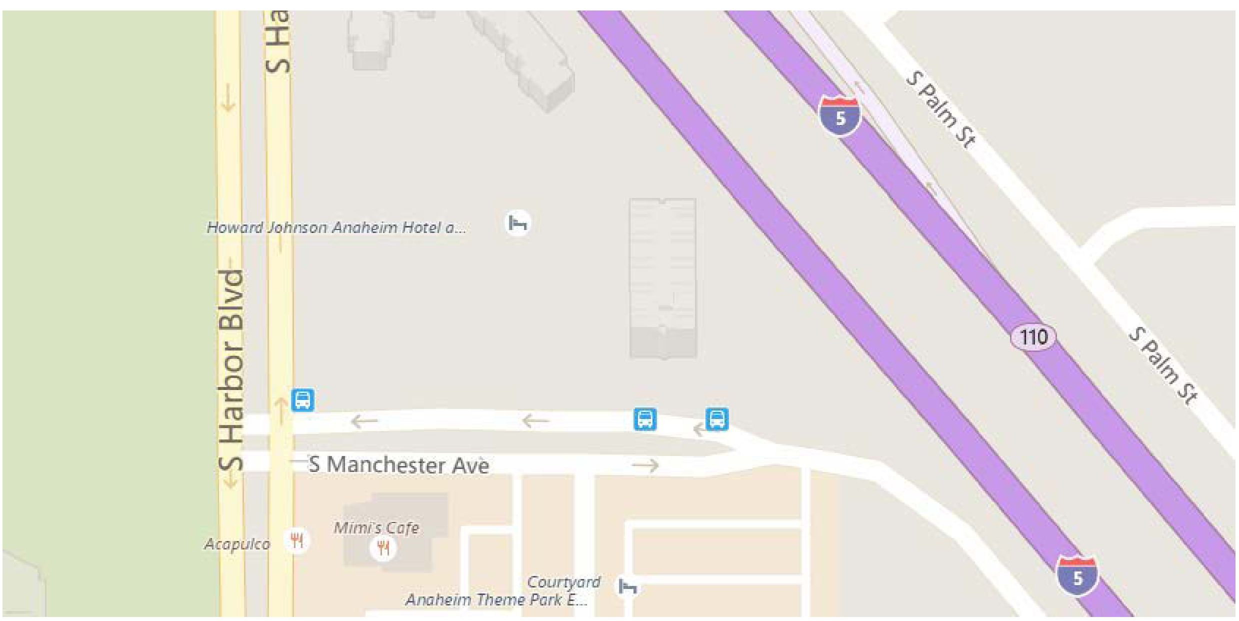


Figure 2.5 Study Area and intersections (Centracs). (Source: Google)

Econolite Centracs System collected sec-by-sec data that included traffic data (vehicle actuations from loop detectors) and signal data (phase data from traffic controllers). The sec-by-sec data (including volume and occupancy) essentially indicates the times when a vehicle arrives at and departs from detectors. Therefore, the raw data can be used to estimate individual vehicle speed and time gap between two consecutive vehicles. Individual vehicle speed is estimated using a calibrated effective vehicle length divided by the occupancy time (i.e. the time that the detector is occupied by a vehicle), and time gap is the time difference between the time when the leading vehicle leaves the detector and the time when the following vehicle arrives at the detector. Time-stamped state changes of ‘phases in use’ and ‘detectors in use’ are stored in binary format on the controller in hourly data files. Data are retrieved from the controller by using a transmission control protocol/Internet protocol network connection, with the controller serving as a file transfer protocol (FTP) server. Once downloaded from the controller, an application is used to convert the data files into comma-separated-value (CSV) files for use in producing the appropriate reports. A sample output file is shown in figure 2.6.

Date	Time	dSec	phs-1	phs-2	phs-3	phs-4	phs-5	phs-6	phs-7	phs-8	det-1	det-2	det-3	det-4	det-5	det-6
11/3/2015	0:00:01	0												-1		
11/3/2015	0:00:01	0.5														-1
11/3/2015	0:00:02	0.3													-1	
11/3/2015	0:00:02	0.5													1	
11/3/2015	0:00:03	0														
11/3/2015	0:00:04	0														
11/3/2015	0:00:04	0.4														
11/3/2015	0:00:04	0.5					3								3	
11/3/2015	0:00:05	0.2														
11/3/2015	0:00:05	0.3														
11/3/2015	0:00:05	0.6										1				
11/3/2015	0:00:05	0.8										-1				
11/3/2015	0:00:06	0.3										1				
11/3/2015	0:00:06	0.5														
11/3/2015	0:00:06	0.9														
11/3/2015	0:00:07	0.4														
11/3/2015	0:00:08	0.4					5								-2	
11/3/2015	0:00:08	0.5						1			1			1	-2	-2
11/3/2015	0:00:08	0.8														
11/3/2015	0:00:09	0.6										-1				
11/3/2015	0:00:09	0.9														
11/3/2015	0:00:10	0.2										1				
11/3/2015	0:00:10	0.3														
11/3/2015	0:00:10	0.5										-1				
11/3/2015	0:00:11	0.2										1				
11/3/2015	0:00:11	0.6														
11/3/2015	0:00:11	0.8														
11/3/2015	0:00:14	0.5										-1				
11/3/2015	0:00:14	0.9														
11/3/2015	0:00:18	0.4														
11/3/2015	0:00:22	0.4		3				3				3				-4
11/3/2015	0:00:23	0.4		5				5				-2				-2

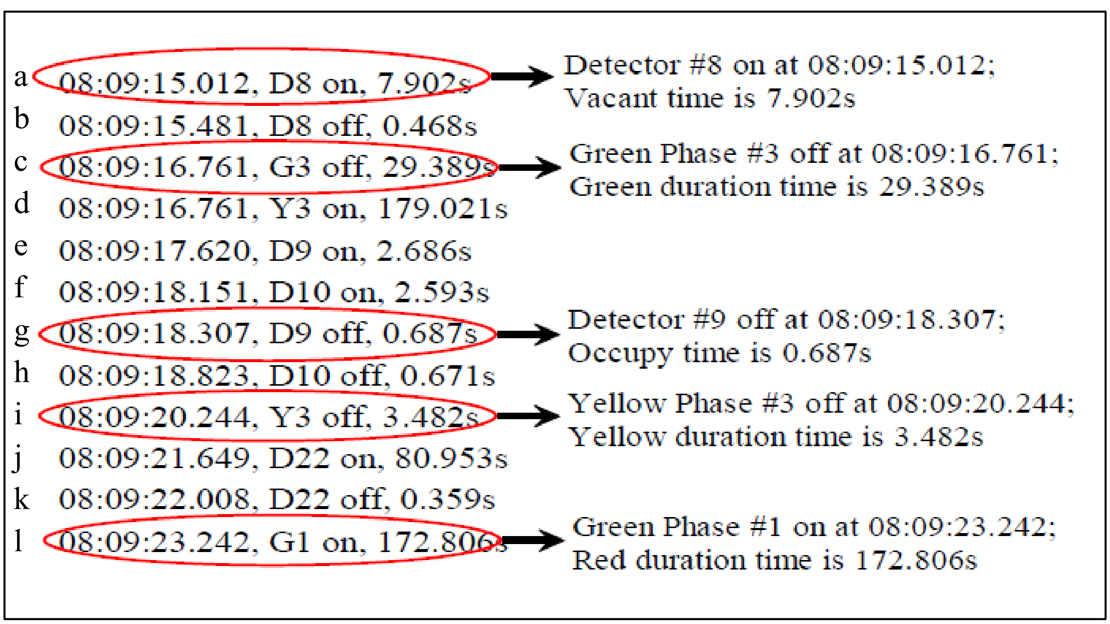
Figure 2.6 Sec-by-sec data format

The file should be interpreted as follows:

For the Phases, the 1, 3, 5 values respectively represent the start of green, yellow and red for the phase indicated in the column heading. The interpretation of the detector values represents the state changes of the detectors. Since multiple state changes can happen in a single time-slice, the states are added up with the counts alternating between positive and negative to indicate whether the last reported state change was ‘on’ (positive) or ‘off’ (negative). As you will observe, there will be cases when you see a ‘1’ followed by a ‘-1’ in a subsequent time-slice. In cases where you observe a ‘-3’ for example, you will know that the signal went from ‘off’ to ‘on’ to ‘off’ within that time-slice. It must be noted that the sign of the value indicates the final state of the detector in the time-slice. A negative value means the vehicle left the detector and thus was ‘off’ at the end of the reporting period. A positive value means that the detector was reporting a vehicle present on the detector at the end of the time-slice. Thus, a ‘-2’ indicates that the last event was an ‘off’ which means it went from ‘on’ to ‘off’ in the time-slice and a value of ‘2’ indicates the detector went from ‘off’ to ‘on’.

2.3 Conversion of sec-by-sec Centracs data to event-based data

Many agencies have started collecting high-resolution data and these two data types have been widely accepted in the past few years. The biggest advantage of these two data types is that they can be converted into each other. This allows application of this research on a large scale. The interconversion of data is explained using an example as follows:



(a)

Record	Date	Time	dSec	phs-1	phs-3	det-8	det-9	det-10	det-22
1	11/3/2015	8:08:46	0.6		1				
2	11/3/2015	8:09:08	0.1			-1			
3	11/3/2015	8:09:15	0			1			
4	11/3/2015	8:09:15	0.5			-1			
5	11/3/2015	8:09:16	0.8		3				
6	11/3/2015	8:09:17	0.6				1		
7	11/3/2015	8:09:18	0.2					1	
8	11/3/2015	8:09:18	0.3				-1		
9	11/3/2015	8:09:18	0.8					-1	
10	11/3/2015	8:09:20	0.2		5				
11	11/3/2015	8:09:21	0.6						1
12	11/3/2015	8:09:22	0						-1
13	11/3/2015	8:09:23	0.2	1					

(b)

Figure 2.7 Interconversion of (a) SMART-SIGNAL data and (b) Centracs data.

Consider the SMART-SIGNAL data (figure 2.7 (a)) that has been converted to Centracs format (figure 2.7 (b)). For the convenience, SMART-SIGNAL and Centracs records has been numbered as a, b, c, etc. and 1, 2, 3, etc., respectively.

- Record a – At 08:09:15.01, detector 8 turned ON after 7.9 sec. The event of turning ON is represented as "1" (record 3) and vacant time can be calculated by subtracting ON timestamp and previous OFF timestamp represented by "-1" (record 3 – record 1).
- Record b – At 08:09:15.48, detector 8 turned OFF after 0.46 sec. The event of turning OFF is represented as "-1" (record 4) and occupied time can be calculated by subtracting OFF timestamp and previous ON timestamp (record 4 – record 3).
- Record c,d - At 08:09:16.76, GREEN at phase 3 turned OFF after 29.38 sec and at the same time YELLOW at phase 3 turned ON. Both events are represented as "3" (record 5) which indicates end of GREEN and start of YELLOW at phase 3. Also, GREEN duration can be calculated by subtracting timestamps indicating START and END of GREEN (record 5 – record 1).
- Record e:g, f:h, j:k – Records e, f, g, h, j, k represents record 6, 7, 8, 9, 11, 12, respectively. Also, occupancy time for detector 9, 10, 22 can be calculated by (record 8 – record 6), (record 9 – record 7) and (record 12 – record 11), respectively.

Chapter 3

Driver's Decision Identification

After collecting high-resolution event data, the second step is to identify YLR, RLR, and FSTP cases to study the behavior of a driver's decision of SoR at the signalized intersection. To better understand our identification method for YLR, RLR, and FSTP, it is necessary to first explain the details of event data.

3.1 High-Resolution event data

The raw event data collected by the SMART-SIGNAL system is first used to estimate individual vehicle speeds and time gaps. As shown in figure 3.1, event data contains information of vehicle's arrival time (T_i^o for i^{th} vehicle) and departure time (T_i^f for i^{th} vehicle). The time difference between T_i^f and T_i^o is occupancy time (t_i^{on}), i.e., the time the detector is occupied by a vehicle. If we assume a known effective vehicle length (l_{eff}), which is the sum of vehicle length and detector length (l_d), the speed of individual vehicles (v_i) can be calculated by the second equation in Eq.(1). When combining the information from two consecutive vehicles, time gap (t_i^{gap}) and headway (t_i^h) can be derived using the third and fourth equations in Eq.(1). As shown in previous research [20, 23], these variables can have significant impacts on drivers' decisions at signalized intersections. In this research, we mainly focus on the impact of the occupancy time (which indicates vehicle's velocity) and time gap (which indicates the relationship between vehicles).

$$\begin{cases} t_i^{on} = T_i^f - T_i^o \\ v_i = \frac{l_{eff}}{t_i^{on}} \\ t_i^{gap} = T_{i+1}^o - T_i^f + \frac{l_d}{v_{i+1}} \\ t_i^h = T_{i+1}^o - T_i^o \end{cases} \quad (1)$$

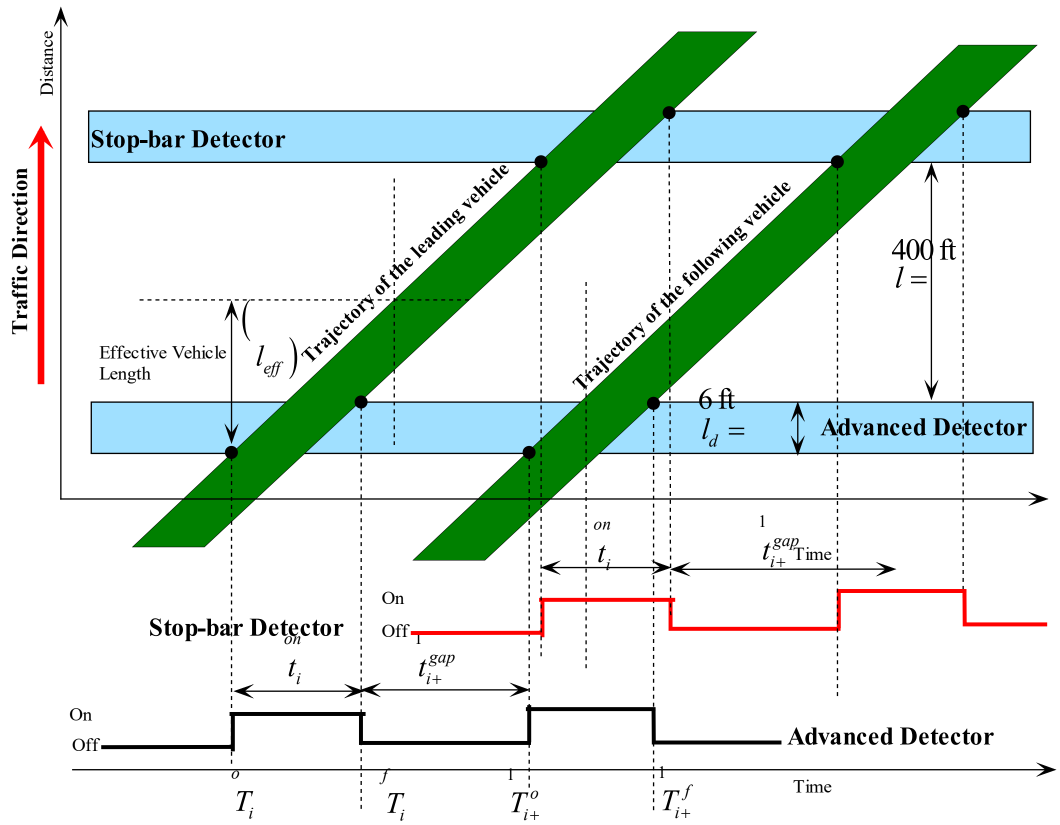


Figure 3.1 Vehicle on-time and time gap

3.2 FSTP/YLR/RLR Identification

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 prediction model can then be used to help avoid collisions by adjusting signal timing. So the first crucial step is to explore whether there is a strong connection between a driver's decision of SoR and driving speed, and signal timing information as well as the information of the driving status of preceding and adjacent vehicles. A driver's decision at the end of green could be first-to-stop (i.e. FSTP; note if there is already a stopped vehicle, the driver's choice is simply 'stop'), or running through the intersections either during yellow (YLR) or red (RLR) signal. These situations represent FSTP, YLR, and RLR cases respectively (See fig 3(a)).

However, merely using an advance detector cannot accurately detect RLR, YLR, and FSTP. Instead, we use stop-bar detectors to identify YLR, RLR, and FSTP cases. The basic idea is very intuitive: if a vehicle approaches stop-bar with a relatively high speed (greater than a threshold value), we conclude that the driver decides to run through the intersection; otherwise, this driver decides to stop. For the stop cases, only the vehicle which is the first to stop before the stop line is defined as "FSTP." For the running situations, if the signal indication is red when the vehicle passes through the stop-bar detector, it is an RLR; and it is a YLR if the signal is yellow. The vehicle's speed is estimated using the effective vehicle length divided by the occupancy time (see the second equation in Eq. (1). We use 25 ft as the effective vehicle length as this value has been calibrated in authors' previous research [29].

If the stop-bar detector is installed directly behind the stop bar (such as Int. Rhode Island, see Fig. 2.1), the threshold value is set as 10 mph. But for some intersections like Int. Boone and Int. Winnetka, the detectors are located about 50 ft. behind the stop-line. Therefore, a different threshold value of 20 mph is applied. Note the limits of 10mph and 20 mph are experimentally verified and relatively conservative to ensure that the identifications of RLR are correct, but with the cost of missing some RLR events.

But only using speed information can mistakenly identify cases. Therefore, two additional conditions are applied to confirm the identification further:

1) Compare the average speed of three preceding vehicles and the speed of the target vehicle. If the speed difference is less than 10mph (and higher than the threshold values as mentioned before), the vehicle can be identified as passing through the intersection. But if the speed difference is larger than 20mph, we consider that the vehicle is slowing down and will stop before the stop line;

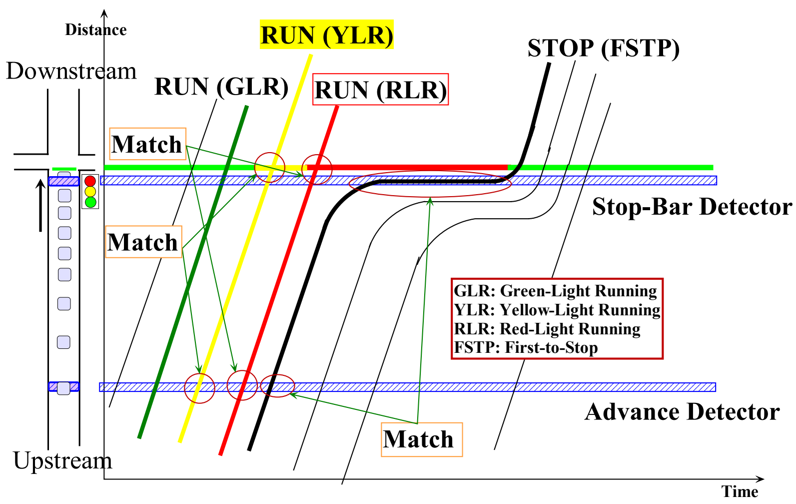
2) Verify if the distance between the stop line and detector is enough for a vehicle to stop safely. This is especially important for the intersections where the detectors are not located right behind the stop line. A simple equation of motion is applied to estimate vehicle's stopping distance (d_i) by assuming a deceleration rate of 10 ft/s^2 (i.e. $a = 10\ ft/s^2$) (the value is suggested by ITE) [44].

$$d_i = \frac{v_i^2}{2*a} \qquad (2)$$

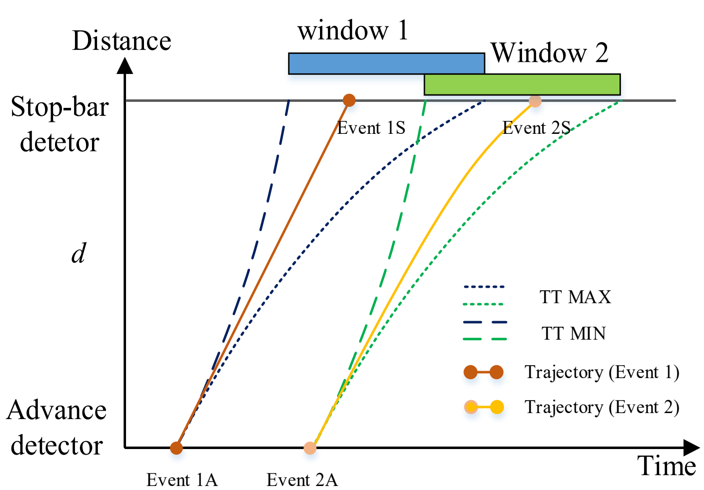
After testing the above conditions, we can safely identify if a vehicle is running through the intersection or stopping before the stop line (FSTP). According to signal timing information, the running cases can be further categorized into RLR and YLR. Some of the identified cases are manually checked using the information from entrance detectors on downstream links to verify the accuracy of the proposed method. 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.

3.3 Event matching between stop-bar and advanced detectors

Another important step is to match FSTP/YLR/RLR events identified by a stop-bar detector with the vehicle events recorded by the corresponding advance detector located on the same lane but 400 feet upstream (Fig.3.2(a)). The information collected by the advance detectors will later be used to estimate vehicle trajectories for conflict analysis and to analyze drivers' decisions of SoR to predict an RLR ultimately. The prediction must be made at the advance detector so that there will be time for performing any conflict or RLR prevention strategy before vehicles reach the intersection. Event data matching becomes tough because cars could accelerate, decelerate, and most importantly, change lanes while traveling from the advance detectors to the stop detectors. However, the matching method assumes no lane changing takes place between the advanced and stop detectors, which is an appropriate assumption considering the short distance between the two detectors, and their proximity to the intersection.



(a)



(b)

Figure 3.2 (a) STOP or RUN event matches; (b) "Window-searching" method.

A simple "window-searching" method is applied to match the events recorded by advance and stop-bar detectors (see Fig. 3.2(b)). 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.

Mathematically, the "window-searching" method can be formulated as follows. If we denote that one vehicle Veh_i^{ad} arrives at the advance detector at time T_i^{ad} but a set of vehicles has been detected by stop-bar detector during the time Δt. If we define this set of vehicles as $S_{stopbar} = \{Veh_1^{sd}, Veh_2^{sd}, \ldots Veh_p^{sd}\}$, the 'match' problem then can be defined as a process to find a vehicle Veh_j^{sd} in set $S_{stopbar}$, which is the same vehicle of Veh_i^{ad}. Here p is the total number of vehicles in the set $S_{stopbar}$; and Δt is the "time window" calculated based on the possible maximum and minimum travel time. Specifically, the maximum travel time (TT_i^{max}) for vehicle i is estimated based on the assumption that the vehicle will fully stop at the stop bar. Also, the minimum travel time (TT_i^{min}) is calculated by assuming a maximum acceleration rate ($a_{constant}$) of 6 $feet/sec^2$ suggested by Long (2000), as shown in the following equation:

$$\begin{cases} TT_i^{max} = \dfrac{2*d}{v_i^{ad}} \\ TT_i^{min} = \dfrac{-v_i^{ad}+\sqrt{(v_i^{ad})^2+2*d*a_{constant}}}{a_{constant}} \end{cases} \quad (2)$$

Where v_i^{ad} is the individual vehicle speed for Veh_i^{ad} when this vehicle arrives at the advance detector. Thus the Δtt can be defined as the time duration of $[T_i^{ad} + TT_i^{min}, T_i^{ad} + TT_i^{max}]$.

For vehicle Veh_i^{ad} and any vehicle Veh_j^{sd} in set $S_{stopbar}$ during Δt, we can set a match strength function, $m_{i,j}$, for these two vehicles:

$$m_{i,j} = \frac{1}{|T_j^{sd} - T_i^{ad} - t_{i,j}|} \quad (3)$$

Where T_j^{sd} is the arrival time of Veh_j^{sd} at the stop-bar detector and $t_{i,j}$ is the expected travel time of vehicle from the advance detector to stop-bar detector. If we assume vehicles keep a constant acceleration or deceleration rate, $t_{i,j}$ can be calculated as

$$t_{i,j} = \frac{2d}{V_i^{ad}+V_j^{sd}} \quad (4)$$

where V_j^{sd} is the vehicle speed when Veh_j^{sd} arrives at the stop-bar detector and d is the distance between the advance detector and stop-bar detector, as shown in in Figure 3.2(b).

To search for the right match, for vehicle Veh_i^{ad}, we calculate $m_{i,j}$ between Veh_i^{ad} and any other vehicle in set $S_{stopbar}$. The vehicle pair with the highest value of $m_{i,j}$ is considered as a right match. Considering that lane changing could bias our matching results, we further use the data collected by downstream link entrance detectors (see Figure 2.1) to ensure that those "running" vehicles (i.e. RLR or YLR) traveled directly from the advance detector to link entrance detectors. In such situation, the possibility of lane-changing activities within such short distance (400 ft) is small.

However, due to the limitation of the loop detector data, our method cannot detect all RLR cases. For example, although it is rare, it is possible that a vehicle that passes an advance detector decides to RLR by passing the car in front of him/her and moves to the other lane. In fact, our purpose is not to detect all RLR cases, but to ensure that all identified RLR cases are exact matches. Also, because we use a large amount of event data, missing some RLR cases will not have a significant impact on our results.

3.4 Data Summary

Some datasets were collected time to time for further research. Note the proposed method was not able to identify all the cases. Due to short link and complicated geometry for Int. Winnetka, the identified cases were much less than the other two intersections. Also, because of the signal progression design, which intends to make vehicles stop at Intersections Boone and Rhode Island, the cases identified in these two intersections had more FSTP and YLR cases. But for all three intersections, the RLR cases were not many. The possible reason could be the actuated signal controls for all three intersections, with long amber times of 5.5 sec due to the high-speed limit of 55 MPH. Therefore, most of "running" cases ended up YLR, not RLR.

All the different aspects of study such as SoR prediction, RLR prediction, conflict estimation and safety evaluation were carried out at different point of time during the research process and hence used different datasets. A summary of various datasets used during this research is given below (see table 3.1, 3.2, 3.3).

Table 3.1 Stop or Run Prediction

Intersection	Months	Total events	RLR events	YLR events	FSTP events
Boone	3	49531	128	22794	27035
Winnetka	3	15238	425	12009	2804
Rhode Island	3	33572	381	22582	10609
Total	9				

Table 3.2 Red Light Running Prediction

Intersection	Months	Total events	RLR events	YLR events	FSTP events
Boone	3	42,227	289	36,155	5833
Winnetka	3	34,127	307	28,788	5032
Rhode Island	3	33,370	266	25,543	7561
Total	9	109,774	862	90,486	18,426

Table 3.3 Conflict estimation and Safety Index

Intersection	Months	Total events	RLR events	YLR events	FSTP events
Boone	7	73,814	637	31,469	41,708
Winnetka	10	72,538	431	39,684	32,423
Rhode Island	4	26,795	322	15,061	11,412
Total	21				

Chapter 4

Stop-or-Run (SoR) Prediction

In this chapter, FSTP/YLR/RLR cases have been analyzed to find factors that significantly impact drivers' SoR decisions. 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 [20, 23, 31]. The factors that impact a driver's decision, according to other studies, include approaching speed, yellow time, traffic flow, estimated travel time to reach the stop-bar at the start of yellow, vehicles on adjacent lanes, etc. Some of these factors have been reinvestigated, and alternatives have been searched using the data collected from advance detectors. Data from three different months are used for this research. (for Boone Ave., the data from Nov., 2008, May, 2009, and Jun., 2009 are used; for Winnetka Ave., the data from Nov., 2008, Jan., 2009, and Jun., 2009 are used; and for Boone Ave., the data from Nov., 2008, Jan., 2009, and Feb., 2009 are used.)

4.1 Impact Factors

The information obtained from advance detectors located 400 feet upstream from the stop-bar has been utilized in the analysis. Vehicle's behaviors including FSTP, YLR, and RLR have been identified using the program introduced in the previous chapter. The information that could impact drivers' decisions and could be directly collected from advance loops includes occupancy time and time gap between consecutive vehicles. This information primarily indicates vehicle's velocity and the (time) distance between the target and leading vehicles at the time when the target vehicle passes the detector. Signal timing has also been collected. But since we are only interested in major approach through movements, most of the signal timing information such as yellow time, cycle length, and all red time are invariable (since all six intersections are coordinated with each other, so they have a constant cycle duration of 180 seconds). Two types of signal timing related information, which we think will significantly impact drivers' decisions, have been chosen to be:

- used yellow time, i.e. the portion of yellow time that elapsed before the vehicle arrives at the advance detector; and
- time to yellow start, i.e. the time left until signal changes to yellow.

Also, to analyze if surrounding vehicles' statuses have an impact on the target vehicle, the data for three preceding vehicles including occupancy times, time gaps, and their decisions (i.e. RLR or YLR) and the information of whether there is a vehicle driving on the adjacent lane were collected. In summary, the following factors have been analyzed in this chapter:

- Occ_A: occupancy time for the target vehicle when passing advance detector;

- Gap_A: time gap between the target vehicle and the nearest preceding vehicle;
- YT_Used_A: the portion of yellow time that elapsed before vehicle arrives at advance detector;
- G2Y_Start_A: the time left until signal changes to yellow;
- Occ_A1, Occ_A2, Occ_A3: occupancy times for the three preceding vehicles (A1 is the nearest one to the target vehicle);
- Gap_A1, Gap_A2, Gap_A3: time gaps for the three preceding vehicles (A1 is the closest one to the target vehicle);
- RLR_A1, RLR_A2, RLR_A3: vehicles' behaviors for three preceding vehicles, category variables with possible values of green-light running (GLR), YLR or RLR;
- RLR_AA: presence of running vehicles in the adjacent lane, a binary variable with "1" represents yes and "0" no;
- EST_TravelTime: estimated travel time to reach the stop-bar at the start of yellow.

Note EST_TravelTime has been determined using vehicle speed (calculated from occupancy time) and signal timing information (including YT_Used_A and G2Y_Start_A). The following equation has been applied to estimate EST_TravelTime value:

$$\mathrm{EST}_{TravelTime\,i} = \begin{cases} YT_{Used_A} + \frac{l}{v_i} \; if\; YT_{Used_A} > 0 \\ max\left\{0, \frac{l}{v_i} - G2Y_{Start_A}\right\} if\, G2Y_{Start_A} > 0 \end{cases} \tag{4}$$

4.2 Mean Values

Before testing the significance of all these factors, it was necessary to visually compare the mean values of some important factors for both "stop" and "run" situations, i.e., FSTP and RUN. Fig. 4 presents the average values of Occ_A, Gap_A, Occ_A1, Gap_A1, G2Y_Start_A, YT_Used_A, and EST_TravelTime for both "FSTP" and "RUN" situations for all three intersections. As indicated in the figure, most factors have significantly different mean values when a driver makes different decisions, especially for time gaps (Figure 4.1(b)), used yellow time (Figure 4.1(f)), and estimated travel time (Figure 4.1(g)). But occupancy values do not show significant changes in two different behaviors (Figure 4.1(a)).

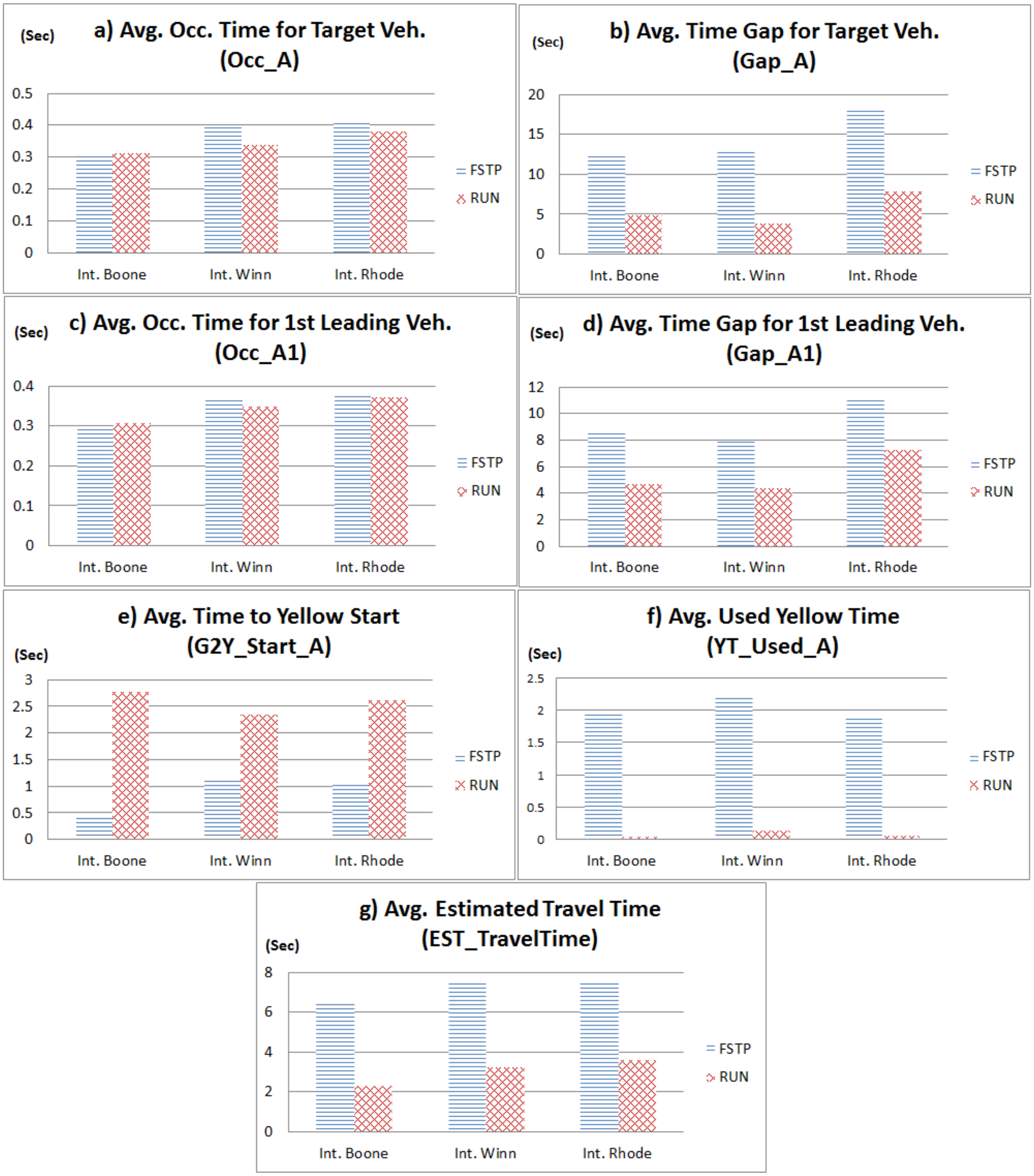


Figure 4.1 Mean values for FSTP and RUN

4.3 Significance Testing

A more rigorous way to test the significance of a factor is to apply a statistics model. Since a driver's SoR decision is a binary variable, a binary logistic regression model has been used to determine the probability that a vehicle will stop or run through the intersection.

$$\frac{p_i}{1-p_i} = e^{\alpha+\overrightarrow{\beta'}\overrightarrow{X_i}} \qquad (5)$$

wherep_i is the probability of stop or run; α is a constant; (β^') ⃗ is a vector of slope parameters; and (X_i) ⃗ is a vector of predictor variables.

All factors mentioned in Section 4.1 (except EST_TravelTime) have been used to fit the model. Some dummy variables have also been introduced to describe categorical variables. For example RLR_A1_Dm1 = (0, 1, 0) where"1" represents "YLR" and RLR_A1_Dm2= (0, 0, 1) where"1" represents "RLR". A similar design has been applied to RLR_A2 and RLR_A3. For RLR_AA, we defined RLR_AA_Dm = (0, 1) where"1" represents the presence of a vehicle driving in an adjacent lane.

After setting up these variables, the binary backward stepwise logistic regression model in SPSS (v. 20) has been applied to analyze the three intersections' data respectively. The results are shown in Table 4.1.

Table 4.1 The Binary Logistic Regression Model (red and bold indicates the factor is significant)

	Int. Boone Ave.		Int. Winnetka		Int. Rhode Island	
	B	Sig.	B	Sig.	B	Sig.
Occ_A	**-7.509**	**.000**	**-8.938**	**.000**	**-5.782**	**.000**
Gap_A	**-.049**	**.000**	**-.075**	**.000**	**-.031**	**.000**
YT_Used_A	**-1.808**	**.000**	**-1.315**	**.000**	**-2.309**	**.000**
G2Y_Start_A	**.678**	**.000**	**.240**	**.000**	**.070**	**.000**
Occ_A1	-.610	.000				
Gap_A1	**-.013**	**.000**	**-.020**	**.000**	**-.011**	**.000**
RLR_A1_Dm1	**.719**	**.000**	**.649**	**.000**	**.291**	**.000**
RLR_A1_Dm2			1.667	.083	3.447	.001
Occ_A2			-.186	.016	.282	.021
Gap_A2	-.015	.000				
RLR_A2_Dm1					3.294	.008
RLR_A2_Dm2						
Occ_A3			.986	.000		
Gap_A3	-.010	.000				
RLR_A3_Dm1	1.808	.000	.682	.074	1.900	.000
RLR_A3_Dm2					8.836	.002

RLR_AA_Dm	**.345**	**.000**	**.788**	**.000**	**-.080**	**.023**
Constant	**2.343**	**.000**	**4.989**	**.000**	**4.292**	**.000**
Sig. of Model	.000		.000		.000	
R Square	.691		.642		.616	

As shown in Table 4.1, significant factors were not the same for all the three models based on data collected from the intersections. However, occupancy time (Occ_A), used yellow time (YT_Used_A), and time to yellow start (G2Y_Start_A) were significant for all three models. These three factors essentially indicate estimated travel time. The time gaps between the target vehicle and the nearest two preceding vehicles (Gap_A and Gap_A1) also showed significance in the model. Also, whether the preceding vehicle passed the intersection during the yellow interval (RLR_A1_Dm1) had a significant impact on the model. The significance of these three factors (Gap_A, Gap_A1, RLR_A1_Dm1) indicated that a driver's decision of SoR is influenced by the status and decisions of the nearest preceding vehicles. Although the vehicles before the nearest preceding vehicles (A2 & A3) could have an impact on drivers' decision making as shown in Table 1 (for example, some factors like Gap_A2 and RLR_A3_Dm1 were significant at some intersections), the significance was not consistent for all intersections. Also, a vehicle passing by in the adjacent lane (RLR_AA_Dm) showed significant impact on the fit model. To be known, all three models were significant with an R-squared value around 0.65. Although the findings were consistent with previous research, the concept of testing the impact of three preceding vehicles introduced in this research is relatively new.

4.4 Estimated Travel Time

As we mentioned, Occ_A, YT_Used_A, and G2Y_Start_A primarily represent the information of estimated travel time to the stop-bar at the yellow start. To investigate the significance of the estimated travel time as done in most of the previous research, Occ_A, YT_Used_A and G2Y_Start_A were replaced by EST_TravelTime and the logistic regression model was re-run. The results are shown in Table 4.2.

Table 4.2 The Binary Logistic Regression Model Using EST_TravelTime Data (red and bold indicates the factor is significant)

	Int. Boone Ave.		Int. Winnetka		Int. Rhode Island	
	B	Sig.	B	Sig.	B	Sig.
Gap_A	**-.055**	**.000**	**-.078**	**.000**	**-.036**	**.000**
EST_TravelTime	**-.946**	**.000**	**-.687**	**.000**	**-.594**	**.000**
Occ_A1	-.209	.000			-.770	.000
Gap_A1	**-.017**	**.000**	**-.017**	**.000**	**-.016**	**.000**
RLR_A1_Dm1	**.773**	**.000**	**.722**	**.000**	**.460**	**.000**
RLR_A1_Dm2						
Occ_A2	.360	.000	-.320	.003		

Gap_A2	-.018	.000	-.007	.093	-.008	.000
RLR_A2_Dm1						
RLR_A2_Dm2						
Occ_A3					.177	.061
Gap_A3	-.013	.000				
RLR_A3_Dm1	1.317	.000			.924	.000
RLR_A3_Dm2					4.279	.004
RLR_AA_Dm	**.407**	**.000**	**.741**	**.000**	**-.197**	**.000**
Constant	3.886	.000	5.150	.000	4.755	.000
Sig. of Model	.000		.000		.000	
R Square	.682		.609		.532	

Without any surprise, the results indicated that estimated travel time was significant to the regression model. Besides, the model fitting results also showed that Gap_A, Gap_A1, RLR_A1_Dm1, and RLR_AA_Dm were significant for all the three intersections, similar to the testing results shown in Table 1. The R-squares of the new models using EST_TravelTime were less than the previous model using Occ_A, YT_Used_A, and G2Y_Start_A, but the differences were slight. This testing merely indicated that we could use the information of Occ_A, YT_Used_A, and G2Y_Start_A, which can be directly measured from loop detectors and signal system, instead of estimated travel time, to predict the possibility of a driver's decision of stop or go through the intersection.

4.5 Prediction Model

A principal purpose of this research was to verify whether we can use loop detector information to predict drivers' SoR behaviors when signal switches to yellow. Based on Eq. 5, the probability of "go" (P_{Go}) and "stop" (P_{Stop}) has been estimated using Eq. 6:

$$\begin{cases} P_{Go} = p_i = \dfrac{e^{\alpha+\overrightarrow{\beta'}\overrightarrow{X_i}}}{1+e^{\alpha+\overrightarrow{\beta'}\overrightarrow{X_i}}} \\ P_{Stop} = 1 - p_i = \dfrac{e^{\alpha+\overrightarrow{\beta'}\overrightarrow{X_i}}}{1+e^{\alpha+\overrightarrow{\beta'}\overrightarrow{X_i}}} \end{cases} \qquad (6)$$

The constant and slope parameters have been determined in Table 1.

To verify the accuracy, we used a new month's data (Dec. 2008) at Int. Rhode Island. Using this month's data and applying our program introduced in Section 2, we were able to identify a total of 11410 cases, of which 3431 were FSTP, i.e. "stop" cases, and 7979 were either YLR or RLR, i.e. "run" cases. Using Eq. 6 and the values presented in Table 1, we correctly estimated 9967 events. The accuracy rate was found to be 87%.

For comparison, we also used the model with EST_TravelTime and the parameter values presented in Table 2 to predict drivers' decisions. This model was able to correctly predict 9727 events with an accuracy rate of 85%.

This experiment clearly demonstrated that using information collected from advance detectors; we were able to accurately predict drivers' decision of "stop" or "run." Since all the information required could be collected in real time, the model presented here could be used to predict drivers' decision in real time and could be tremendously beneficial for real-time collision avoidance.

Chapter 5

Red-Light-Running (RLR) Prediction

5.1 Influential factors for RLR

The large amount of matched event data presented in Chapter 2 record the specific driving behaviors of each vehicle (i.e. RLR, YLR, or FSTP). From a statistical point of view, a driver's current driving conditions (i.e. speed and the time gap between the target and leading vehicles), together with surrounding traffic conditions (i.e. the driving behaviors of surrounding vehicles) and signal timing situations (i.e. signal status of green, yellow, or red and their durations), would directly or indirectly lead to the driver's later behavior of RLR and Non-RLR (note YLR and FSTP has been combined as Non-RLR class). Therefore, by statistically analyzing a large amount of event data, the inner correlation between all these impact factors and driver's RLR or Non-RLR behavior could be derived; and such correlation then could be used to predict RLR through some statistical methods, such as logistic regression.

5.1.1 A binary logistic regression model

Since a driver's RLR behavior is a binary variable, a binary logistic regression model has been applied to describe the correlation between all impact factors and driver's RLR behavior (RLR or Non-RLR). If we had to define RLR and Non-RLR as "1" and "0", respectively, a standard logistic regression model for RLR could be described as:

$$logit[\pi_i] = log\left[\frac{\pi_i}{1-\pi_i}\right] = \alpha + \beta_c x_i \qquad (4)$$

where π_i is the probability that the target vehicle i is a RLR and $1 - \pi_i$ is the probability that the target vehicle i is a Non-RLR; x_i represents a vector of all control factors which impact the behavior of vehicle i (the details about control factors will be discussed in next section); α is an intercept parameter; and β_c is a vector of the coefficients of the corresponding control factors. α and β_c have been estimated by the method of maximum likelihood estimation (MLE). The likelihood function was constructed as Eq. (5). By maximizing the log-likelihood expression shown in Eq. (6), the estimate of the new intercept parameter α and coefficients vector β_c has been obtained accordingly. Note the "logit" function in the Stata (v. 10) was used to get all the results.

$$l(\beta_c) = \prod_{i=1}^{n}\{\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}\} \qquad (5)$$

$$LL(\beta_c) = \ln\big(l(\beta_c)\big) = \sum_{i=1}^{n}\{y_i ln[\pi(x_i)] + (1-y_i)ln[1-\pi(x_i)]\} \qquad (6)$$

where y_i represent that whether vehicle i would run the red light, with a value of either 0 or 1 only; $i = 1,2, \ldots n$ and n is the total number of observed vehicles.

5.1.2 Influential factors and significance

To apply Eq. (4), the first critical step was to figure out influential or control factors (i.e. x_i in Eq. (4)), which had direct or indirect impacts on drivers' RLR or Non-RLR behaviors. Much previous research has studied this problem (e.g. Bonneson et al., 2001; Gates et al., 2007; Yang and Najm, 2007; Elmitiny et al., 2010;, etc.). In this research, we considered all the potential factors extracted from the high-resolution traffic and signal event data which could impact a driver's RLR behavior, as shown in Table 2. This list included the information of occupancy time and the time gap between consecutive vehicles, which essentially indicate the vehicle's velocity and the (time) distance between the target and the leading vehicles at the time at which the target vehicle passes the detector. Signal timing has also been included in the list. Two types of signal timing related information that significantly affects drivers' behaviors have been chosen to be used here:

(a) the used yellow time, that is, the portion of yellow time that elapsed before the vehicle arrives at the advance detector; and

(b) the time to yellow start, that is, the time left until the signal changes to yellow.

Also, to analyze whether the status of the surrounding vehicles had an effect on the target vehicle, the data for three preceding vehicles had been collected. The data included occupancy times, time gaps, and their behaviors (i.e., RLR or YLR) and the information on whether there was a vehicle driving in the adjacent lane. In summary, this list includes most of the potential factors such as the driving conditions of the target vehicle itself, driving conditions of surrounding vehicles, and signal timing information. The significance of these factors has been determined during the regression process.

Table 5.1 Descriptive statistics of the potential variables

Variable	Description
Occ_A	Occupancy time for the target vehicle when passing advance detector (s)
Gap_A	Time gap between the target vehicle and the nearest preceding vehicle (s)
Yt_Used_A	The yellow time that elapsed before vehicle arrives at advance detector (s)
G2Y_Start_A	The green time left until signal changes to yellow (s)
Occ_A1 Occ_A2 Occ_A3	Occupancy time for three preceding vehicles (A1 is the nearest one to the target vehicle) (s)
Gap_A1 Gap_A2 Gap_A3	Time gap for three preceding vehicles (A1 is the nearest one to the target vehicle) (s)
YLR_A1_Dm YLR_A2_Dm	Vehicles' YLR behaviors for three preceding vehicles (A1 is the nearest one to the target vehicle); a binary variable with "1" represents yes (i.e. YLR) and "0" no.

YLR_A3_Dm	
RLR_A1_Dm RLR_A2_Dm RLR_A3_Dm	Vehicles' RLR behaviors for three preceding vehicles (A1 is the nearest one to the target vehicle); a binary variable with "1" represents yes (i.e. RLR) and "0" no.
AA_Dm	Presence of running vehicles on the adjacent lane, a binary variable with "1" represents yes and "0" represents no.

After identifying all possible influential factors as shown in Table 5.1, we then use the data extracted from Chapter 3 to test the significance of these influential factors statistically. Note for each intersection, three-month data were collected, but only two-month' data were used to derive intercept parameters and coefficient vectors; the other month's data were used later for evaluating RLR prediction. All factors described in Table 5.1 were used to fit a binary logistic regression model. The results are shown in Table 5.2. The "logit" function in the Stata (v. 10) was used to obtain all the results.

Table 5.2 Standard binary logistic regression model

	Int. Boone Ave.		Int. Winnetka		Int. Rhode Island	
Variable	Parameter	P-value	Parameter	P-value	Parameter	P-value
Constant	**-2.663**	**P<0.01**	**-2.140**	**P<0.01**	**-2.263**	**P<0.01**
Occ_A	**-1.275**	**P<0.01**	**-0.707**	**P<0.01**	**-1.867**	**P<0.01**
Gap_A	**-0.004**	**P<0.05**	**-0.002**	**P<0.05**	**-0.010**	**P<0.01**
Yt_Used_A	**0.282**	**P<0.01**	**0.296**	**P<0.01**	**0.526**	**P<0.01**
G2Y_Start_A	**-3.052**	**P<0.01**	**-1.616**	**P<0.01**	**-3.437**	**P<0.01**
Occ_A1	/	/	1.146	P<0.01	1.947	P<0.01
RLR_A1_Dm	/	/	-3.371	P<0.01	/	/
Gap_A2	0.008	P<0.01	/	/	/	/
YLR_A2_Dm	**0.670**	**P<0.01**	**0.965**	**P<0.01**	**0.266**	**P<0.05**
Gap_A3	/	/	/	/	0.010	P<0.01
YLR_A3_Dm	0.628	P<0.01	/	/	0.591	P<0.05
AA_Dm	**0.481**	**P<0.05**	**0.213**	**P<0.05**	**0.354**	**P<0.01**

Note: *: Significance of model = .000; "/" indicates that variable is not significant.

Only variables which are statistically significant at the significance level of 0.05 has been included in the table. The results showed that Occ_A, YT_Used_A, and G2Y_Start_A were common significant factors for all three intersections. This was consistent with our understanding of RLR behaviors. As we know, Occ_A indicates the speed of the target vehicle. This factor, together with used yellow time (YT_Used_A), and time to yellow start (G2Y_Start_A), certainly has a significant impact on whether a driver decides to pass the intersection within the rest of yellow time. The results also showed that Gap_A (i.e. time gap) was a common significant factor for all three intersections that indicated vehicles' car-following behavior, i.e., the shorter the gap time, the more likely the following vehicle would follow the leading vehicle and run through the intersection even traffic light is red. Interestingly, the results also showed that the YLR behavior of the second preceding vehicle (i.e. YLR_A2_Dm) was significant for all three intersections which probably indicated platoon behavior. That means when the first vehicle in a platoon runs through the intersection during yellow; the following several vehicles very likely end up either YLR or RLR since long yellow time (i.e. 5 sec) was provided for these three intersections.

Also, whether there was a vehicle passing through on the adjacent lane (AA_Dm) showed a significant effect on RLR for all three intersections. This demonstrated that the "following" behavior of drivers was not only impacted by preceding vehicles, but also by adjacent vehicles. Note that all fittings were statistically significant.

5.2 RLR prediction using rare events logistic regression

Based on Eq. (4), the probability that the target vehicle was a RLR could be described by the logistic distribution as shown in the following equation:

$$p(y_i = 1|x_i) = \pi_i = \frac{\exp(\alpha + \beta_c x_i)}{1+\exp(\alpha + \beta_c x_i)} \qquad (7)$$

where yy_{ii} denotes the behavior of vehicle i. $yy_{ii} = 1$ means the vehicle i is a RLR.

The above logistic regression model worked well for YLR prediction (Lu et al., 2015). However, when applying to RLR prediction, the prediction accuracy was very low due to the rare events nature of RLR. As shown in Table 1, the proportions of RLR events were small for all three intersections: 0.7% at Int. Boone, 1.0% at Int. Winnetka, and 0.8% at Int. Rhode Island. With such small portions of RLR, applying the standard binary logistic regression method sharply underestimated the probability for RLR. To address this challenging issue, a modified rare events binary logistic regression, originally developed by King and Zeng (2001 and 2002), has been applied in this research to improve RLR predictions.

5.2.1 Rare events logistic regression method & fitting results

The proposed rare events logistic regression method essentially included the following three-step correction procedure (King and Zeng, 2001 and 2002):

Step 1: The first step was to apply a choice-based data collection strategy to select Non-RLR cases randomly. This move would form a new dataset, in which all RLR cases would be included but only a portion of randomly selected Non-RLR cases would be included. In this new dataset, the ratio of RLR events to Non-RLR events was recommended to set around 1:10 as suggested by other rare events studies (e.g. Beguería, 2006 and Guns and Vanacker, 2012). With the new dataset, the MLE technique could be applied to estimate new intercept parameter $\hat{\alpha}$ and coefficients vector $\hat{\beta}_c$.

Step 2: However, the use of choice-based sampling strategy could have significantly biased the estimation of intercept parameter $\hat{\alpha}$. Therefore, the second step was to apply a prior correction to avoid sampling bias. In this step, the corrected intercept parameter $\tilde{\alpha}$ was calculated based on the uncorrected intercept parameter $\hat{\alpha}$, as in Eq. (8):

$$\tilde{\alpha} = \hat{\alpha} - \ln[(\frac{1-\tau}{\tau}) (\frac{\varepsilon}{1-\varepsilon})] \qquad (8)$$

where τ is the actual fraction of "1"s (i.e. RLRs) in the whole population, and ε is the observed fraction of "1"s in the new dataset.

Step 3: With corrected intercept $\tilde{\alpha}$ and updated coefficients vector $\hat{\beta}_c$, Eq. (7) could be used to calculate the updated probabilities $\hat{P}_i$. However, $\hat{P}_i$ was still an underestimated probability, because the estimation uncertainty of the coefficients vector $\hat{\beta}_c$ was neglected. So the third step was to correct $\hat{P}_i$ by introducing a correction factor C_i to $\hat{P}_i$. The final corrected probability P_i was obtained as:

$$\begin{cases} P_i = \hat{P}_i + C_i \\ C_i = (0.5 - \hat{P}_i)\, \hat{P}_i (1 - \hat{P}_i) X_i V(\beta) X_i' \end{cases} \qquad (9)$$

where $\beta = (\tilde{\alpha}, \hat{\beta}_c)$, $X_i = [1, x_i]$, β and X_i have same dimensions, X_i' is the transpose of X_i, and $V(\beta)$ is the estimated variance-covariance matrix of the estimated coefficients.

The rare events logistic regression model was applied to fit the high-resolution event data collected earlier. The regression fitting results are displayed in Table 5.3. Note two same months' data were used to derive intercept parameter and coefficients vector, and the other month's data were used for evaluation. To implement the rare events logistic regression model, we used the "relogit" function in the Stata (v.10). The results are shown in Table 5.3.

Table 5.3 Rare events logistic regression model

	Int. Boone Ave.		Int. Winnetka		Int. Rhode Island	
Variable	Parameter	P-value	Parameter	P-value	Parameter	P-value
Constant	**-1.998**	**P<0.01**	**-1.885**	**P<0.01**	**-1.537**	**P<0.01**
Occ_A	**-1.371**	**P<0.01**	**-1.685**	**P<0.01**	**-1.830**	**P<0.01**
Gap_A	**-0.001**	**P<0.01**	**-0.010**	**P<0.01**	**-0.017**	**P<0.01**
Yt_Used_A	**0.342**	**P<0.01**	**0.176**	**P<0.01**	**0.721**	**P<0.01**
G2Y_Start_A	**-3.331**	**P<0.01**	**-1.745**	**P<0.01**	**-3.074**	**P<0.01**
Occ_A1	/	/	1.068	P<0.01	1.035	P<0.01
RLR_A1_Dm	/	/	-2.337	P<0.01	/	/
Gap_A2	0.039	P<0.01	/	/	/	/
YLR_A2_Dm	**0.640**	**P<0.01**	**1.106**	**P<0.01**	**0.169**	**P<0.05**
Gap_A3	/	/	/	/	0.011	P<0.01
YLR_A3_Dm	1.202	P<0.01	/	/	1.263	P<0.05
AA_Dm	**0.803**	**P<0.05**	**0.436**	**P<0.05**	**0.162**	**P<0.01**

Note: *: Significance of model = .000; "/" indicates that variable is not significant.

The significant impact factors of the rare events model were consistent with those factors for the standard logistic regression model. In addition, there were not any factors which flipped the signs after applying the rare events logistic regression model. This further indicated the consistency between the standard and rare events logistic regression models. But the values for all significant factors have been changed.

5.2.2 Prediction Comparison

For comparison, both the regression models presented in Table 5.2 and Table 5.3 were used to predict RLR based on the event data collected from the third month. The variables that were statistically significant at the significance level of 0.05 were included in the prediction models. As we mentioned before, for each intersection, three month's data were collected, but only two-month' data were used to derive intercept parameters and coefficient vectors; the other month's data were used for evaluating prediction accuracy. Table 5.4 presents all the prediction results. Note "standard" and "rare" represent standard binary logistic regression model and rare events binary logistic regression model respectively.

Table 5.4 RLR predictions

	Int. Boone		Int. Winnetka		Int. Rhode	
Regression Model*	Standard	Rare	Standard	Rare	Standard	Rare
Total Events	19960		14730		11400	
Actual RLR Number	119		128		74	
Predict RLR Number	247	405	211	357	169	293
Correct RLR Number	58	90	63	99	32	52
False Alarm Rate	0.9%	1.6%	1.0%	1.7%	1.2%	2.3%
Accurate Predication Rate	48.7%	**75.6%**	49.2%	**77.3%**	43.2%	**70.3%**

Note: *: Standard and Rare represent standard binary logistic regression model and rare events binary logistic regression model, respectively.

From Table 5.4, the result showed that the rare events regression models performed significantly better than standard models. The accurate prediction rate (defined as the ratio between correct RLR prediction number and actual RLR number) jumped from around 45% to near 80% after applying the rare events binary logistic regression methods. Although the false alarm rate (defined as the falsely predicted RLR number over total event number) has been increased, an overall 2.0% false alarm rate is acceptable for real applications.

To further compare the performance of all these models, the Receiver Operating Characteristics (ROC) curves were plotted and the area under the curves (AUC) of these models were calculated. ROC curves and AUCs are frequently used to compare the performance of different methods (Ling et al., 2003). ROC curve directly shows that which model is dominating, and the model with larger AUC also indicates better performance. Fig.5.1 (a), (b) and (c) shows the ROC curves and AUC values for the standard and rare events binary logistic regression models for three intersections, respectively. The figures clearly demonstrated that the rare events logistic regression model was significantly superior to the standard binary logistic regression model.

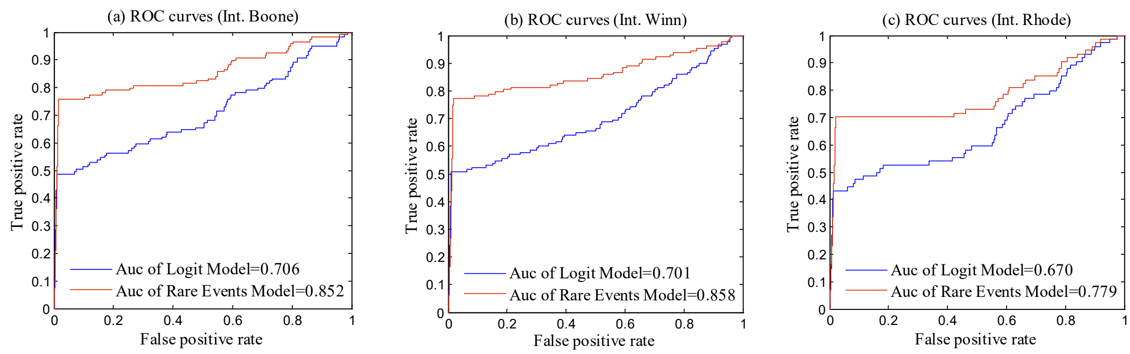


Figure 5.1 ROC plots and AUC values for standard and rare events models

5.3 Summary

For prediction and prevention of the potential RLR, it was important to gain a better understanding of the relationship between RLR and the impact factors which contribute to drivers' RLR behaviors. A large amount of high-resolution traffic and signal data collected from loop detectors was used to extract 9-month's RLR events from three signalized intersections, and then identify the influential factors that significantly affected RLR behaviors. The data analysis indicated that occupancy time, time gap, used yellow time, time left to yellow start, whether the preceding vehicle runs through the intersection during yellow, and whether there was a vehicle passing through the intersection in the adjacent lane are significant factors for determining RLR behaviors.

Furthermore, this research addressed the rare events issue of RLR prediction by developing a rare events binary logistic regression model. To be noted, it was the first time to apply rare events logistic regression for RLR study according to our limited knowledge. The results showed that rare events logistic regression model performed significantly better than standard logistic regression model. The accurate prediction rate jumped from about 45% for standard logistic regression models to near 80% for rare events regression methods. Although the false alarm rate has been increased, an overall 2.0% false alarm rate was still acceptable for real applications. Moreover, the proposed RLR prediction methods were purely based on the loop detector data collected from single advance detectors located 400 feet away from stop-bar. This demonstrated that the proposed models have great potential for future field applications since loops have been widely implemented in most of the intersections and can automatically collect data in real time.

Therefore, this research is expected to contribute to the improvement of intersection safety significantly.

However, although the rare events logistic regression model was superior to the standard logistic regression model, the accuracy of the RLR prediction rate was still not very high, and the false alarm rate was relatively high. One potential reason was that some important factors which could not be collected in real time (i.e. the gender of the driver, age, vehicle type, etc.) were not included in our analysis. Also, some other classification models such as decision trees, linear discriminant analysis (LDA), vector support machine (VSM) could also be applied to generate better prediction results. All these will be left for our future research.

Chapter 6

Conflict Estimation and Intersection Safety Index

All rear-end traffic conflicts for this study were counted for the major road (TH-55) only. Conflicts for the minor approaches were only considered during the crossing conflict stage.

6.1 Rear-end Conflict Identification

A conflict has been identified by calculating the difference in projected distance traveled between two following vehicles. If the projected distance traveled over a period of time, for the leading vehicle, is less than that of the following vehicle, then there is a rear-end conflict. The projected distance could be calculated using simple physics shown in Eq.(1).

$$\boldsymbol{d_P = v * TTC_{max}} \tag{1}$$

where d_P is projected distance, v is individual vehicle speed, and TTC_{max} is maximum time-to-collision. The maximum TTC determines how far into the future the distance will be projected. This value can be set by the user to any value they deem appropriate; however, the farther the distance is projected, the more time there is for a change in trajectory to occur in the real world, which will not be accounted for in the projected distance. A maximum TTC that is too high would result in inaccurate results. A value of 1.5 seconds has typically been suggested in previous research and the SSAM Validation Report [43].

With vehicle speeds calculated as described in Section 2.1, and the gap time between succeeding vehicles known from the detector data, the trajectory of the leading vehicle could be calculated by modifying Eq.(1) to include the gap time as shown in Eq.(2).

$$\boldsymbol{d_1 = v_1 * (g + TTC_{max})} \tag{2}$$

where d_1 is the distance traveled between the time the leading vehicle left the detector and the time the following vehicle arrived plus the specified maximum TTC, v_1 is the leading vehicle's speed, and g is the gap time between the leading and following vehicles. The distance traveled by the following vehicle, d_2, was calculated without the gap time included, similar to Eq.(1). These two distances were then compared, and if the following vehicle's projected distance was greater than the leading vehicle, a rear-end conflict had occurred. Figure 6.1 illustrates Eq.(2) in detail, assuming the following vehicle's speed is higher than the leading vehicle's speed.

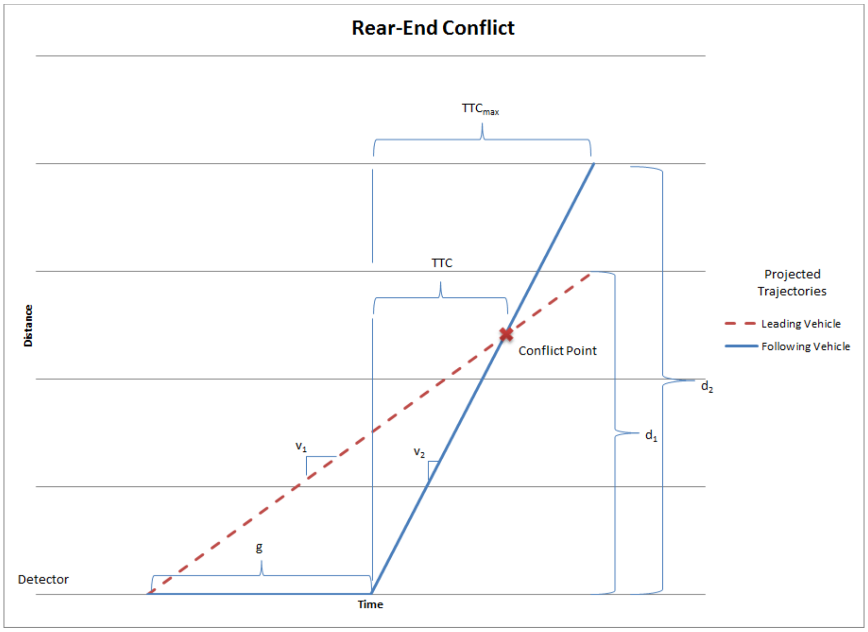


Figure 6.1 Rear-end conflicts

Since TTC max is only a threshold value to define a conflict and not necessarily the actual TTC (the initial time of conflict), TTC must be calculated per Eq.(3)

$$\boldsymbol{TTC} = \frac{\boldsymbol{g * v_1}}{\boldsymbol{v_2 - v_1}} \tag{3}$$

Acceleration data could be added by estimating the following vehicle's trajectory. Since driver's behavior is dynamic and constantly adjusting to the existing situation, assuming a constant speed during the conflict event was not an accurate approach. To help account for this, acceleration rate could be estimated using the "match" method. As explained earlier, the stop detector could be matched to the corresponding advance detector to get two points on the vehicle's trajectory. With data known at two points, constant acceleration could now be applied to determine a more accurate projection for conflict identification. The new calculation for the following vehicle's projection could now be modified as shown in Eq.(4). This step, however, requires data from multiple detectors.

$$\boldsymbol{d_2} = \frac{\boldsymbol{v_A + v_S}}{\boldsymbol{2}} \times (\mathbf{TTC_{max}}) \tag{4}$$

6.2 Crossing Conflict Identification

Crossing conflicts are determined by first splitting the intersection into 4 "Conflict Zones." When two vehicles with crossing trajectories are projected to be in the same Conflict Zone at the same time, a crossing conflict is identified. Figure 6.2 shows how the conflict zones are split.

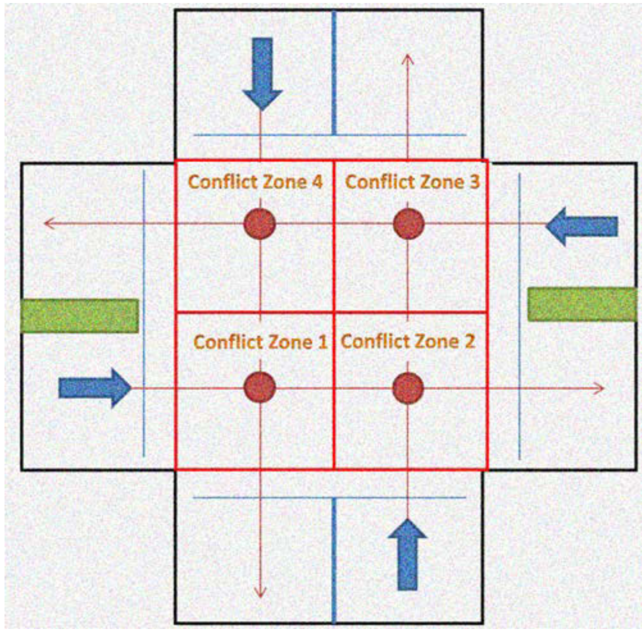


Figure 6.2 Conflict Zones.

To see if two crossing vehicles were inside the same zone, arrival times must be calculated for each vehicle. Two possible trajectory types could be identified here: 1) running through the intersection without stopping, and 2) accelerating from a stopped position. The first case was simple to calculate and was done by Eq.(5).

$$\boldsymbol{AT = Event\ Start\ Time + (d_i/v_S)} \tag{5}$$

where AT is arrival time, event start time is given by the detector data, d_i is the distance to the desired Conflict Zone, and v_S is the speed of the through vehicle at the stop detector.

If the driver stops, then the Entrance detector past the intersection must be used to estimate acceleration from a stopped position. In this case, the speed of the stop-bar detector from Eq.(5) was replaced with the average speed of the Entrance detector and the event start time was replaced with the red phase end time, as shown in Eq.(6).

$$\boldsymbol{AT = Phase\ End\ Time + (2 * d_i/v_E)} \tag{6}$$

After calculating the arrival time for every conflict zone that the driver would be passing, their arrival times were compared to the arrival times of the crossing vehicles. If the arrival time for a conflict zone is within the same range as the arrival time of a crossing vehicle for the same conflict zone, then a crossing conflict is identified. It should be noted that every case of crossing conflict must be the result of an RLR either from the minor or the major road. This method, however, required the use of stop-bar and entrance detectors. It would be more economical if merely the advance detectors could be used to estimate crossing conflicts. If only one detector has to be used to determine crossing conflicts, then it is necessary to predict the RLR as discussed in detail in previous chapters.

6.3 Intersection Safety Evaluation

The overall intersection safety level was measured by the number of conflicts (including both rear-end and crossing) and RLR events. This section provides our evaluation results for three target intersections.

6.3.1 Conflict and RLR Counts

Applying the methods introduced in Section 3, we were able to estimate the number of potential conflicts and RLR events. A complete summary of each intersection is shown in Table 6.1**Error! Reference source not found.**.

The results indicated that Winnetka would have the most rear-end accidents. Due to such low number of crossing conflicts detected, it is hard to determine which intersection would experience the highest number of right-angle crashes. A large number of RLR cases at Rhode indicated that there could be a high risk of right-angle crashes, but since it was a T-intersection, no crossing conflicts were tested using the current methodology (which could only test two through crossing events, not turning). Overall, Winnetka had the highest conflict counts and second largest RLR counts. This puts Winnetka at the greatest risk of future collision when compared to Boone and Rhode. Boone had nearly double the rear-end conflicts experienced at Rhode, indicating that it has a much higher risk of rear-end collision than Rhode.

Table 6.1 Summary of conflict and RLR counts.

		General Info		Conflicts		Total Red-Light Runners
		Total Events	Matched Events	Adjusted* Rear-end	Crossing	
Boone	May	730581	491143	7483	3	279
	Jun	899404	545307	8577	0	380
	Nov	757919	477848	6915	0	289
	Dec	728995	461945	4815	0	307
	Feb	713062	476277	6246	0	245
	Jan	615181	368554	3894	0	219
	Jul	520697	341678	5400	0	205
	Average	**709406**	**451822**	**6190**	**0**	**275**
Winn	Jan	553937	268292	5546	0	169
	Jun	752600	369237	9122	2	445
	Nov	596693	294523	6895	4	630
	Apr	625528	332482	8134	1	250

	Aug	719709	369379	9126	0	210
	Mar	488305	199878	4840	0	129
	May	749312	368628	9101	2	444
	Dec	688102	355429	6187	0	304
	Jul	609887	297965	7327	0	199
	Feb	379794	113929	2895	0	86
	Average	**616387**	**296974**	**6917**	**1**	**287**
Rhode	Jan	748598	463677	3171	-	520
	Feb	721246	432237	3462	-	291
	Dec	811497	520209	3175	-	736
	Nov	653244	351207	3082	-	355
	Average	**733646**	**441833**	**3223**	**-**	**476**

*Average daily traffic adjustments were made to account for the discrepancies in the number of successful matches vs. actual traffic to get a conflict count which better reflects the intersection.

6.3.2 Crash Rate Estimation & Comparisons to Actual Crash Data

Conflicts were be converted to crashes per year by Eq.(7) from the SSAM Validation Report [9]. The results are shown in Table 6.2.

$$\textbf{Crashes per year} = \mathbf{0.119} * \left(\frac{\textbf{Conflicts}}{\textbf{Hour}}\right)^{\mathbf{1.1419}} \qquad \textbf{(7)}$$

The Minnesota Department of Transportation Metro District was able to provide crash data between the years 2006 through 2011 for the three intersections being studied. The data was obtained from the Minnesota Department of Public Safety (DPS) database; therefore, only those crashes reported to the DPS were included in the data. A summary of the crash count taken from data collected from the Minnesota DOT Metro District is shown in Table 6.2. These counts were for intersection-related crashes only. Rear-end collisions were considered only in the east and west directions to compare to the conflicts recorded by the detector data on TH55.

Table 6.2 Estimated number of rear-end crashes based on conflict results.

Intersection Name	Rear-end Conflicts			Actual Number of Rear-end Crashes (2006-2011)
	Estimated Conflicts per hour	Estimated Crashes per year	Estimated Crashes in 6 years	
Boone	8.60	2.52	15	11
Winnetka	9.61	2.95	18	15
Rhode	4.48	1.00	6	2

A comparison of the rear-end conflict counts to the historical crash data showed an excellent correlation. The rear-end conflicts matched up very well. It would have taken about six years of crash data to notice that Winnetka has the highest risk for rear-end accidents. But automatically collected conflict information, revealed the pattern in only a few months. The numbers do tend to fluctuate year to year, however, which further supports the need for a real-time safety evaluation that can address the current safety concerns before a crash would happen.

Chapter 7

Intersection Safety Evaluation System Implementation - Website Development

As an outcome of this project, a website (www.intersectionsafety.com) was designed to further support and share the work done so far. Although it does not use real time data, it features the data used in this project and provides the user with an interface to find the number of RLR cases and conflicts for various intersections. All the essential features of the website have been presented in this chapter.

7.1 Homepage

The site introduces the concept of intersection safety using high-resolution traffic data. It allows direct access to web pages where you can understand and estimate RLR, conflicts, trajectories, queue length, safety index and traffic pollution based on data collected from certain intersections. Figure 7.1 shows a preview of website homepage.

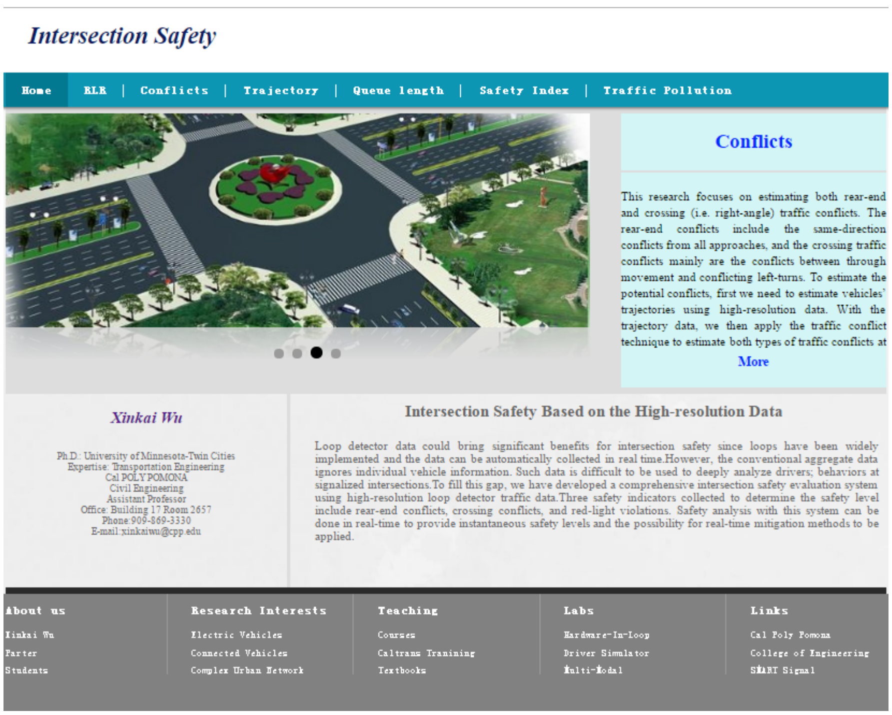


Figure 7.1 Homepage: www.intersectionsafety.com

The website is in development mode and being updated time-to-time based on research done so far. The main sections directly related to the research project that allows user to estimate various aspects of intersection safety through a Graphic User Interface (GUI) has been discussed briefly in the following sections.

7.2 RLR

Red Light Running is an important issue while evaluating safety index of an intersection. This kind of driving behavior directly or indirectly acts as a hazard to other vehicles and often ends up as a collision. This web page allows a user to estimate the number of RLR cases in all directions at an intersection based on the high-resolution event data collected by loop detectors. It gives a brief information about the idea of detecting RLR as discussed in previous chapters. A user can follow simple steps to estimate RLR cases:

1) On the left side pane, choose an intersection.
2) Fill the desired start time and end time for estimation. Make sure the entered period is within the range of available data. You can check the dates for available data on the top right corner of the map.
3) Click the button: RLR number.

The map will display the number of RLR in each direction next to the arrows on the map. An estimate of RLR for one day at Boone Ave is shown in figure 7.2.

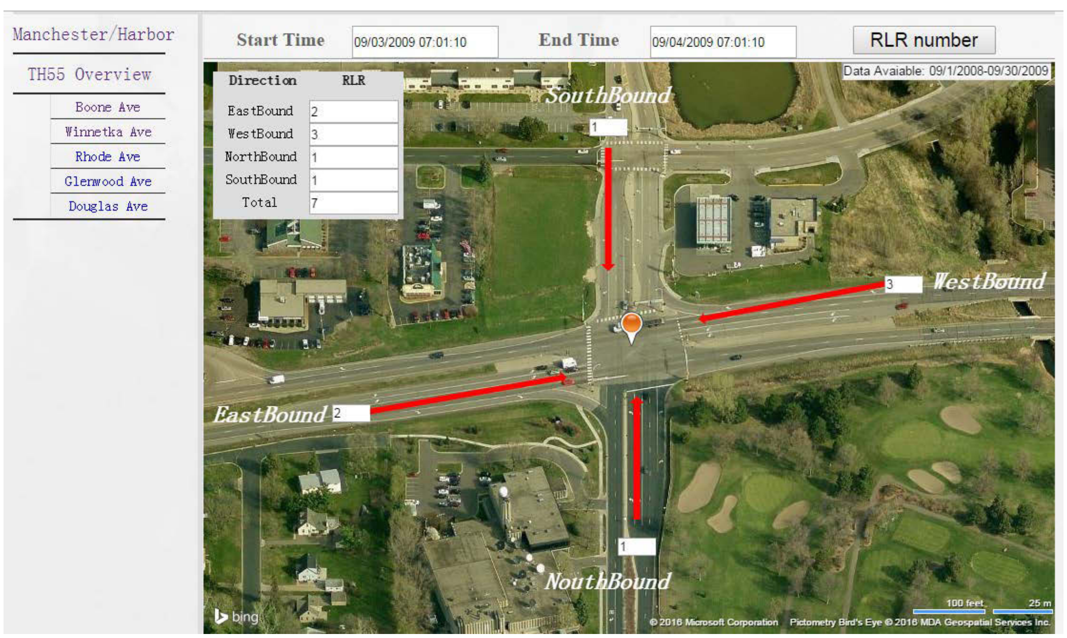


Figure 7.2 RLR estimate at Boone Ave.

Using the data from website, RLR cases were analyzed at Manchester Ave. and Harbor Blvd. for October 2015 as shown in Figure 7.3. This intersection is a T-intersection with traffic moving Northbound, Southbound and Westbound. It was found that Most of the RLR cases occurred in Northbound-Southbound direction. Also, RLR cases were generally higher on Friday, Saturday and Sunday.

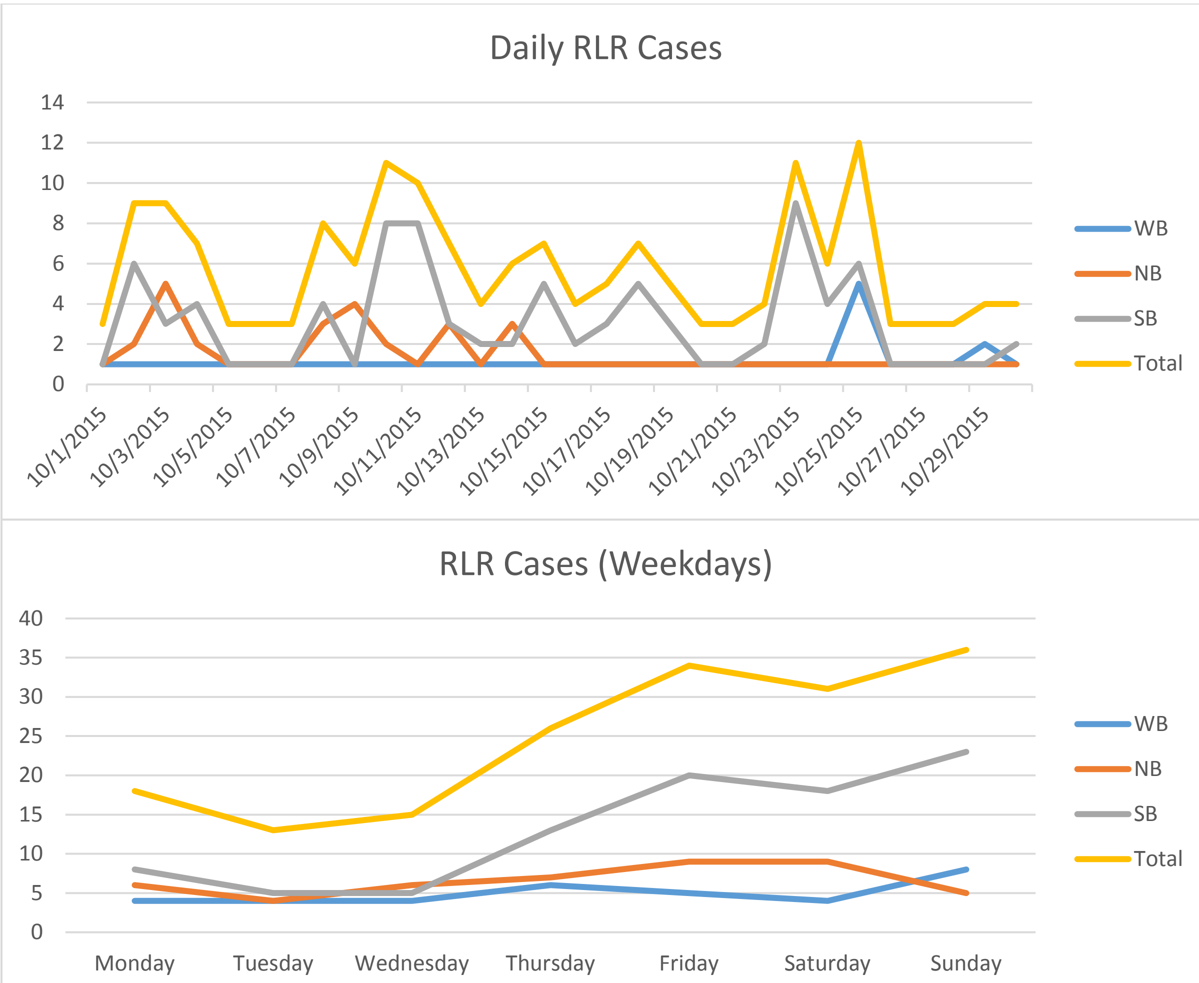


Figure 7.3 RLR cases for 1 month at Manchester Ave. and Harbor Blvd.

7.3 Conflicts

Another important aspect from safety point of view is potential conflicts that can occur at an intersection. Traffic signal design well addresses most of the conflict zones, but events like RLR or YLR can potentially lead to a collision in these conflict zones. Hence, it is good to have an estimate of a potential number of conflicts at an intersection to apply protective measures. The web page for conflicts also gives an introduction the methodology to estimate conflicts. As mentioned in previous chapters, all rear-end traffic conflicts has been counted for the major road (TH-55) only. Conflicts for the minor approaches were only considered during the crossing conflict stage. Similar steps can be followed to estimate conflicts as described for RLR. An example showing the number of conflicts for one day at Winnetka Ave is provided in Figure 7.4.

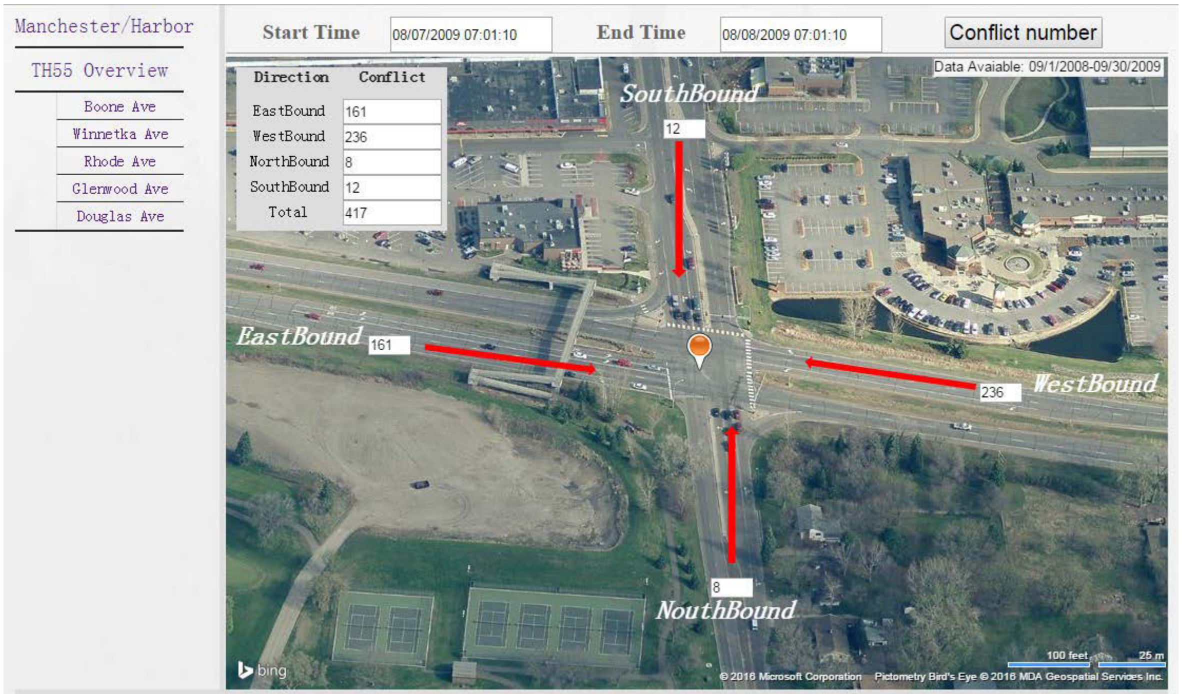


Figure 7.4 Conflict estimate at Winnetka Ave.

Using the data from website, RLR cases were analyzed at Manchester Ave. and Harbor Blvd. for October 2015 as shown in Figure 7.5. It was found that Maximum number of conflicts occurred mostly on Fridays. In addition, majority of conflits were in Nothbound-Southbound Direction. This may be due to the T-shape of the intersection. Vehicles on perpendicular leg of intersection have to be more careful and hence conflicts were less as compared to NB direction.

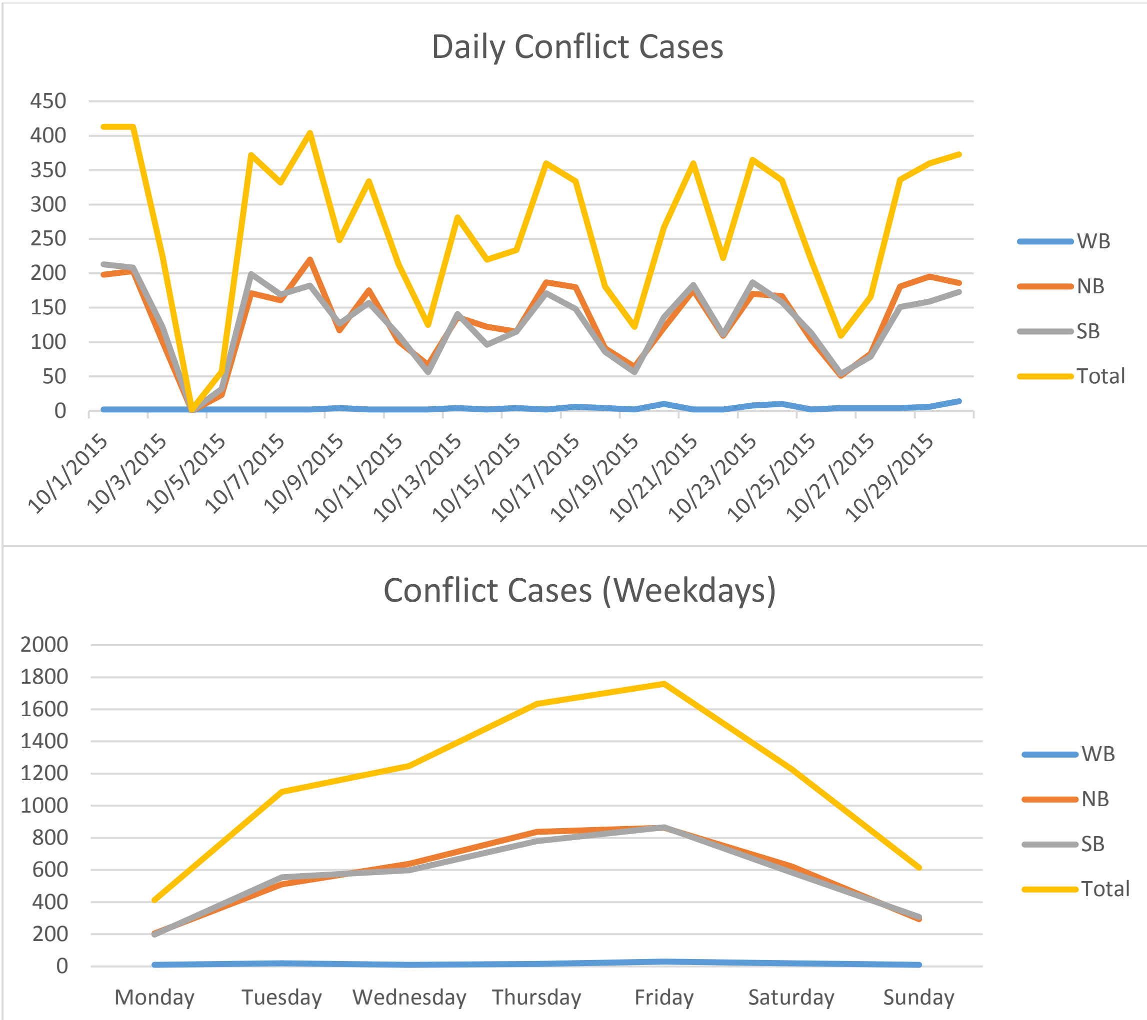


Figure 7.5 Conflicts for 1 month at Manchester Ave. and Harbor Blvd.

7.4 Safety Index

Safety Index is a combination of conflicts and RLR events occurring at an intersection that gives us an overall estimate of the level of safety for an intersection. Similar steps need to be followed as in case of RLR or conflict estimation to find the safety index for an intersection. Figure 7.6 shows an example of Safety Index evaluation at Glenwood Ave for 2:15 hours.

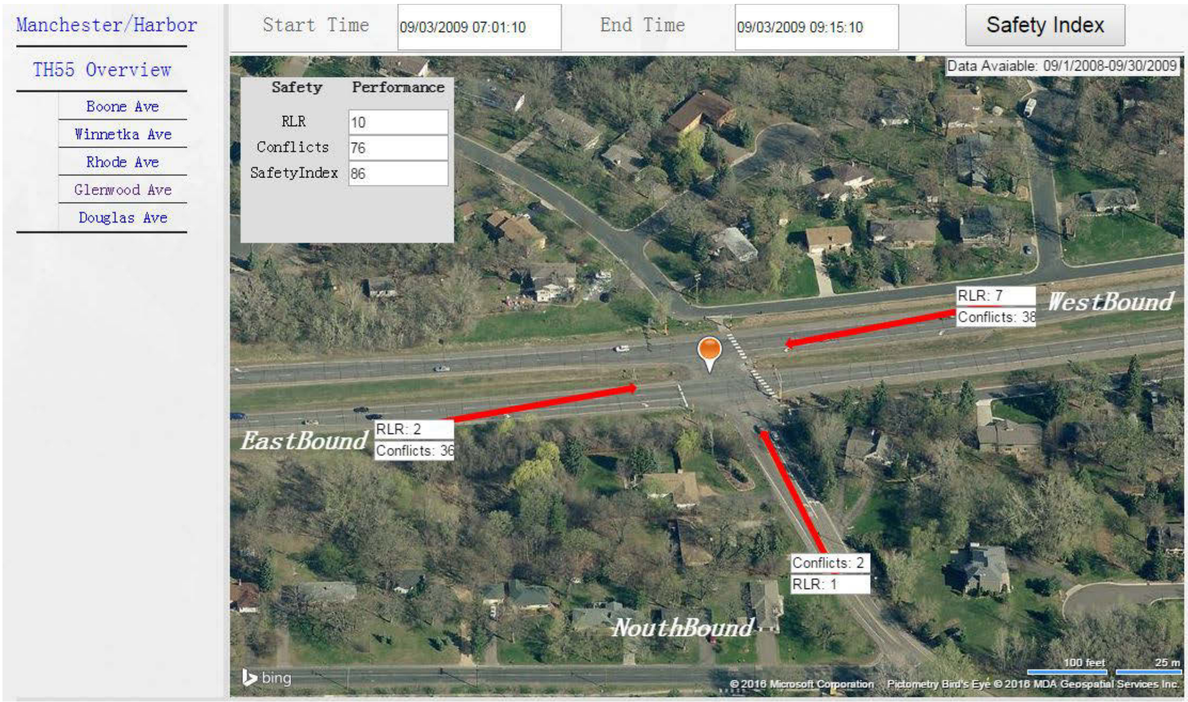


Figure 7.6 Safety index at Glenwood Ave

Also, Safety Index was analyzed at Manchester Ave. and Harbor Blvd. for October 2015 as shown in Figure 7.5. It was found that Safety Index on Fridays was very high. This is consistent with high number of conflicts on the same intersection. As seen earlier in case of conflicts, safety index was higher in Nothbound-Southbound Direction. The reason may be, again, the T-shape of the intersection. Vehicles on perpendicular leg of intersection have to be more careful and hence conflicts were less as compared to NB direction.

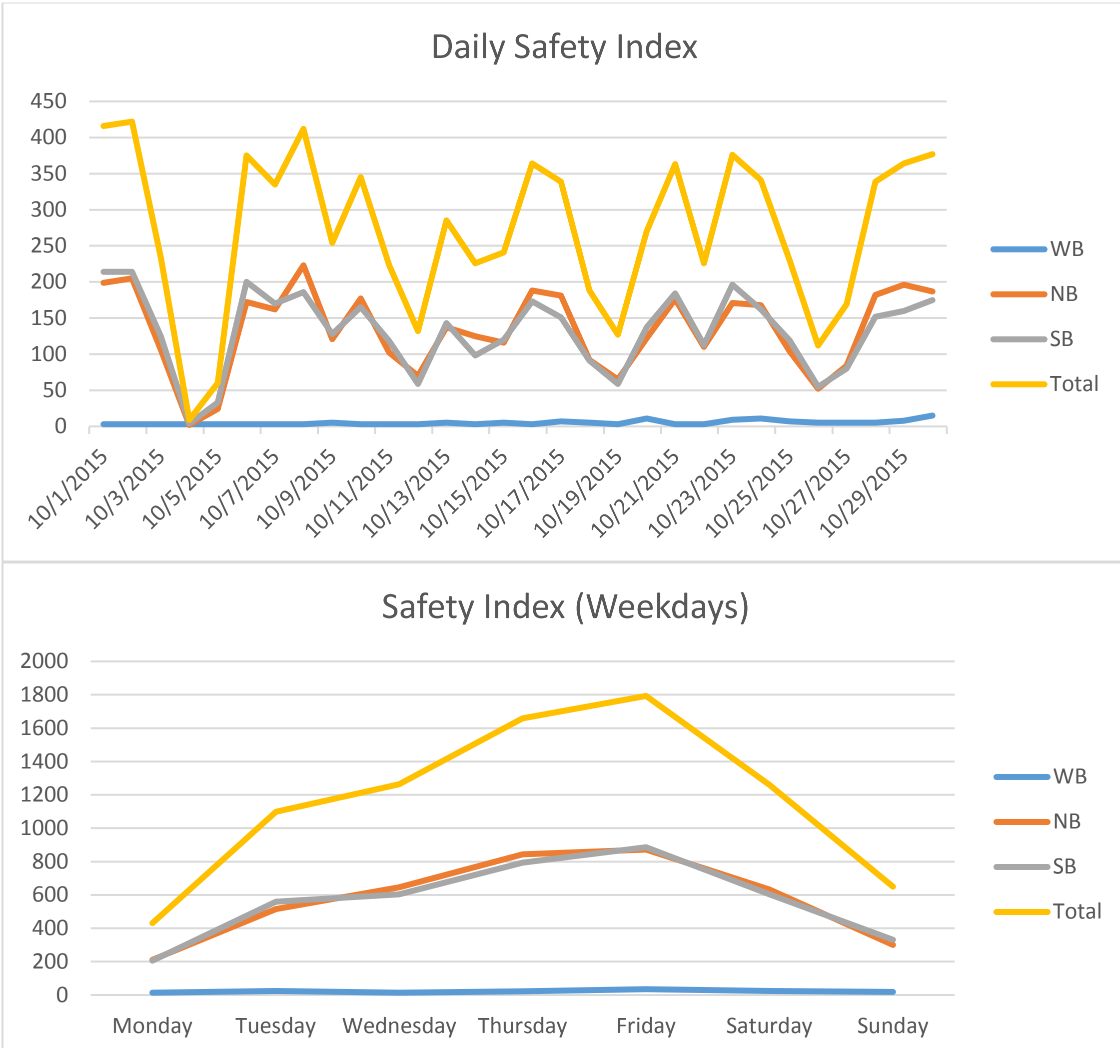


Figure 7.7 Safety Index for 1 month at Manchester Ave. and Harbor Blvd.

7.5 Trajectories

The vehicle trajectory data, as a type of traffic microdata, has attracted more and more researchers' attention. It can be used for various transportation research, e.g. car following behavior, vehicle lane changing behavior and road safety analysis, etc.

A vehicle detection and tracking system based on CCTV cameras is developed for collecting trajectory data. A deep learning based method, called convolutional neural network (CNN), is used to achieve high precision vehicle detection and tracking as shown in figure 7.8. This system can identify multi-type vehicles, e.g. car, vans, truck, and motorcycle. Pedestrian detection and tracking are also supported. This system is supposed to build a big-data system for traffic research.

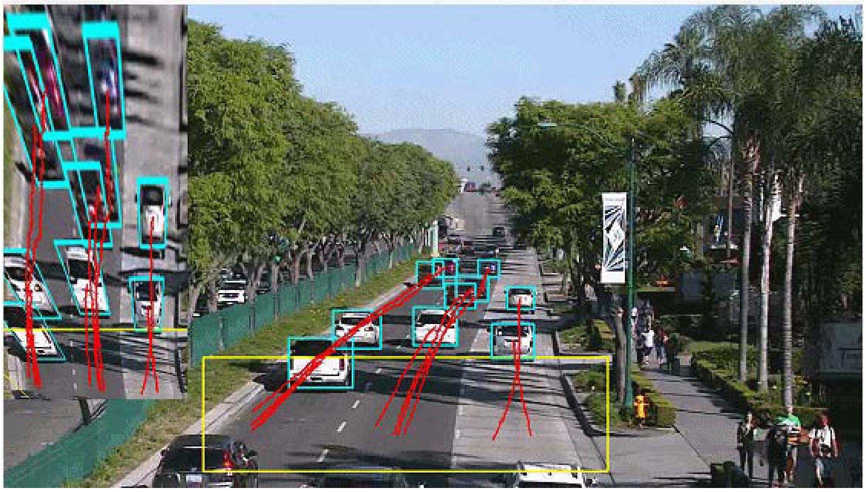


Figure 7.8 Trajectory data collected by CCTV cameras.

7.6 Queue Length

Using the event-based data, a set of arterial performance measures, especially intersection queue length and arterial travel time, can be estimated. Queue length is unarguably the most important performance measure at a signalized intersection since other performance measures such as average delay, and level of services can be easily derived from queue length information. A major shortcoming of traditional input-output models for estimating queue length has been their inability to determine queue length under saturated conditions—i.e. when the queue of cars waiting to pass through an intersection extends beyond the upstream vehicle detector. Under saturated conditions, data on incoming traffic flow are no longer available, and the input side of the input-output model breaks down. This limitation was overcome by developing a new algorithmic approach to estimate queue length based on traffic shockwave theory.

This method first utilizes the high-resolution data collected by the SMART-Signal to identify the changes of traffic states and then applies Lighthill-Whitham-Richards (LWR) shockwave theory to construct shockwave profiles. The figure below demonstrates the shockwave pattern within a cycle. This pattern consists of four shockwaves, and the shockwave motion will repeat from cycle to cycle. Using the high-resolution data, the changes of traffic states, i.e. the "breakpoints" (A, B, and C), can be identified as shown in figure 7.9. The time-dependent queue length including the maximum and minimum queue (if existing) can then be easily derived from the constructed shockwave profile.

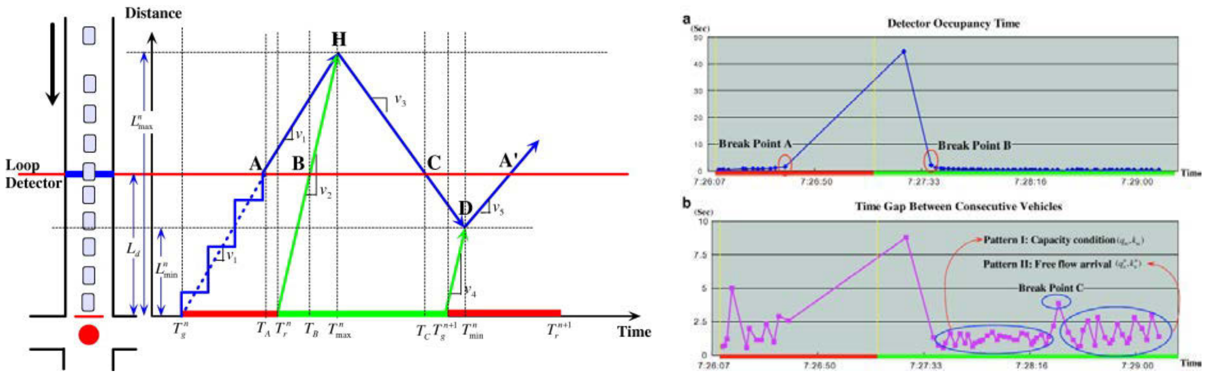


Figure 7.9 Queue length estimation.

The user can estimate the queue length at various intersections using the queue length web page. Instructions to estimate queue length are as follows:

1) On the left side pane, choose an intersection.
2) Fill the desired start time and end time for estimation. Make sure the entered period is within the range of available data. You can check the dates for available data on the top right corner of the map.
3) Check mark the direction in which you want to estimate the queue length.
4) Click the button: Queue.

A time-space graph will be plotted with estimated length of queue during the period entered. An example showing estimated queue length for Boone Ave is as illustrated in figure 7.10.

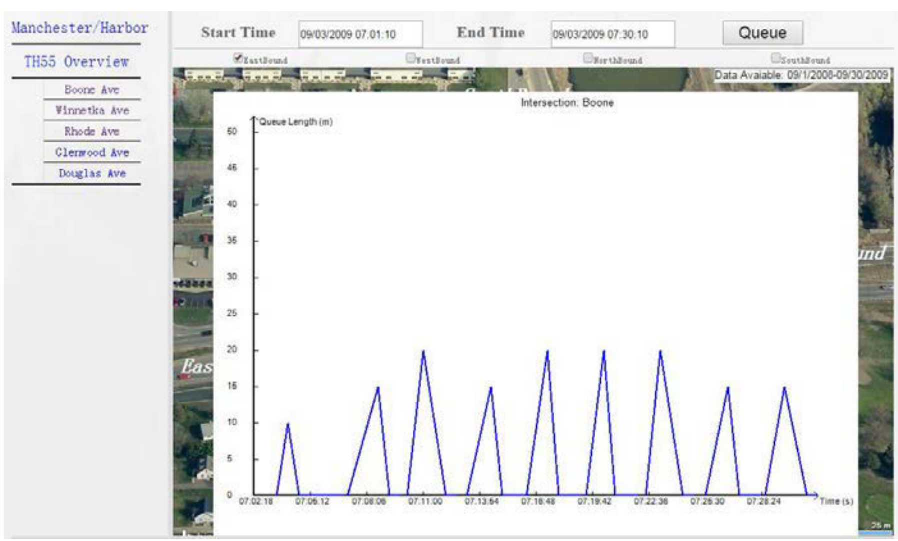


Figure 7.10 Estimated queue length for Boone Ave.

Chapter 8

Conclusions and Future Research

This report presented our recent findings on the safety evaluation for signalized intersections. First, this research studied drivers' stop-or-run (SoR) behavior at signalized intersection when the signal changes to yellow. In contrast to most previous research, this research uses high-resolution traffic and signal event data collected from loop detectors. A simple method is first proposed to identify first-to-stop, yellow-light running and red-light running cases using the information from both stop bar and advance detectors. Then the data collected from advance detectors (including occupancy time and time gap), signal information (including used yellow time and the time left to yellow start), as well as the information from three preceding vehicles and vehicles on adjacent lanes are applied to identify the factors that significantly impact drivers' decision making. A binary logistical regression method is implemented to analyze the significance of all these factors. The investigation results show that occupancy time, time gap, used yellow time, time left to yellow start, the time gap between the first two preceding vehicles, whether the nearest preceding vehicle runs through intersection during yellow, and whether there is a vehicle passing through on an adjacent lane show significant impact on drivers' decision. A prediction model, which predicts if a driver SoR through the intersection, is also developed using the information collected by advance detectors. The testing experiment shows the accuracy of the model is as high as 87%.

Second, this research presented a new method to predict red-light-running (RLR). To predict and prevent the potential RLR, it is important to gain a better understanding of the relationship between RLR and the impact factors which contribute to drivers' RLR behaviors. This research addresses the rare events issue of RLR prediction by developing a rare events binary logistic regression model. To be noted, it is the first time to apply rare events logistic regression for RLR study according to our limited knowledge. The results show that rare events logistic regression model perform significantly better than standard logistic regression model. The final prediction rate jumps from about 45% for standard logistic regression models to near 80% for rare events regression methods. Although the false alarm rate has been increased, an overall 2.0% false alarm rate is still acceptable for real applications. Moreover, the proposed RLR prediction methods are purely based on the loop detector data collected from single advance detectors located 400 feet away from stop-bar. This demonstrates that the proposed models have great potential for future field applications since loops have been widely implemented in most of the intersections and can automatically collect data in real time.

Third, this research developed a comprehensive Intersection Safety Evaluation System using high-resolution loop detector traffic data. The safety indicators collected to determine the safety level include traffic conflicts (both rear-ending and crossing conflicts) and red-light violations.

All these measurements are automatically collected by the detection system using the programmed algorithms. Safety analysis with this system can be done in real-time to provide immediate safety levels and the possibility for real-time mitigation methods to be applied. Results provide valuable information for evaluating the safety of each intersection. Future rear-ends crashes can be estimated and potentially avoided if countermeasures are taken at high-risk intersections. Crossing conflicts and RLR counts can indicate driver behavior patterns at each intersection and can be used to optimize signal timing operations for safety (i.e. synchronize signals at intersections with high RLR to reduce the number of vehicle arrivals during yellow).

Overall, this research laid a solid foundation for intersection safety evaluation. There were some successes, such as predicting the SoR behavior of drivers with an 88% accuracy rate, and accurately portraying rear-end accidents through rear-end conflicts. The prediction of red-light running produced decent success rates of 65%-75%, but the current model predicts too many "false positives" (i.e., too many FSTP cases are predicted as RLR). The SoR model could be improved to include other relevant impact factors, and travel time estimation could be enhanced with the use of additional loop detectors to determine the vehicle's trajectory better, and some other classification models such as decision trees, linear discriminant analysis (LDA), vector support machine (VSM) could also be applied to generate better RLR prediction results.

Overall, this research showed that high-resolution loop detector data has a lot of potential for use in intersection safety applications. Not only can intersection-related data be monitored continuously, but also future applications could make use of the real-time nature of the data for collision avoidance or warning systems. Future research could focus on the optimal placement of loop detectors for better safety analysis. Advanced loop detectors could be placed closer to the stop bar to avoid confusing a left turning vehicle with a through vehicle. This would improve the detector matching algorithm as well as trajectory estimations. The quality of data could also be improved by adding more information relating to traffic safety. Information such as illumination levels and weather conditions could help with understanding driver behaviors. Further development of the algorithms must also include turning maneuvers.

More importantly, the proposed system can be applied to build a real-time intersection collision avoidance system, and the network-wide intersection safety information can be used to rank the safety severity for all intersections in the network, therefore helps agencies identify the intersections which need most improvement. So our next research topic will be to incoportate the proposed safety evaluation system with emerging connected vehicle (V2V and V2I) technologies to build a dynamic system that could alert drivers of emerging and impending hazardous situations at signalized intersections.

References

1. FHWA (2009), http://safety.fhwa.dot.gov/intersection/, visited on July 30th, 2012
2. National Highway Traffic Safety Administration (2008) *NHTSA's Fatality Analysis Reporting System (FARS) reports*, http://www.fars.nhtsa.dot.gov/Main/index.aspx,visited July 30th, 2012
3. National Highway Traffic Safety Administration (2008) *Traffic Safety Facts 2008 Report*, http://www-nrd.nhtsa.dot.gov/Pubs/811170.PDF, visited on July 30th, 2012
4. National Highway Traffic Safety Administration (2004), *The National Survey of Speeding and Other Unsafe Driver Actions Vol. 2: Findings*, Report No. DOT HS 809 730, National Highway Traffic Safety Administration, May 2004.http://www.nhtsa.gov/people/injury/aggressive/unsafe/att-beh/cov-toc.html, visited July 30th, 2012
5. Retting, R. A., Williams, A. F., Preusser, D. F., and Weinstein, H. B. Classifying urban crashes for countermeasure development. Accident Analysis & Prevention, 1995, 27(3), 283-294.
6. Retting, R. A., and Williams, A. F. Characteristics of red light violators: results of a field investigation. Journal of Safety Research, 1996, 27(1), 9-15.
7. Wang, X., Yu, R., and Zhong, C. A field investigation of red-light-running in Shanghai, China. Transportation Research Part F: Traffic Psychology and Behaviour, 2016, 37, 144-153
8. FHWA (2009) Intersection Safety Issue Briefs FHWA-SA-10-005, available at http://safety.fhwa.dot.gov/intersection/resources/fhwasa10005/brief_2.cfm
9. Bonneson, J. A., Middleton, D.R., Zimmerman, K.H., Charara,H.A., & Abbas, M.M. (2002) *Intelligent Detection-Control System for Rural Signalized Intersections*. Report 4022-2. Texas Transportation Institute, College Station.
10. Ragland, D. R. & Zabyshny, A.A. (2003). *Intersection Decision Support Project: Taxonomy of Crossing-Path Crashes at Intersections Using GES 2000 Data*. Report UCB-TSC-RR-2003-08. University of California, Berkeley.
11. Neale, V.L., Perez, M.A., Lee, S.E., & Doerzaph, Z.R. (2007).*Investigation of Driver–Infrastructure and Driver–Vehicle Interfaces for an Intersection Violation Warning System*. Journal of Intelligent Transportation Systems, Vol. 11, pp. 133–142.
12. Wang, L., Zhang, L., Zhou, K., Zhang, W., and Wang, X. Prediction of Red-Light Running on Basis of Inductive-Loop Detectors for Dynamic All-Red Extension. Transportation Research Record: Journal of the Transportation Research Board, 2012(2311), 44-50.
13. FHWA (2004) Signalized Intersections: Informational Guide, FHWA-HRT-04-091, available at http://www.fhwa.dot.gov/publications/research/safety/04091/04091.pdf
14. AASHTO (2010). Highway Safety Manual. AASHTO, Washington, D.C., 2010.
15. Retting, R. A., Ulmer, R.G., & Williams, A.F. (1999).*Prevalence and characteristics of red light running crashes in the United States*. Accid. Anal Prev., 31(6), pp.87-694.
16. Wissinger, L.M., Hummer, J.E., & Milazzo II, S.M. (2000) *Using Focus Groups to Investigate Issues of Red Light Running*, Transportation Research Record 1734, pp.38-45
17. Porter, B. E. & T. D. Berry (2001).*A Nationwide Survey of Self-Reported Red Light Running: Measuring Prevalence, Predictors, and Perceived Consequences*. Accident Analysis and Prevention, 33, pp. 735-741.
18. Bonneson, J. A., Brewer,M.A., & Zimmerman, K.H. (2001).*Review and Evaluation of Factors that Affect the Frequency of Red-Light-Running*. Report FHWA/TX-02/4027-1. Texas Transportation Institute, College Station.
19. Bonneson, J. A. & Zimmerman, K.H. (2004).*Effect of Yellow-Interval Timing on the Frequency of Red-Light Violations at Urban Intersections*, Transportation Research Record 1865, pp. 20-27.

20. Gates, T. J., Noyce, D. A., Laracuente, L., & Nordheim, E.V. (2007). *Analysis of Driver Behavior in Dilemma Zones at Signalized Intersections*. Transportation Research Record 2030, pp. 29-39.
21. David, C.Y. & Najm, W.G. (2007) *Examining Driver Behavior using Data Gathered from Red Light Photo Enforcement Cameras*, Journal of Safety Research 38, pp. 311–321
22. Porter, B. E., and England, K. J. Predicting red-light running behavior: a traffic safety study in three urban settings. Journal of Safety Research, 2000, 31(1), 1-8.
23. Bonneson, J. A. & Son, H.J. (2003) *Prediction of Expected Red-Light-Running Frequency at Urban Intersections*, Transportation Research Record 1830, pp. 38-47
24. Yang, C. D., and Najm, W. G. Examining driver behavior using data gathered from red light photo enforcement cameras. Journal of safety research, 2007, 38(3), 311-321.
25. Wang, L., Zhang, L., Zhang, W.B. and Zhou, K. Red light running prediction for dynamic all-red extension at signalized intersection. Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on (pp. 1-5). IEEE.
26. Zhang, L., Zhou, K., Zhang, W., and Misener, J. A. Prediction of Red Light Running Based on Statistics of Discrete Point Sensors. TRANSPORTATION RESEARCH RECORD, 2009(2128), 132-142.
27. Elmitiny, N., Yan, X., Radwan, E., Russo, C., and Nashar, D. Classification analysis of driver's stop/go decision and red-light running violation. Accident Analysis & Prevention, 2010, 42(1), 101-111.
28. Smaglik, E.J., Sharma, A., Bullock, D.M., Sturdevant, J.R., & Duncan, G. (2007).*Event-Based Data Collection for Generating Actuated Controller Performance Measures*. Transportation Research Record 2035, pp. 97–106.
29. Liu, H., Wu, X., Ma, W., & Hu, H. (2009). *Real-Time Queue Length Estimation for Congested Signalized Intersections*. Transportation Research-Part C. 17(4), pp. 412-427.
30. Chatterjee, I. & Davis, G. (2011) *Using High-Resolution Detector and Signal Data to Support Crash Identification and Reconstruction*, presented at the 90th Transportation Research Board (TRB) Annual Meeting, Washington, D.C.
31. Sheffi, Y. & Mahmassani, H. (1981).*A Model of Driver Behavior at High Speed Signalized Intersections*. Transportation Science, 15(1) pp. 50–61.
32. Bonneson, J. A., McCoy, P.T., & Moen., B.A. (1994*)Traffic Detector Design and Evaluation Guidelines*. Report TRP-02-31-93. Nebraska Department of Roads, Lincoln.
33. Haque, M. M., Ohlhauser, A. D., Washington, S., and Boyle, L. N. Decisions and actions of distracted drivers at the onset of yellow lights. Accident Analysis & Prevention, 2015.
34. Newton, C., Mussa, R. N., Sadalla, E. K., Burns, E. K., and Matthias, J. Evaluation of an alternative traffic light change anticipation system. Accident Analysis & Prevention, 1997, 29(2), 201-209.
35. Papaioannou, P. Driver behavior, dilemma zone and safety effects at urban signalized intersections in Greece. Accident Analysis & Prevention, 2007, 39(1), 147-158.
36. Lu, G., Wang, Y., Wu, X., Liu, H. (2015) Analysis of yellow-light running at signalized intersections using high-resolution traffic data. Transportation research part A: policy and practice 73:39-52.
37. NHTSA. (2009). Crash Factors in Intersection-Related Crashes: An On-Scene Perspective, DOT HS 811 366, H, Washington, DC: Available at http://www-nrd.nhtsa.dot.gov/Pubs/811366.pdf
38. Chawla, N. V., Japkowicz, N., and Kotcz, A. Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter, 2004, 6(1), 1-6.
39. He, H., and Garcia, E. A. Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on, 2009, 21(9), 1263-1284.
40. King, G., and Zeng, L. Logistic regression in rare events data. Political analysis, 2001, 9(2), 137-163.
41. FHWA, (1990). Traffic conflict techniques for safety and operations. Course Materials. NHI course 38059. USDOT, Washington, DC.

42. Perkins, D. & Harris, J. (1967). Criteria for traffic conflict characteristics, Report GMF 632. Warren, MI. General Motors Corp.
43. Gettman, D., Pu, L., Sayed, T., and Shelby, S. (2008), Surrogate Safety Assessment Model and Validation: Final Report, FHWA-HRT-08-051, available at http://www.fhwa.dot.gov/publications/research/safety/08051/08051.pdf. Accessed on February 26, 2013.
44. Institute of Transportation Engineers (1999).*Traffic Engineering Handbook*. Institute of Transportation Engineers, Washington, DC.