Center for Advanced Multimodal Mobility Solutions and Education

Project ID: 2019 Project 06

CORRIDOR LEVEL ADAPTIVE SIGNAL CONTROL CONTINUATION

Final Report

by

Hao Liu, Ph.D. (ORCID ID: http://orcid.org/0000-0002-5319-9514)
The University of Texas at Austin

and

Randy Machemehl, Ph.D., P.E. (ORCID ID: https://orcid.org/0000-0002-6314-2626)
Associate Professor, Department of Civil and Environmental Engineering
The University of Texas at Austin
301 E. Dean Keeton Street, Stop C1761, Austin, TX 78712
Phone: 1-512-471-4541; Email: rbm@mail.utexas.edu

for
Center for Advanced Multimodal Mobility Solutions and Education
(CAMMSE @ UNC Charlotte)
The University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223

September 2020

ACKNOWLEDGMENTS

This project was funded by the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE @ UNC Charlotte), one of the Tier I University Transportation Centers that were selected in this nationwide competition, by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT), under the FAST Act. The authors are also very grateful for all of the time and effort spent by DOT and industry professionals to provide project information that was critical for the successful completion of this study.

DISCLAIMER

Table of Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

With the increasing number of vehicles, most cities around the world are suffering from traffic congestion. Transportation practitioners can use a variety of strategies to address the congestion problem, such as traffic signal control, constructing new roads, allocating dedicated lanes for public transit, and road space rationing. Out of these solutions, traffic signal control provides the most cost-effective way to mitigate congestion. Traffic signal control strategies maximize traffic mobility without the monetary cost of acquiring new infrastructure.

Signal control falls into two groups: pre-timed signal control and adaptive signal control. In the pre-timed control method, historical data informs signals timing decisions for specific times of day. Unlike the pre-timed control, adaptive signal control optimizes signal timing plans based on real-time traffic conditions. Thanks to its ability to adjust and respond to the prevailing traffic condition, adaptive control has superior capabilities to pre-timed control. Although adaptive signal control offers promise in reducing traffic congestion, it is not very popular in urban networks because of its complexity. For instance, the City of Austin manages about 1,020 traffic signals and almost all of them are pre-timed. Therefore, adaptive signal control is still a very worthy research topic.

The core elements of an adaptive signal control method include a traffic volume prediction model and a signal optimization model. An adaptive control method must produce accurate predictions. In many adaptive signal control methods, upstream traffic detectors frequently predict traffic counts. However, many locations do not have traffic detectors widely deployed. A common way practitioners deal with this situation is by using historical data to predict traffic volumes. To derive a simple traffic prediction model, the vehicle arriving at the target intersections are usually approximated as simple patterns such as linear arrivals (constant interarrival time) and as platoon forms with known platoon size and arrival time. However, the real arrival process may disagree with these assumptions especially when the traffic volumes are low. So, any errors in the traffic volume prediction will reduce the efficiency of the adaptive signal control.

Using queueing theory, this study proposes an adaptive signal control algorithm to optimize signal timing for an isolated intersection under low traffic volume conditions. The algorithm models vehicle arrivals for all approaches at the intersection as Poisson processes, i.e., the interarrival times are exponential distributions. In addition, the algorithm assumes departure rates for all approaches are constant. Thus, we model the queues at the target intersection as M/D/1 queues. Note that the departure rates of these queues are not constant since traffic conditions switch between saturation flow and zero flow, matching the traffic signal switching between green and red indications. For simplicity, we call these queues as M/D/1 queues. In queueing theory, a polling system is a system consisting of multiple queues accessed in cyclic order by a single server. Previous work has proposed a number of analytical models [1]–[3] to study properties of polling systems. Three commonly used service policies for a polling system include: exhaustive service policy where a queue continuous to receive service until it is emptied; gated service policy where only the customers in the queue at the instant that the server arrived, but new arrivals during this service time must wait until the next server visit; and limited service policy where a fixed maximum number of customers can be served in each visit. Under

most cases, the exhaustive service policy has the best performance in terms of reducing the waiting time or delay. We demonstrate the efficiency of the proposed model by comparing its traffic delay to that resulting from the Webster's model, which is a fixed time control method, and to that resulting from an exhaustive policy adaptive signal control method. The comparison shows that the proposed method can reduce the traffic delay significantly under low traffic volumes.

Chapter 1. Introduction

1.1 Problem Statement

With increasing car ownership, congestion has become a worldwide problem in urban traffic networks. Transportation entities can use various methods to mitigate congestion, such as dedicated lanes, road pricing, and access restriction. Compared to these methods, traffic signal control is a very cost-effective solution to improve the efficiency of the traffic network. In past decades, traffic signal control has drawn a lot of research attention.

Traffic signal control techniques can be divided into two groups based on the control mechanism: pre-timed control and adaptive control. In pre-timed control, different signal parameters, including cycle length, phase splits, and phase sequences, are designed from historical data and implemented at predetermined periods of the day. This method works efficiently if the traffic conditions do not vary significantly from day to day. On the other hand, adaptive signal control provides optimal signal timing for the prevailing traffic condition. Adaptive signal timing changes according to traffic volume predictions to optimize specific measures of effectiveness such as travel delay, stop delay, and throughput rate. Adaptive control functions more flexibly in comparison to pre-timed control thanks to its ability to adjust to varying traffic conditions.

For a signalized intersection, traffic flow prediction and traffic signal optimization are two core components of the adaptive control method. Generally, current state measurements from sensors and approximation models predict traffic flow. After that, adaptive control method uses the optimization model and the traffic volume prediction to produce optimal signal timings. A favorable adaptive control method needs to possess two qualities: high prediction accuracy and fast computation speed. The first quality ensures the optimal signal timings match the real upcoming traffic volume; the second quality is necessary to change the signal timing plans in time.

1.2 Objectives

This study (1) develops a stochastic traffic delay model based on busy period analysis in queueing theory; (2) proposes a traffic optimization model based on the stochastic traffic delay model; (3) proposes a rolling horizon scheme to update signal timing on a phase-phase base; and (4) demonstrates the efficiency of the proposed model through case studies.

1.3 Expected Contributions

To accomplish these objectives, we have undertaken several tasks. First, based on busy period analysis, we propose stochastic traffic delay models for both the currently served phase (green phase) and the currently idle phase (red phase). We derive and then aggregate the traffic delays in all the subperiods within a signal cycle. Second, we derive a signal optimization model to minimize the average delay per vehicle in the current cycle. Third, the stochastic traffic delay model uses a rolling horizon scheme to adjust the signal timing in time to make full use of the available information, i.e., queue lengths. Fourth, we executed simulations to demonstrate the

effectiveness of the proposed framework. The simulations show that compared to the Webster's formula and another proposed adaptive signal control method, the proposed model can reduce traffic delay significantly.

1.4 Report Overview

The remainder of this report presents the following chapters: Chapter 2 presents a comprehensive review of the queueing theory based adaptive signal control methods and preliminary knowledge of M/D/1 queues. Chapter 3 provides a derivation of the stochastic traffic delay models based on the busy period analysis of M/D/1 queues. Chapter 4 proposes the signal optimization model and the rolling horizon scheme to implement the optimal signal timings on a phase-to-phase base. Chapter 5 demonstrates the performance of the proposed model by comparing its traffic delay vehicle to that resulting from the Webster's formula and another adaptive signal control method. Chapter 6 summarizes the work and discusses the future research directions.

Chapter 2. Literature Review

2.1 Introduction

This chapter provides a comprehensive review of the literature on various adaptive signal control methods.

2.2 Existing Queueing Theory Based Adaptive Signal Control Methods

Adaptive signal control methods aim to provide optimal signal timings for present traffic condition, therefore the importance of using an accurate arrival pattern cannot be overstated. In most cases, especially when the traffic volume is low, the arrival processes are random, making it challenging to model the measures of effectiveness (i.e., traffic delay). Commonly, modelers deal with the uncertainties in the arrival process by adding a stochastic term in the delay model [4]. However, such a static or time-invariant delay model may not be applicable when the volume-capacity ratio (v/c) is high since the traffic system becomes transient. Therefore, it is essential to analyze the traffic system using statistical methods based on the prevailing condition.

Queueing theory, widely used in traffic engineering, is a mathematical study of relationships of characteristics, such as arrival and departure processes, queue length and waiting time, in a system of queues. Traditionally, Kendall's [5] notation represents and classifies a queue. The notation involves three letters A/B/n that represent the distribution of interarrival time, service time, and the number of servers, respectively. The common letters for distributions include M (exponential distribution), D (deterministic) and G (general distribution). Many modelers have proposed traffic delay models based on queueing theory. Webster [6] proposed a well-known traffic model which is the sum of three terms: delay of a D/D/1 queue, delay of an M/D/1 queue and an adjustment factor which approximates 10 percent of the sum of the first two terms. However, Webster's model does not reflect of the impact of overflow when the v/c approaches 1 or the variance of the arrival distribution is large. To overcome this drawback, Miller [7] proposed two delay models that perform better when overflow occurs. Newell [8] developed a model considering the general distributions of arrival and departure processes. Akcelik's model [9] which is an approximation to Miller's model, has been used in the Australian Road Research Board's signalized intersection manual. The Highway Capacity Manual 2000 (HCM 2000) model [10] captures the delay from an existing queue at the beginning of the study period. A more detailed review of the existing delay models can be found in [11].

Based on queueing theory, Newell [12] conducted a comprehensive study depending on a steady-state analysis, however, Lo [13] pointed out that the peak hours could end before the intersection reaches steady state, and Newell's method may not be practical under this situation. To overcome this, he proposed a reliability framework called phase clearance reliability (PCR) to analyze the probability of phase failure for consecutive cycles given the arrival rate distribution. Based on this framework, Li et al. [14] developed an adaptive signal control model for coordinated arterial streets. Although the arrival rate is random in [13], [14], a constant interarrival time is employed, and this assumption is questionable under most situations.

2.3 Conditions for M/D/1 Queues

The distribution of time headways plays an important role on our analysis. The commonly used distributions include exponential distribution [6], [15]–[17], shifted exponential distribution [18]–[20], Gamma distribution [21], Lognormal distribution [13], [22] etc. In this discussion, we adopt the exponential distribution for the interarrival time, which has been extensively used by traffic flow researchers, especially those working on traffic signal control, since Adams [16]. The main advantage of this distribution is its memoryless property.

Definition 1 (Memoryless Property). A non-negative random variable X is said to have the memoryless property if

$$P(X > s+T|X > T) = P(X > t) \quad s,t \geq 0$$

In addition, the exponential interarrival time, with arrival rate of λ, leads to a Poisson arrival process, which indicates $N(t) \sim \text{Pois}(\lambda t)$, where $N(t)$ represents the number of arrivals in time t. A formal definition is as follows.

Definition 2 (Poisson Process). The counting process $N(t), t \geq 0$ generated by $X_n, n \geq 1$ is called a Poisson process with parameter (or rate) λ if $X_n, n \geq 1$ is a sequence of i.i.d $\exp(\lambda)$ random variables.

An important property of a Poisson process:

Theorem 1 (Stationary and Independent Increments). A Poisson process has stationary and independent increments.

Definition 3 (Stationary and Independent Increments). Let $X(t), t \geq 0$ be a continuous-time real-valued stochastic process. For given $s,t \geq 0$, $X(s+t)-X(s)$ is called the increment over the interval $(s, s+t]$. $X(t), t \geq 0$ is said to have stationary and independent increments if

(i) the distribution of the increment over an interval $(s, s+t]$ is independent of s,

(ii) the increments over non-overlappig intervals are independent.

These properties are very useful in computing statistics related to a Poisson process.

The main disadvantage of this exponential assumption is that the probability density of unrealistically short interarrival time is too large, and it increases with the arrival rate. Therefore, this assumption is valid only when the traffic volume is low. Wattleworth, in Baerwald [23], suggests that the traffic volume should be equal or less to 500 veh/h/lane. Luttinen [21] suggests a traffic volume of 100 veh/h/lane or less.

In addition to the arrival process, the model assumes that a queue's departure rate is constant if the downstream is uncongested. Equivalently, the model assumes that the time a vehicle needs to cross an intersection is constant. This model does not consider the start-up delay at the beginning of green time.

Then, we model an intersection as a system of M/D/1 queues. Each queue is formed by the vehicles from the same approach and each queue is served by the same phase. Note that the departure rates of these queues are not constant since it switches between saturation flow and zero along with signal switching between green and red. For the reason of simplicity, we call these queues M/D/1 queues. In queueing theory, a *polling system* consists of multiple queues accessed in cyclic order by a single server. Previous papers have proposed a number of analytical models [1]–[3] to study properties of polling systems. Three commonly used service policies for a polling system include: *exhaustive service policy* where a queue is served until it is emptied, *gated service policy* where only the customers in the queue at the instant that the server arrived are served, but new arrivals during this service time must wait until the next server arrives, and *limited service policy* where a fixed maximum number of customers can be served in each visit. Under most cases, the exhaustive service policy preforms best in reducing the waiting time or delay.

Unlike the exhaustive service policy which makes decision on server's movement based on the existence of a queue in the currently served channel, an adaptive signal control method usually takes traffic states of all phases into account and provides an optimal signal timing for a period in advance, such as green splits for the next cycle. This study proposes a stochastic traffic delay model for a signalized intersection and an adaptive signal control model using a rolling horizon scheme.

2.4 Summary

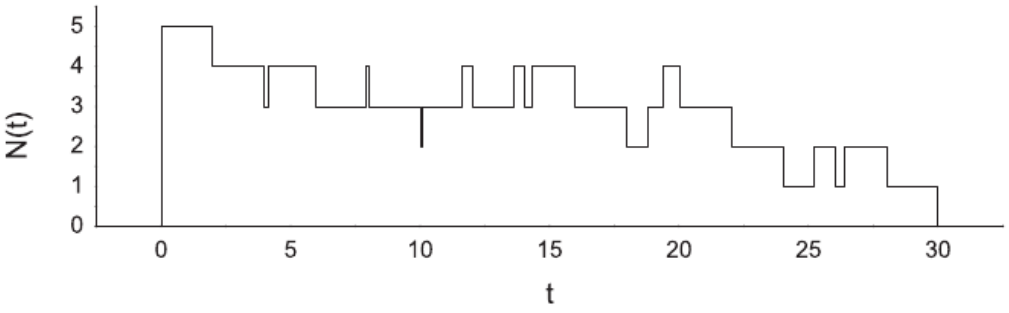
A comprehensive review of past research of on the queueing theory based adaptive signal control problem has been discussed in the preceding section. These models make assumptions on the traffic arrival processes such as uniform arrivals, arrivals in platoons. However, these assumptions deviate from the reality especially when the traffic volume is low where we expect the adaptive signal control techniques to reduce more traffic delay than when the traffic volume is high. According to previous studies, the exponential distribution is an accurate assumption for time headways under low traffic volume conditions. Therefore, to produce a more accurate traffic delay model and a more sufficient signal control algorithm, we model the queues as M/D/1 queues and propose a corresponding adaptive signal control method in this report.

Chapter 3. Stochastic Traffic Delay Models

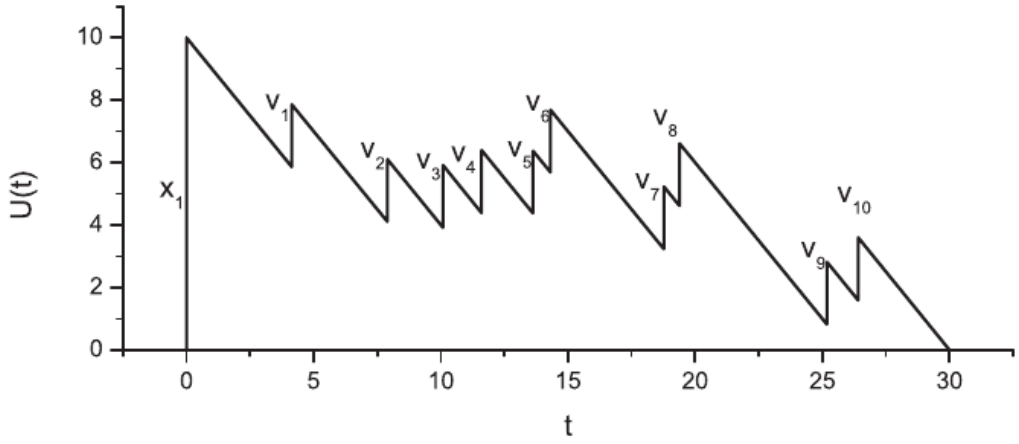
3.1 Busy Period Analysis

3.1.1. The Probability Function of a Busy Period

Borel [24] introduced the concept of a busy period as the time duration between the instant the signal begins serving a queue until the instant the queue is cleared, and this time duration is closely relates to the traffic delay model in the Section 3.2. Figure 3-1 shows an example of a busy period of an M/D/1 queue. The arrivals are served in a First-In-First-Out (FIFO) order. In this example, the arrival rate $\lambda = 0.35$, the service rate $\mu = 0.5$, there are 5 vehicles at the beginning of the analysis period, and the busy period ends at $t = 30$ s when the queue is cleared. Figure 3-1(a) represents the number of vehicles in the queue, and the solid line in Figure 3-1(b) shows the corresponding remaining time to empty the queue. The jumps in Figure 3-1(a) denote arrivals at random times while the drops indicate departures which occur every 2 seconds since the departure rate is 0.5 vehicles per second. Similarly, the jumps in Figure 3-1(b) denote the arrivals and the jump step size is 2 seconds, which means the remaining time to empty the queue increases by 2 seconds when a new vehicle joins the queue. The slope of the oblique sections is -1. In Figure 3-1(b), $t = x_1$ is the time required to clear the initial queue, and ten new arrivals are served during this busy period.



(a) Number of vehicles in a queue



(b) Remaining time to clear a queue

Figure 3-1 The busy period.

Let λ and μ denote the arrival rate and departure rate, respectively; N denote the number of vehicles at the instant the signal starts serving the queue and n represent the number of vehicles arrived and served during the busy period. The event that a busy period serves N+n vehicles needs to satisfy the following conditions:

(i) there are exact n vehicles arrived during time $\frac{N+n}{\mu}$;

(ii) the arrival pattern should be 'admissible'. The term 'admissible' indicates that at least one vehicle must arrive during time N/μ, at least two vehicles must arrive during time $\frac{N+1}{\mu}$, and so on, so that the busy period will always be occupied for time $\frac{N+n}{\mu}$.

Borel [25] developed the probability of the length of a busy period of a M/D/1 queue. Tanner gave another derivation and generalized [26] that distribution, which is called Borel-Tanner distribution [27].

$$P_n = \frac{1}{n!}\left((N+n)\,\rho\right)^n \exp(-\rho(N+n))\frac{N}{N+n}, \tag{1}$$

where P_n is the probability of that the busy period is $\frac{N+n}{\mu}$ given the initial queue length equal to N, $\rho = \frac{\lambda}{\mu}$ is the server utilization or degree of saturation. Note that all the statistics shown in this chapter are conditional on the initial queue length N. We omit it from the expressions.

Let T denote the length of a busy period, its expectation and variance are

$$\mathrm{E}[T] = \frac{N/\mu}{1-\rho}, \tag{2}$$

$$\mathrm{Var}[T] = \frac{1}{\mu^2}\frac{\rho N}{(1-\rho)^3} \tag{3}$$

3.1.2. **Average Delay per Vehicle in a Busy Period**

To derive the traffic delay model, we need to obtain the average delay per vehicle d_n in a busy period given $T = \frac{N+n}{\mu}$. The numerical values of d_n's can be achieved by utilizing the Campbell's theorem.

Theorem 2 (Campbell's Theorem). Let a_i be the ith event time in a $PP(\lambda)\{N(t), t \geq 0\}$. Given $N(t) = n$,

$$(a_1, a_2, ..., a_n) \sim (\tilde{U}_1, \tilde{U}_2, ..., \tilde{U}_n)$$

where $(U_1, U_2, ..., U_n)$ are n i.i.d random variables that are uniformly distributed over [0,t], and $(\tilde{U}_1, \tilde{U}_2, ..., \tilde{U}_n)$ such that $\tilde{U}_1 \leq \tilde{U}_2 \leq ... \leq \tilde{U}_n$ is the order statistics of $(U_1, U_2, ..., U_n)$.

Then, d_n's can be obtained through the following steps:

(i) draw n random samples $(u_1, u_2, ..., u_n)$ from the uniform distribution unif $(0, (N+n)/\mu)$ and arrange them in an ascending order such that the joint sample $(\tilde{u}_1, \tilde{u}_2, ..., \tilde{u}_n)$. Repeat this step M times, where $M = 10^6$ in this dissertation;

(ii) remove the joint samples that do not satisfy the 'admissible' condition, and let m denote the number of remaining joint samples;

(iii) for the ith joint sample, let d_j^i denote the delay of vehicle $j, j \in [1,n]$, and $d_1^i = N/\mu - u_1^i, d_2^i = (N+1)/\mu - u_2^i, ..., d_n^i = (N+n-1)/\mu - u_n^i$. Let $d^i = \frac{\sum_{j=1}^{n} d_j^i}{n}$ be the average delay of the ith sample;

(iv) set d_n equal to the average of d^i's, i.e. $d_n = \frac{\sum_{i=1}^{m} d^i}{m}$.

An interesting finding is that given n, the average delay is dependent on μ and not dependent on λ.

3.2 Traffic Delay Analysis

At the instant when signal starts serving a phase, we observe the queue lengths of all phases and propose a stochastic traffic delay model conditional on these queue lengths. Figure 3-2 shows an example of green splits of a four-phase `cycle'. Without loss of generality, we assume $t_0$ is the time the signal starts to serve phase 1. In the following `cycle' $t_0 + c$, every phase i consists of three parts: leading red time r_{il}, green time g_i and following red time r_{if}. Note that the first phase does not have a leading red time, but the last phase has a following red time due to the all red time.

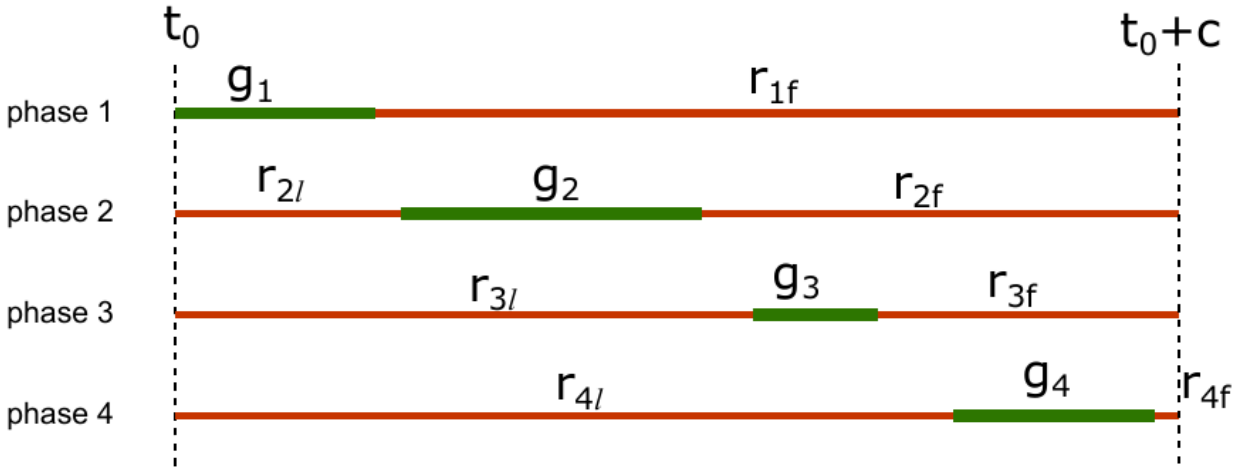


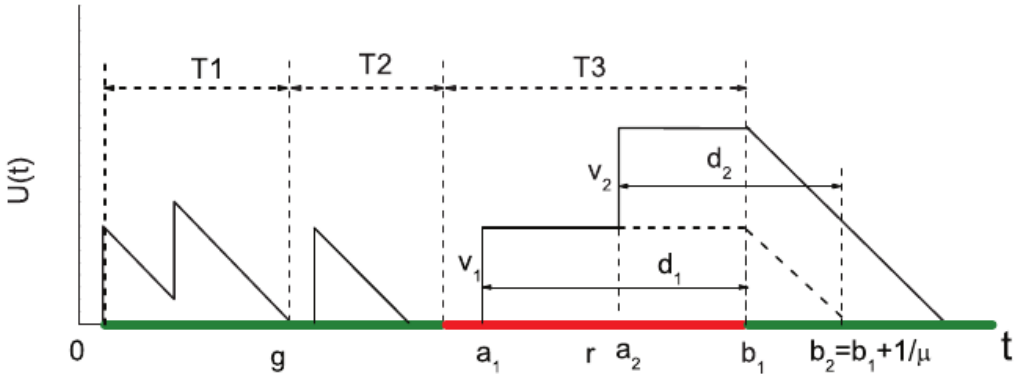
Figure 3-2 Schematic of a horizon

In the proposed model, cycle length is allowed to change during the simulation time. Therefore, the definition of a `cycle' is not appropriate. Instead, since the adaptive signal control model utilizes the rolling horizon scheme, the term of a 'project horizon' is adopted in the following to represent the time interval between the instant when the signal starts to serve a certain phase and the instant when the signal finishes serving all the phases and visits that phase again. Before we derive the traffic delay model, we classify the phases into two groups: currently

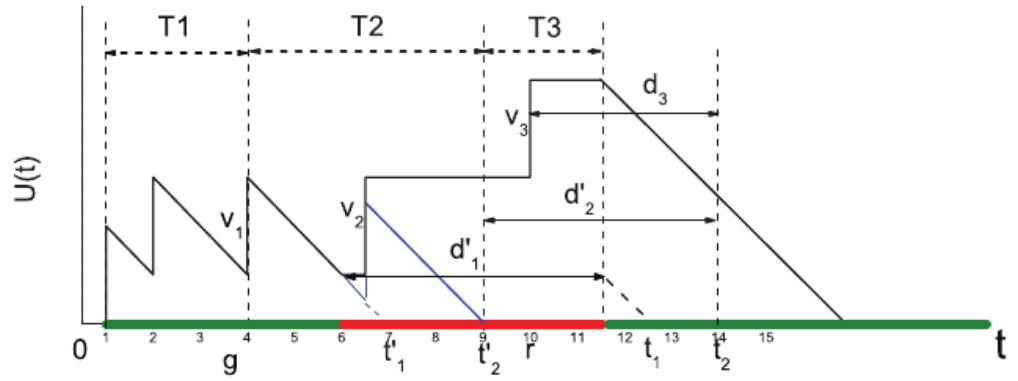
served phase and currently idle phase. For example, phase 1 in Figure 3-2 is the currently served phase and all others are currently idle phase. N_i and n_i denote the queue length of movement i at time t_0 and the number of arrivals in the busy period initiated by the queue at the instant when signal starts to serve movement i, respectively.

3.2.1 **Traffic Delay Model for Currently Served Phase**

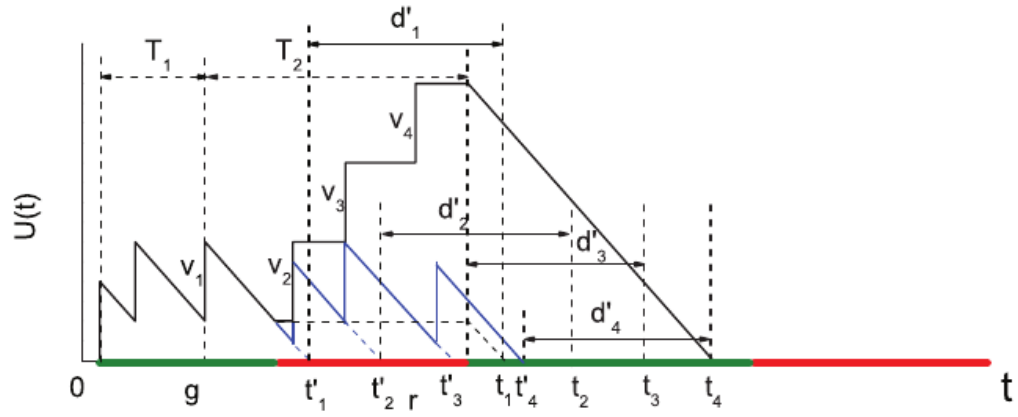
For a currently served phase, there are three cases based on the relationship between the busy period and the phase splits, which are shown in Figure 3-3. Let c denote the project horizon. For simplicity, the subscripts are omitted. In addition, we assume the green time is sufficient to proceed the initial queue.



(a) $g \geq \frac{N+n}{\mu}$



(b) $g \leq \frac{N+n}{\mu} \leq g + r$



(c) $g + r \leq \frac{N+n}{\mu}$

Figure 3-3 Delay per vehicle analysis scenarios

Case (a): $g \geq \frac{N+n}{\mu}$

Figure 3-3(a) shows the remaining time to clear the current queue. For simplicity, by busy period, we mean the first busy period after the signal starts to serve the corresponding movement. In this case, the busy period $T1 = \frac{N+n}{\mu}$ is shorter than green time g; after the busy period ends, more vehicles will arrive and be served during the rest of green time $T2 = g - T1$; then vehicles arrived during the red time T3 will depart in next green time. Let d_{a1}, d_{a2} and d_{a3} denote the average delay per vehicle in T1, T2 and T3, respectively, we can compute the expectation of these delays conditional on n. First of all, $E[d_{a1}|n]$ is obtained from the numerical value in section 3.1.2.

As mentioned before, the exponential interarrival time is valid when the arrival rate λ is small, so we assume that all the vehicles arrived during T2 can be served before the end of green time. In addition, based on queueing theory, the expected queue length of an M/D/1 queue in a long run is equal to $\frac{\rho^2}{2(1-\rho)\mu^2}$, which is less than 0.5 when $\rho \leq 0.5$. Therefore, for simplicity, we assume there no vehicle left at the end of green time. $E[d_{a2}|n]$ is approximated as the average delay of an M/D/1 queue in the long run. Based on queueing theory,

$$\mathrm{E}[d_{a2}|n] = \frac{\lambda}{2(1-\rho)\mu^2} \tag{4}$$

To calculate the average delay in T3, we need to first condition $E[d_{a3}|n]$ on the number of arrival, n', during the red time and the arrival time of these n' vehicles: $a_1, a_2, ..., a_{n'}$. It is clearly shown in Figure 3-3(a) that the departure times of these vehicles, shown as b_i's, are independent of the arrival times and equal $g + r$, $g + r + \frac{1}{\mu}$, …, $g + r + \frac{n'-1}{\mu}$. Therefore,

$$\mathrm{E}[d_{a3}|n, n', a_1, a_2 ... a_{n'}] = \frac{\sum_{i=1}^{n'} \left(g + r + \frac{i-1}{\mu} - a_i\right)}{n'} = \frac{n'\left(g + r + \frac{n'-1}{2\mu}\right) - \sum_{i=1}^{n'} a_i}{n'}. \tag{5}$$

Since arrivals during the red time is a Poisson process, based on Theorem 2, $(a_1, a_2, ..., a_{n'})$ is an order statistic of the uniform distribution $\mathrm{unif}(0, r)$. By un-conditioning the arrival times, we obtain

$$\mathrm{E}[d_{a3}|n, n'] = \frac{n'\left(g + r + \frac{n'-1}{2\mu}\right) - n'\left(g + \frac{r}{2}\right)}{n'} = \frac{r}{2} + \frac{n'-1}{2\mu}. \tag{6}$$

It is known that n' is a Poisson distribution with parameter of λr. Based on the law of total expectations,

$$\mathrm{E}[d_{a3}|n] = \frac{r}{2} + \frac{\lambda r - 1}{2\mu}. \tag{7}$$

Let n_1, n_2 and n_3 be the expectations of arrivals during T1, T2 and T3, respectively, then

$$n_1 = n, \quad n_2 = \lambda\left(g - \frac{N+n}{\mu}\right), \quad n_3 = \lambda r. \tag{8}$$

In addition to the delay in these three periods, we also need to consider the impact of the queue at $t = t_0 + c$ on the delay in the next cycle. We assume that it only influences the departure time of the vehicles arrived in the busy period. From Equation (2), we know that the expected number of affected vehicles is

$$\mathrm{E}[d_{a4}|n] = \frac{n_3}{2\mu}. \tag{9}$$

Then, the total delay of the currently served movement in this project horizon is approximated as

$$\mathrm{E}[d_a|n] = \sum_{i=1}^{4} \mathrm{E}[d_{ai}|n]n_i. \tag{10}$$

Case (b): $g \le \frac{N+n}{\mu} \le g + r$

Before we show the delay models, it is necessary to explain how we define n in this case. As shown in Figure 3-3(b), the solid blue line represents the evolution of the unfinished work pretending the signal is always green. At time $t_2^{'}$, the busy period ends since the unfinished work reaches zero. n is defined as the number of arrivals served during this hypothetically busy period. For example, n=3 in Figure 3-3(b). We divide this project horizon into 3 parts: T1 is the period in which the vehicle arrived and can be served during the current green time, T2 denotes the rest of the hypothetically busy period and T3 represents the rest of the project horizon. Different from case (a), let d_{b1} denote the average delay in the busy period assuming the signal is always green, d_{b2} represent the extra delay occurred by the vehicles arrived in T2 and d_{b3} indicate the average delay in T3. Like case (a), $E[d_{b1}|n]$ is obtained from the numerical value in section 3.1.2.

To calculate $E[d_{b2}|n]$, let $t_i^{'}$ denote the hypothetical departure time of the vehicle i that arrived during T2, and t_i indicate its real departure time. It is evident in Figure 3-3(b) that compared to the all green time case, the real departure time of all vehicles is postponed by r, i.e. $t_1 - t_1^{'} = t_2 - t_2^{'} = ... = r$. Therefore,

$$\mathrm{E}[d_{b2}|n] = r. \tag{11}$$

Like case (a), we need to condition $E[d_{b3}\,|\,n]$ on the number of arrivals during T3 and their arrival time. After the signal turns green, these vehicles have to wait until all the vehicles arrived in T2 depart, which takes time of $\frac{N+n}{\mu}-g$. Therefore,

$$\begin{aligned}\mathrm{E}[d_{b3}|n,n^{'},a_1,a_2,...,a_{n'}] &= \frac{\sum_{i=1}^{n^{'}}\left(g+r+\frac{N+n}{\mu}-g\frac{i-1}{\mu}-a_i\right)}{n^{'}} \\ &= \frac{n^{'}\left(r+\frac{N+n}{\mu}+\frac{n^{'}-1}{2\mu}\right)-\sum_{i=1}^{n^{'}}a_i}{n^{'}}.\end{aligned} \tag{12}$$

By conditioning the arrival times and n', we obtain

$$\begin{aligned}\mathrm{E}[d_{b3}|n,n^{'}] &= \frac{n^{'}\left(r+\frac{N+n}{\mu}+\frac{n^{'}-1}{2\mu}\right)-\sum_{i=1}^{n^{'}}a_i}{n^{'}} \\ &= \frac{r+\frac{N+n}{\mu}-g}{2}+\frac{n^{'}-1}{2\mu}.\end{aligned} \tag{13}$$

and

$$\mathrm{E}[d_{b3}|n] = \frac{r+\frac{N+n}{\mu}-g}{2}+\frac{\lambda\left(g+r-\frac{N+n}{\mu}\right)-1}{2\mu}. \tag{14}$$

Let n_1, n_2 and n_3 be the expectations of arrivals during the hypothetical busy period T1+T2, T2 and T3, respectively, then,

$$n_1 = n, \quad n_2 = n-g\mu+N, \quad n_3 = \lambda(g+r-\frac{N+n}{\mu}). \tag{15}$$

Similar to case (a), let n_4 denote the number of vehicles arrived during next busy period, and $E[d_{b4}\,|\,n]$ indicate their delay resulted from the queue at the end of this project horizon,

$$n_4 = (n_2+n_3)\frac{\rho}{1-\rho}. \tag{16}$$

$$\mathrm{E}[d_{b4}|n] = \frac{n_2+n_3}{2\mu}. \tag{17}$$

The expression of the total delay in this project horizon is approximated as

$$\mathrm{E}[d_b|n] = \sum_{i=1}^{4}\mathrm{E}[d_{bi}|n]n_i. \tag{18}$$

Case (c): $g+r \le \frac{N+n}{\mu}$

Period T3 is disappeared in this case. Contrary to cases (a) and (b), in which the busy periods are shorter than the project horizon, the busy period exceeds the project horizon in this case. Since we only assess the delay of vehicles arrived during the current project horizon, we approximate this number as

$$n_t = n\frac{g+r}{\frac{N+n}{\mu}}. \tag{19}$$

where $\frac{g+r}{\frac{N+n}{\mu}}$ is the ratio of the project horizon to the busy period. By employing Equation (19), we assume the number of arrivals is proportional to the time window. Like the two cases above, $E[d_{c1}|n]$ is obtained from the numerical values in section 3.1.2, and

$$\mathrm{E}[d_{c2}|n] = r, \quad \mathrm{E}[d_{c3}|n] = 0, \quad \mathrm{E}[d_{c4}|n] = \frac{n_t - (g\mu - N)}{2\mu}. \tag{20}$$

The corresponding numbers of arrivals are

$$n_1 = n_t, \quad n_2 = n_t - (g\mu - N), \quad n_3 = 0, \quad n_4 = (n_t - (g\mu - N))\frac{\rho}{1-\rho}. \tag{21}$$

The total delay is estimated as,

$$\mathrm{E}[d_c|n] = \sum_{i=1}^{4} \mathrm{E}[d_{ci}|n]n_i. \tag{21}$$

Overall, the total delay for a currently served phase can be expressed as

$$\mathrm{E}[d|n] = \delta_a \mathrm{E}[d_a|n] + \delta_b \mathrm{E}[d_b|n] + \delta_c \mathrm{E}[d_c|n]. \tag{22}$$

where δ_a, δ_b and δ_c are binary variables satisfying

$$\begin{cases} \delta_a + \delta_b + \delta_c = 1, \\ \delta_a = 1, & \text{if } g \geq \dfrac{N+n}{\mu} \\ \delta_b = 1, & \text{if } g \leq \dfrac{N+n}{\mu} \leq g + r \\ \delta_c = 1, & \text{if } g + r \leq \dfrac{N+n}{\mu} \end{cases} \tag{23}$$

3.2.2 **Traffic Delay Model for Currently Idle Phase**

As shown in Figure 3-2, an idle phase contains a leading red time. Let us call this period T0. Compared to the delay models for a currently served phase, we need to add the extra delay in this period for an idle phase. Let N denote the queue length at $t = t_0$, the expected number of vehicles arrived during T0 and the corresponding average delay are

$$n_0 = \lambda r_l \tag{24}$$

$$\mathrm{E}[d_0|n] = \frac{N}{\mu} + \frac{\lambda r_l - 1}{2} \tag{25}$$

For the remaining time of the project horizon, we can use the models developed in section 3.1.3 to calculate the delay by replacing N by $N + n_0$. Therefore, the total delay for an idle phase can be expressed as,

$$\mathrm{E}[d|n] = \delta_a \mathrm{E}[d_a|n] + \delta_b \mathrm{E}[d_b|n] + \delta_c \mathrm{E}[d_c|n] + n_0 \mathrm{E}[d_0|n] + N\left(r_l + \frac{N/\mu}{2}\right) \tag{26}$$

in which the last term is the total delay of the initial N vehicles. The reason why we do not include this term for the currently served phase is that we assume the initial queue will be cleared in the current project horizon, and their delay will be a constant under this assumption.

Chapter 4. Adaptive Signal Control Model

Based on the stochastic traffic delay models, an adaptive signal control model using rolling horizon scheme is proposed in this section. At the beginning of each phase, the queue lengths of all movements at an intersection are obtained through deployable technologies such as loop detectors and video image processors. Then, optimal phase splits for the current project horizon is provided. However, only the green time for the first phase is implemented. This procedure is repeated at the beginning of each phase until the end of the study. Although only one phase is applied in each iteration, the optimization model is processed for the whole project horizon. This can help avoid a shortsighted control. Additionally, by using the rolling horizon scheme, the newest information, i.e. queue lengths at the intersection, is exploited and this ensures that the optimal solutions are adaptive to the most current traffic condition.

Although the arrival times are random, the average arrival rates are assumed to be known. The optimization model can be expressed as

$$
\begin{aligned}
\min_{\boldsymbol{g},c} \quad & \frac{\sum_{i=1}^{n_m} \mathrm{E}[d_i]}{\sum_{i=1}^{n_m} \lambda_i c + \sum_{i \notin I_1} N_i} \\
\text{s.t.} \quad & \sum_{i=1}^{n_m} g_i + n_p l = c, \\
& r_{il} = \sum_{j=1}^{j=i-1} g_j + (i-1)l, \quad \forall i \in [2, n_p] \\
& r_{if} = c - r_{il} - g_i, \qquad \forall i \in [1, n_p - 1] \\
& g_i \geq \frac{N_i + \lambda_i r_{il}}{\mu_i}, \qquad \forall i \in [1, n_p] \\
& c \leq c_{\max}
\end{aligned}
\tag{27}
$$

where g_i is the green time for phase i, c is the project horizon, n_p is the number of phases, n_m is the number of movements, I_i is the set of movements served by phase i, l is the lost time, which is equal to the sum of the all-red time, startup delay and deceleration delay, λ_i and μ_i are the average arrival rate and saturation flow of phase i, respectively, r_{il} and r_{if} are the leading and following red time of phase i, $c_{\max}$ is the predetermined maximum project horizon length.

The green time and project horizon are decision variables. The objective function is to minimize the average delay per vehicle in the current project horizon. The first constraint indicates that the project horizon is equal to the sum of green time and lost time of all phases; the second constraint represents the leading red of a phase is equal to the sum of green time and lost time of its preceding phases; the third constraint means the following red time of phase i is equal to the project horizon subtracted by its leading red and green time; the fourth time forces the

leading time of each phase to be longer than the required time to clear the queue formed at the beginning of green time; the fifth constraint sets an upper bound for the project horizon.

By the law of total expectation, $E[d_i]$ can be expressed as

$$\mathrm{E}[d_i] = \sum_{j=1}^{n_{\max}} P_j \mathrm{E}[d_i|j] \tag{28}$$

where j is the number of vehicles arrived during a busy period, P_j is the Borel-Tanner distribution Equation (1), and $n_{\max}$ is the truncation point. Let $n_{\max} = 100$ in this chapter.

Chapter 5. Case Study

The performance of the proposed model is tested using an isolated intersection with two one-way streets. Let the maximal cycle time be 80 s. The simulation time is 3 h. Under scenarios with different degrees of saturation ρ's, we compare the traffic delay from our model to Webster's model [6] and the exhaustive policy [3] of a polling system. Since Webster's model is a fixed signal design model which calculates the phase splits and cycle length based on ρ's, one may argue it is unfair to compare the efficiency of an adaptive signal control model to a fixed control method. Therefore, another adaptive control logic is employed in this section: at the beginning of each phase, the queue length of the currently served phase is observed and the expectation of a busy period from Equation (2) is offered for its green time. The reason why this method is chosen is that based on our results, the exhaustive policy, which moves the signal to next movement until the current queue is cleared, performs the best. In this policy, the signal's stay time is, in fact, the busy period. Therefore, we regard setting the green time equal to the expectation of the busy period as a good and fair method to be compared. We call this method `Adaptive BP'. Figure 5-1-Figure 5-3 show the average delay per vehicle from Webster's formula, `Adaptive BP', the proposed method and the exhaustive service policy, under various scenarios.

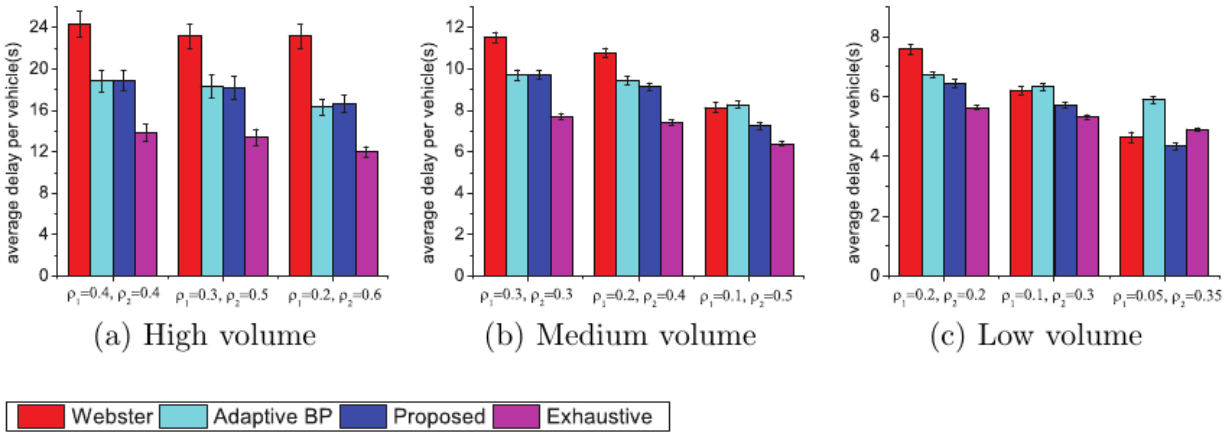


Figure 5-1 Performance comparison: $\mu_1 = \mu_2 = 1$ **veh/s**

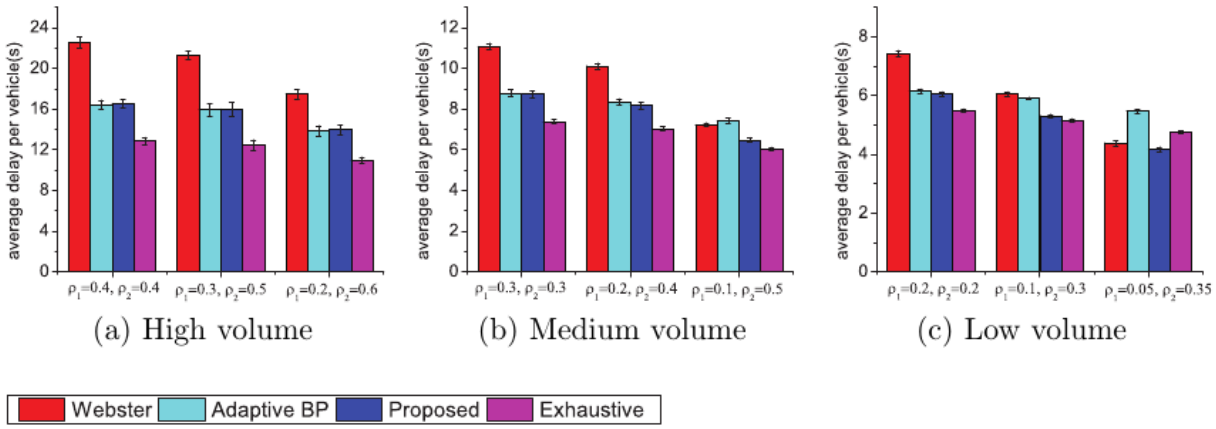


Figure 5-2 Performance comparison: $\mu_1 = \mu_2 = 2$ **veh/s**

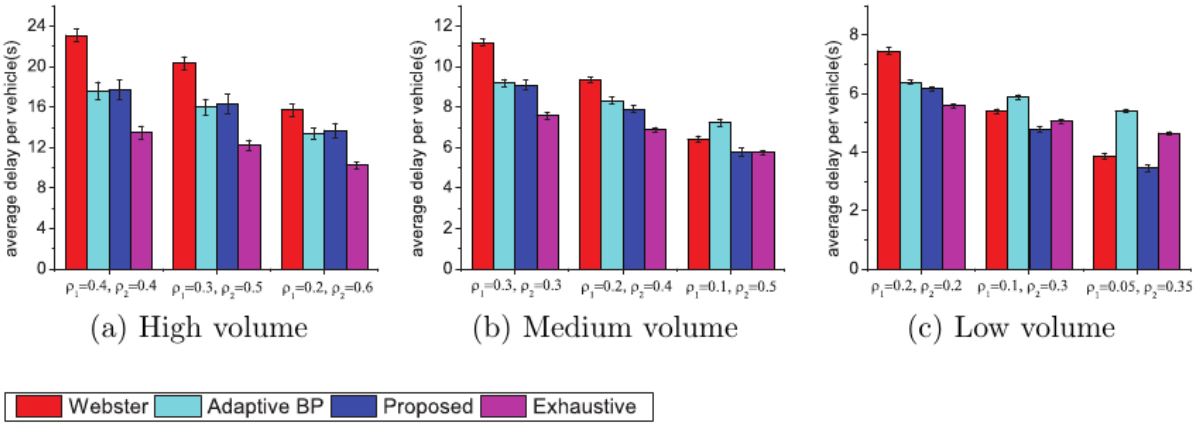


Figure 5-3 Performance comparison: $\mu_1 = 1$ veh/s, $\mu_2 = 2$ veh/s **veh/s**

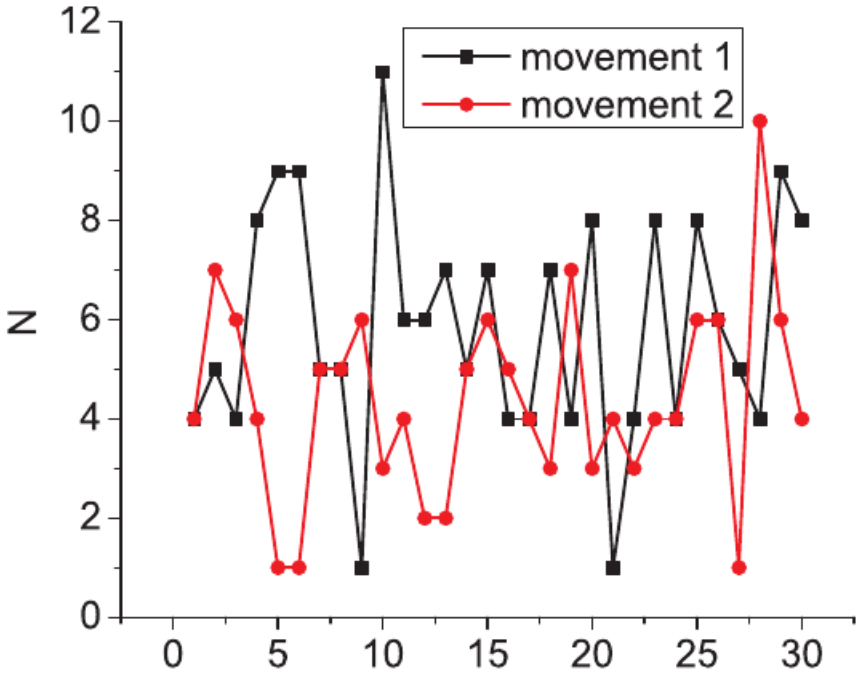
Each figure has different saturation flows. Figure 5-1 and Figure 5-2 represent the symmetric cases, in which the saturation flows of two directions are equal, while Figure 5-3 is the asymmetric case. In each case, we test three levels of overall v/c ratios: 0.8, 0.6 and 0.4. Furthermore, for each overall v/c, we consider the balanced case and unbalanced cases. For each scenario, 20 runs with different random seeds are executed. They show that under most cases, the exhaustive service policy has the best performance while Webster's model has the worst performance. The efficiency of our proposed model and the `Adaptive BP' method fall in between them. The traffic delay from the proposed model is always equal to or less than Webster's method and the `Adaptive BP' method. Although the exhaustive service policy is superior to the proposed model, it cannot be implemented as an adaptive signal control since it is not able to provide the optimal signal timings for a time period in advance. In addition, for a coordinated traffic system, it is difficult to utilize this logic since this method does not encode the dependence between consecutive intersections. Compared to Webster's formula, our model can reduce the traffic delay significantly under all scenarios. Compared to the `Adaptive BP' method, the reduction is not obvious under a high volume. With the decrease of traffic volume, the proposed model is capable of decreasing traffic delays notably, especially for the unbalanced cases. When $\rho_1 = 0.05$ and $\rho_2 = 0.35$, our method dominates all other three models. By using

the `Adaptive BP' method as a baseline, Table 5-1 shows the total delay reduction in one hour profited from the proposed model.

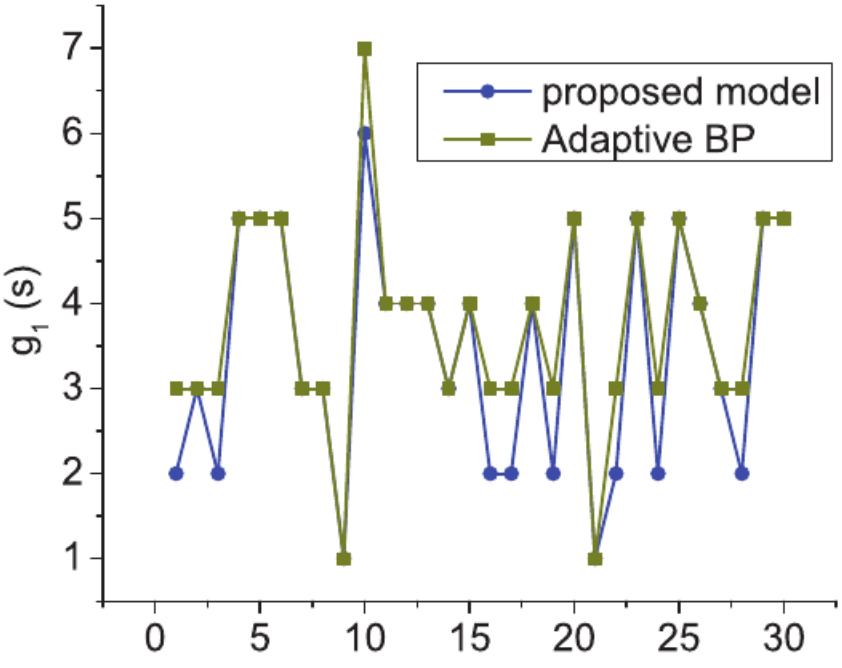
Table 5-1 Total delay reduction compared to the 'Adaptive BP' method (h).

	medium			low		
	$\rho_1 = 0.3$	$\rho_1 = 0.2$	$\rho_1 = 0.1$	$\rho_1 = 0.2$	$\rho_1 = 0.1$	$\rho_1 = 0.05$
	$\rho_2 = 0.3$	$\rho_2 = 0.4$	$\rho_2 = 0.5$	$\rho_2 = 0.2$	$\rho_2 = 0.3$	$\rho_2 = 0.35$
$\mu_1 = \mu_2 = 1$	0.01	0.17	**0.61**	0.11	**0.25**	**0.62**
$\mu_1 = \mu_2 = 2$	0.07	**0.23**	**1.13**	0.09	**0.47**	**1.04**
$\mu_1 = 1, \mu_2 = 2$	0.09	**0.44**	**1.59**	0.13	**0.77**	**1.47**

For the balanced case, the reduction is less than 0.13 h. For the unbalanced case, it can reach up to 1.59 h. This is because our model considers the queues, arrivals and saturation flows of all movements when it optimizes the signal timings while the `Adaptive BP' only takes the currently served movement into consideration. Figure 5-4 - Figure 5-5 show the queue lengths and the green times from both models under the case of $\mu_1 = \mu_2 = 2$, and $\rho_1 = 0.1$, $\rho_2 = 0.5$. While the pattern of the green time from the `Adaptive BP' method is the same as the corresponding queue length, our model takes both queues to make a decision. Figure 5-4 shows that our method allocates a shorter green time to the movement with a lower arrival rate than the `Adaptive BP' model. On the contrary, it is shown in Figure 5-5 that the proposed model is inclined to offer more green time to the busy direction. Since the proposed models insert the influence of residual queue on the next cycle, a queue in the direction with heavy traffic can block more vehicles and lead to higher delay than the other direction. In addition, the model Equation (3) tells that variation of a busy period increases with the arrival rate. Since the objective is to minimize the delay per vehicle, the proposed model provides longer green time for the heavy traffic direction to reduce the probability of a residual queue and further lower the expectation of overall traffic delay. In other words, in this case, the benefits from increasing green time for a heavy direction is more significant than the delay increase caused by cutting green time for a light direction.

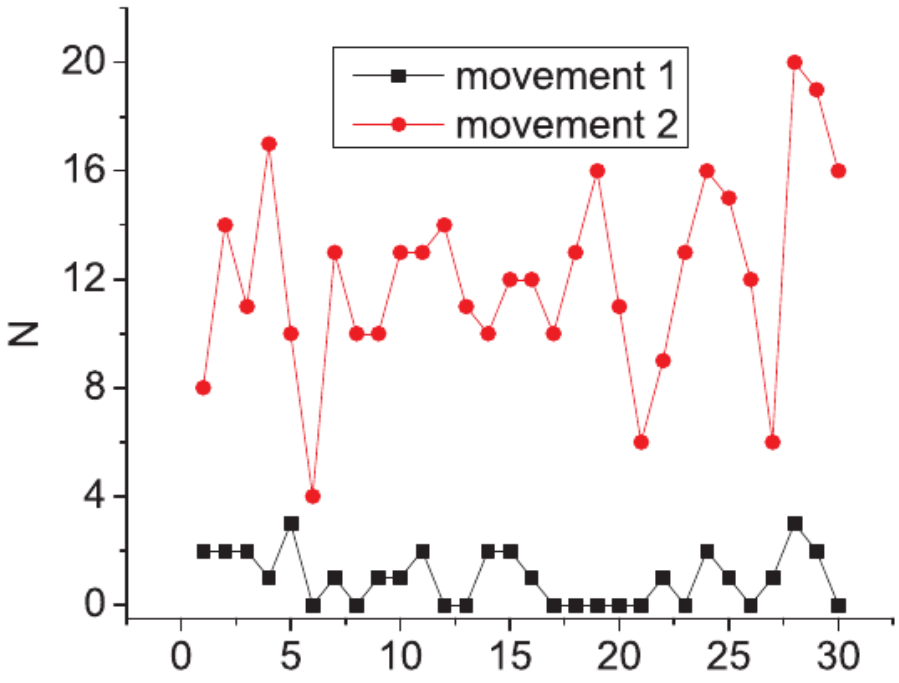


(a) Queue lengths when signal moves to movement 1

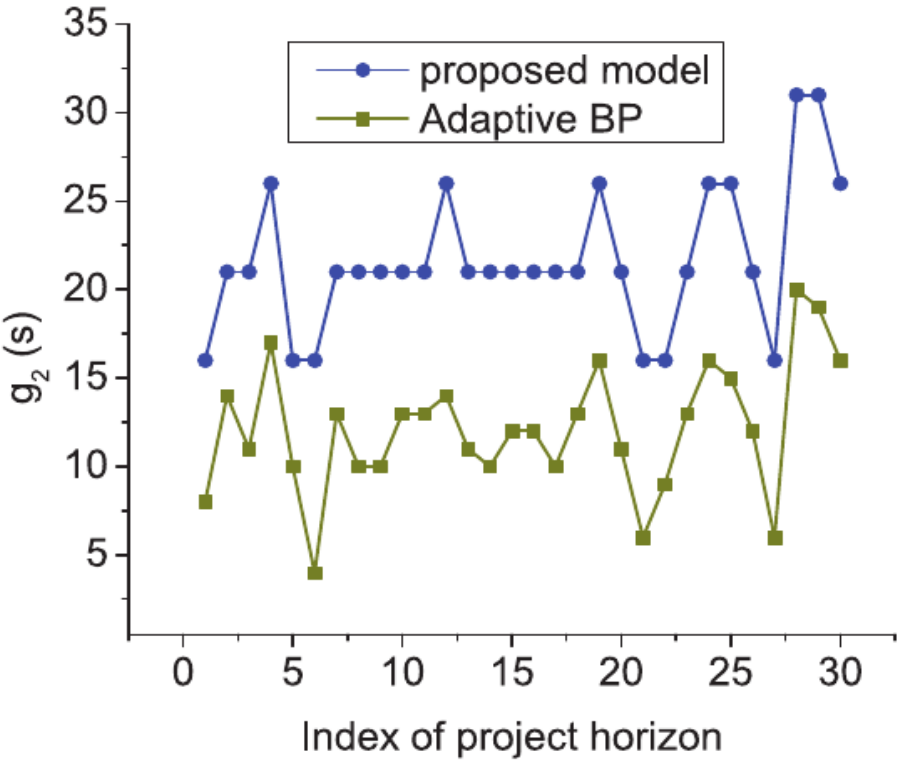


(b) Green time for movement 1

Figure 5-4 Green time for movement 1



(a) Queue lengths when signal moves to movement 2



(b) Green time for movement 2

Figure 5-5 Green time for movement 2.

Chapter 6. Conclusion

Based on queueing theory, we model a signalized intersection as a system of M/D/1 queues. Built upon the busy period analysis, this report derives the stochastic traffic delay models for both currently served and idle movements. Then, this report explains the adaptive signal control model which minimizes the delay per vehicle. The proposed model uses the rolling horizon scheme to maximize the utilization of available information, i.e. queue lengths, to improve the performance. Compared to Webster's formula and another adaptive signal control method inspired by the exhaustive service policy, our model can reduce delay significantly, especially for low and unbalanced traffic volume condition.

Although the proposed model can reduce traffic delay significantly, the signal optimization model Equation (27) is not convex, and the computation speed is low. In the future, researchers will need to find a way to improve the computational speed of the proposed model to make it ready for practical use.

References

[1] L. Kleinrock and H. Levy, "The analysis of random polling systems," *Oper. Res.*, vol. 36, no. 5, pp. 716–732, 1988.

[2] C. . F. . Daganzo, "Some properties of polling systems," vol. 6, pp. 137–154, 1990.

[3] H. Takagi, "Queuing analysis of polling models," *ACM Comput. Surv.*, vol. 20, no. 1, pp. 5–28, 1988.

[4] B. G. Heydecker, "TREATMENT OF RANDOM VARIABILITY IN TRAFFIC MODELLING.," *Work. TRAFFIC Granul. FLOW HLRZ, FORSCHUNGSZENTRUM, JULICH, Ger. Oct. 9-11, 1995*, 1996.

[5] D. G. Kendall, "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain," *Ann. Math. Stat.*, vol. 24, no. 3, pp. 338–354, Sep. 1953.

[6] F. Webster, "Traffic signal settings, road research technical paper," *Road Res. Lab.*, 1958.

[7] A. J. Miller, "Australian road capacity guide: provisional introduction and signalized intersections," no. 4, Jun. 1968.

[8] G. F. Newell, "STATISTICAL ANALYSIS OF THE FLOW OF HIGHWAY TRAFFIC THROUGH A SIGNALIZED INTERSECTION*," 1956.

[9] R. Akcelik, "The highway capacity manual delay formula for signalized intersections," *ITE J.*, vol. 58, no. 3, pp. 23–27, 1988.

[10] "Highway Capacity Manual," in *Transportation Research Board*, 2000.

[11] A. P. Akgungor and A. Bullen, "ANALYTICAL DELAY MODELS FOR SIGNALIZED INTERSECTIONS," 1999.

[12] G. F. Newell, "Theory of Highway Traffic Signals," 1989.

[13] H. K. Lo, "A reliability framework for traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 250–260, 2006.

[14] L. Li, W. Huang, and H. K. Lo, "Adaptive coordinated traffic control for stochastic demand," *Transp. Res. Part C Emerg. Technol.*, vol. 88, pp. 31–51, Mar. 2018.

[15] P. B. Mirchandani and N. Zou, "Queuing Models for Analysis of Traffic Adaptive Signal Control," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 50–59, Mar. 2007.

[16] W. F. Adams, "Road traffic considered as a random series," *Oper. Res. Q.*, vol. 1, no. 1, p. 9, 1950.

[17] A. S. Al-Ghamdi, "Analysis of Time Headways on Urban Roads: Case Study from Riyadh," *J. Transp. Eng.*, vol. 127, no. 4, pp. 289–294, Aug. 2001.

[18] F. B. Lin and S. Shen, "Relationships between queueing flows and presence detectors," *ITE J.*, vol. 55, no. 8, pp. 42–46, 1985.

[19] F.-B. Lin, *Evaluation of Queue Dissipation Simulation Models for Analysis of Presence Mode Full-Actuated Signal Control*. TRB, 1985.

[20] E. A. A. Shawaly, R. Ashworth, and C. J. D. Laurence, "A comparison of observed, estimated and simulated queue lengths and delays at oversaturated signalised junctions," *Traffic Eng. Control*, vol. 29, no. 12, 1988.

[21] R. T. Luttinen, "Statistical properties of vehicle time headways," *Transp. Res. Rec.*, p. 92, 1992.

[22] I. Greenberg, "The log normal distribution of headways," *Aust. Road Res.*, vol. 2, no. 7, 1966.

[23] J. E. Baerwald, M. J. Huber, and L. E. Keefer, *Transportation and traffic engineering handbook*. 1976.
[24] É. Borel, “Sur l’emploi du théoreme de Bernoulli pour faciliter le calcul d’une infinité de coefficients. Application au probleme de l’attentea un guichet,” *CR Acad. Sci. Paris*, vol. 214, pp. 452–456, 1942.
[25] É. B. Paris, “Sur l’emploi du théoreme de Bernoulli pour faciliter le calcul d’une infinité de coefficients. Application au probleme de l’attentea un guichet,” *CR Acad. Sci.*, vol. 214, pp. 452–456, 1942.
[26] J. C. Tanner, “A Derivation of the Borel Distribution,” *Biometrika*, vol. 48, no. 1/2, p. 222, Jun. 1961.
[27] B. Y. F. A. Haight and M. A. Breuer, “The Borel-Tanner distribution,” *Biometrika*, vol. 47, no. 1, pp. 143–150, 1960.