Data-Driven Methods for Improved Estimation and Control of an Urban Arterial Traffic Network

by

Leah Adrian Anderson

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alexandre M. Bayen, Chair
Professor Alexey Pozdnukhov
Professor Laurent El Ghaoui

Spring 2015

Data-Driven Methods for Improved Estimation and Control of an Urban Arterial Traffic Network

Abstract

Data-Driven Methods for Improved Estimation and Control of an Urban Arterial Traffic Network

by

Leah Adrian Anderson

Doctor of Philosophy in Engineering - Civil and Environmental Engineering

University of California, Berkeley

Professor Alexandre M. Bayen, Chair

Transportation is a field which is universal in our society: people from every country, culture or background are familiar with the challenges of getting around in our built environment. Yet what is not always so obvious to the average traveler is how the techniques and tools of designing, observing, and controlling our modern transportation networks are derived. In fact, the theory of traffic engineering has many gaps and unknowns are the topic of ongoing research efforts in the academic community. This work presents a collection of theoretical and practical methodologies to advance the state of traffic flow modeling, state estimation, and control of signalized roadways in particular. It uses theory from traditional transportation engineering, but also demonstrates the application of new tools from control theory and computer science to the specific application of signalized traffic networks.

First, two numerical modeling dynamics representing traffic flows on signalized arterials are presented: the well-known Cell Transmission Model, a discretization of the physical hydrodynamic laws believed to govern vehicle flows, and a new Vertical Cell Model which resembles classical "store-and-forward" models with the addition of transit delays and finite buffer capacities. Each of these models is implemented in a common software framework, which provides an ideal experimental platform for direct comparison of the competing dynamics. A chapter in this dissertation contributes a validation and comparison of the two models against real vehicle trajectory data on an existing signalized road network.

Accuracy and confidence in such traffic models requires complimentary methods of observing true traffic conditions to provide initial conditions and real-time state estimates. Yet there are many technological deficiencies in existing urban roadway detection systems that prevent the acquisition of a real-time estimate of arterial link state (or queue length) at signalized intersections. Hence this thesis also contains methodology to improve the estimates obtained from existing hardware by combining data from typical infrastructure sensors with new sources of Lagrangian probe measurements into a detailed model of flow dynamics. This technique was previously proposed for continuous-flow (freeway) networks, but required novel adaptions to be applied to interrupted-flow networks such as signalized road networks.

This dissertation next explores advancements in theoretically optimal control algorithms for statistically-modeled signalized queueing networks. In the context of a large body of previous work on flow-impeding control for vertical queueing networks, the practical challenges of traffic signal control are highlighted. Some of these challenges are tackled in the specific case of the *max pressure* controller, an algorithm derived from the field of communications networks that has been shown to optimize through-flow in an idealized network model.

The lack of adequate measurements or demand-volume data has historically been a major limitation in advancing research on signalized arterial traffic networks. Yet the current revolution of inexpensive storage and processing of "big data" shows promise for improving daily operations of existing road traffic networks without the need for expensive new hardware systems. One example of this potential appears is the case of traffic signal control. Existing traffic signals are capable of operating more efficiently by changing signal plans based on real-time demand measurements through a traffic responsive plan selection (TRPS) mode of operation (rather than depending on a rigid schedule for plan changes). However, this mode is rarely used in practice because its calibration process is not accessible or intuitive to traffic technicians. This dissertation presents application of statistical learning techniques to improve the process of calibrating and implementing an existing TRPS mechanism. A proof-of-concept implementation using historical sensor data from a busy urban intersection demonstrates that real operational improvements may be immediately achievable using existing sensing infrastructure.

Contents

Acknowledgments

I don't think anyone enters a Ph.D. program truly knowing how they will come out on the other end, but in my case the journey has taken so many turns that I can't really even remember what it was like at the origin. It's been a rough ride—but a worthwhile one for sure! What has made my time at Berkeley most rewarding are all of the inspiring people who I have met along the way. I have spent the last six years constantly surrounded with mentors, colleagues, and peers who have not only influenced my academic future with their wisdom and guidance, but have also exposed me to new world-views and personal philosophies that have completely changed my perspective on life.

I first decided to come to UC Berkeley after meeting with Alex Bayen between back-to-back appointments in a very hectic conference room at the CCIT Bancroft offices. I was initially impressed by this lively research environment, as well as Alex's personal enthusiasm for his projects and his interest in doing work that simultaneously contributes to both the theoretical community and the practical world—it did not take me long in the Ph.D program to appreciate how rare and incredibly challenging that prospect is!

I also soon realized how lucky I was to have been accepted into Alex's academic family. Not every Ph.D advisor would have taken so much time to truly understand my strengths and find ways to improve or overcome the areas in which I struggled. I cannot imagine being where I am today without his support, advice, and willingness to put up with my occasional periods of slow progress or sudden changes of focus. Furthermore, I have continued to be impressed at how he puts such a priority on making time for each of his students while juggling so many other academic and institutional responsibilities (as well as a growing family!). Alex has served as more than an academic advisor during my time in graduate school; he has also been an amazing role model.

Alexey Pozdnukhov has provided valuable advice in the completion of the work included in this dissertation, especially in the data-related field in which it seems I may build my future career. He specifically contributed to the chapter on learning techniques for traffic responsive control. I have also valued my candid discussions with Laurent El Ghaoui on the impacts and future implications of my work; he has provided unique insights into what I can expect in the next phase of my career. Other professors who have provided guidance and advice during my graduate school career include Adib Kanafani, Roberto Horowitz, Alexander Skabardonis, Pravin Varaiya, Joan Walker, and Steven Glaser. I feel fortunate to have been supported by such distinguished members of the academic community.

When I first arrived in Berkeley, I immediately found myself surrounded by the most brilliant (or insane...) group of people I had ever met. I would like to thank all of my colleagues and lab-mates who have served as teachers, tutors, partners, co-authors, fellow explorers, supporters, organizers, lunch buddies, and friends throughout my time at Berkeley: Andrew Tinka, Kevin Weekly, Carlos Oroza, Jon Beard, Christian Claudel, Qingfang Wu, Samitha Samaranayake, Jack Reilly, Tim Hunter, Aude Hofleitner, Walid Krichene, Jerome Thai, Cathy Wu, Francois Belletti, George Netscher, and Steve Yadlowsky—as well as all of the interns who dedicated many months of their educational careers to contribute to our

work. Thanks to all of my other grad student friends in the systems group and beyond for being a welcoming community and helping me find my way in the academic (and real) world: Anu, Brenda, Travis, Eloi, Fabien, Alex, Andre, and Farzana. A special thanks to Timmy Siauw for the countless lunches at Aki's and so many hours deliberating on the validity of our work or sharing thoughts on all of life's little/big challenges. There were times when I was not sure that I could make it through graduate school, but I always felt better after you helped me keep things in perspective.

Many of the graduate students, researchers, and administrators at California PATH have provided priceless technical and administrative support during my studies and furthermore have been a pleasure to work with. I would like to mention some of them here, in no particular order: Thomas Schreiter, Ethan Xuan, Gabriel Gomes, Anthony Patire, Greg Merritt, Francois Dion, Dimitrios Triantafyllos, Dongyan Su, Ajith Muralidharan, Matthew Wright, Yi Zhou, Rene Sanchez, Alexander Kurzhanskiy, Bill Sappington, Brian Peterson, and Joe Butler.

I also couldn't have finished my PhD without all of the people that have kept me sane outside of the (often seemingly-insurmountable) walls of the university. Shout outs to the Apes, Absolute, Team TaTas, the Flash, and all the yellow ladies who make life fun every weekend (go Drops!). Thanks to the staff at BSR for letting me put my artistic hobbies to good use for a few years. To Alicyn, Autumn, and Seanna, who somehow made it through North Dorm and four years of college with me (and yet we still want to hang out each other!): I am so impressed to see where life has taken us all, and I look forward to seeing what comes next. To those friends with whom I have lost touch over recent years, especially Beth, Stacie, Tom, Mara, Jenn, Ivan, Miriam, Dariya, and Neel: your contributions to my success has not been forgotten.

My husband David continues to be my biggest supporter in anything and everything that I want to do in life. Taking the time to get to know you was the best decision of my life. Your constant optimism and confidence provided the incentive that I needed to renew my dedication to my academic pursuits and many other life goals. With you, every day is a new adventure—and I still don't want it to slow down! Also to the rest of our family, Avery, Audrey, and Chloe: thank you for making me smile every day.

Last (but definitely not least!), I cannot forget the amazing upbringing provided to me by my parents, Barbara and Richard, my brothers, James and Phillip, and my grandmother, Marge. It seems too trivial to only say that I would not have achieved everything that I have accomplished today without your countless sacrifices and continuing influences on my life. I specifically want to thank you for all of the times that you showed me how good things come from hard work and dedication. I will be forever grateful for your enduring love and support.

Chapter 1

Introduction

1.1 Motivation

Traffic congestion is a growing problem in modern urban areas, causing an annual loss in productivity and fuel of over $120 billion in recent years [182]. While there has been a recent emphasis put on reducing traffic jams on our freeways, the operations of signalized roadways has not been given as much attention—even when approximately two-thirds of all miles driven in the United States are on non-freeway roads with traffic signals [64].

Today, there are more than 311,000 traffic signals in the United States [150]. A single busy intersection could easily serve more than 100,000 vehicles each day; in California a full ten percent of urban intersections see total volumes of over 60,000 vehicles per day [94]. Hence existing inefficiencies in signal control could impact the daily commutes of a significant number of travelers. Recent efforts to improve regional traffic signal management and operations have demonstrated benefit-cost ratios exceeding 40-to-1 [151]. This is an incredible economic efficiency compared to that expected from infrastructure projects aimed solely at expanding capacity without addressing other operational needs.

This work presents a collection of theoretical and practical methodologies to advance the state of traffic flow modeling, state estimation, and control of signalized roadways from a network-wide perspective. It uses theory from traditional transportation engineering, but also demonstrates the application of new tools from control theory and computer science to the specific application of signalized traffic networks.

Integrated corridor management

Integrated corridor management (ICM) refers to a comprehensive initiative by the United States Department of Transportation (USDOT) to find novel ways to make use of any underutilized capacity available in a regional transportation network with the objective of optimizing its throughput and reducing overall congestion.

A typical *corridor* to be managed in an ICM project involves multiple local traffic jurisdictions which all have varying technical standards and modes of operation, but share a

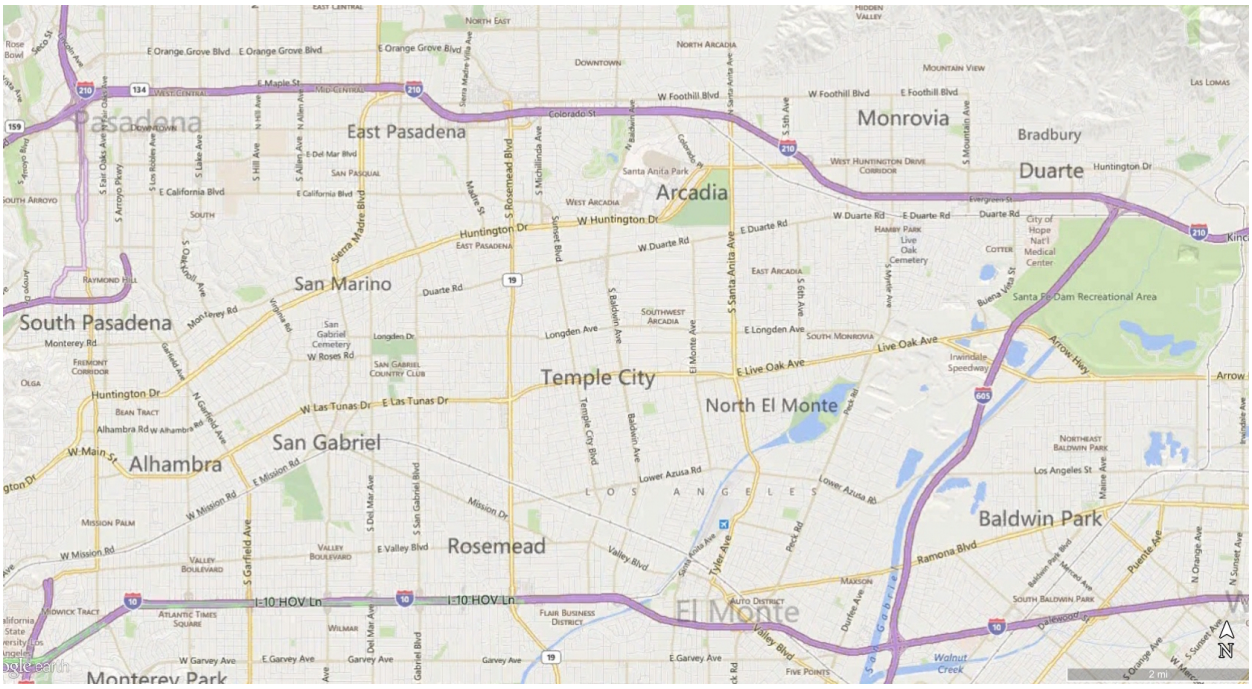

Figure 1.1: The I-210 corridor involves multiple freeways/major arterials as well as a commuter rail and transit system connecting at least nine independently-operated traffic jurisdictions. During morning and evening peak periods, this area suffers from major congestion on all East-West freeways. Dense urban development and geographical barriers prevent significant expansion of the existing road network.

significant portion of traffic going across or within a relatively small geographic region.

Figure 1.1, for example, illustrates a highly congested commuter corridor along Interstate 210 in the eastern suburbs of Los Angeles, California. The depicted area surrounding the I-210 is a prime example of an urban region with prominent peak-hour congestion on major throughways, yet no land available to expand the capacity of the existing road network. An ongoing ICM project sponsored by the California Department of Transportation (Caltrans) and the relevant local municipalities aims to coordinate the control mechanisms of the state-operated freeways and locally-operated arterials to shift demand patterns in a way that better utilizes the existing physical network capacities. This is believed to be achievable via development of a Decision Support System (DSS) which integrates active management with technical strategies such as predictive traffic modeling, intelligent controller design, novel monitoring strategies, advanced traveller information systems, and an increased emphasis on transit alternatives.

This dissertation develops a set of solutions for one specific aspect of an ICM project: the management of signalized road networks.

In the traffic community, high-capacity signalized roadways which serve as major thor-

oughfares in urban areas are referred to as *arterials* (in the sense that they are arteries of the mainline freeway flow). Both technological and theoretical advancement on the operations of arterials has lagged behind those being made on the modeling and control of freeways. One reason is that it is more difficult to extract meaningful observations of performance from point sensor measurements on arterials than on freeways because of the typical dense platooning that occurs downstream of signals [31]. This expected behavior also makes the dynamics of traffic on signalized roads much more difficult to represent mathematically than those of the uninterrupted freeway flows. Furthermore, intersection signal control involves many more parameters and degrees of freedom than single-stream ramp metering that is used to mitigate freeway congestion. These complexities have left significant room for advancement in the field of arterial traffic. Enterprising ICM initiatives then provide a clear opportunity to develop and test new arterial monitoring and control options that can contribute to network-wide performance improvements.

Arterial traffic management: modeling, estimation, and control

A comprehensive arterial management system addresses three main objectives:

1. analyze and model the dynamic characteristics of demands and queues on a network where flows are artificially impeded by signals,

2. obtain a real-time or near real-time estimate of the current traffic state to inform this model given existing (limited) sensing capabilities, and

3. use knowledge gained from the state observations and modeled predictions to limit queues and minimize unnecessary congestion via informed signal control.

The relationship between these components is presented in Figure 1.2. Effective control requires a dependable model of traffic dynamics, which in turn relies on an accurate estimate of traffic state.

Ideally, predictive modeling, estimation, and responsive control all operate using the same assumptions on network dynamics and a single underlying mathematical model. For example, freeway operators can use the Cell Transmission Model, (CTM) a numerical approximation of the kinematic wave model typically used to represent continuous flow, for an accurate representation of traffic queueing behaviors on freeways [48, 49]. CTM is attractive to researchers because of its ability to efficiently calculate the dynamics of *macroscopic* or large-scale flow behaviors that are observed at different vehicle densities. A large body of research has developed advanced estimation [132, 148, 28] and control [78, 79] capabilities which operate on CTM dynamics. Yet inherent characteristics of signalized roadways challenge the theoretical assumptions and validity of CTM for use on arterials.

First, CTM divides a continuous road into discrete lengths called "cells" where vehicle density can be considered uniform. However on signalized roads, periodic red signals generate rapid "stop-and-go" behaviors that can only be captured by extremely fine spatial

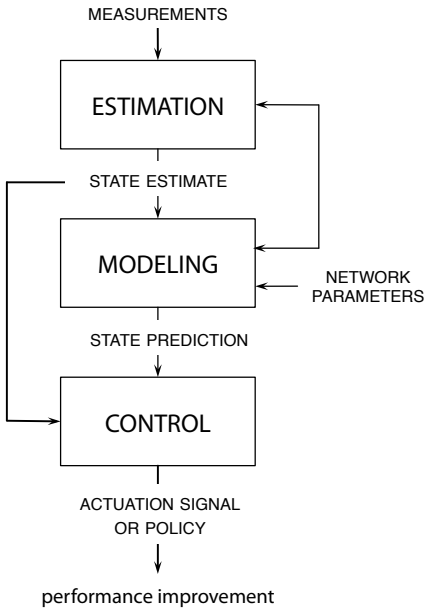

Figure 1.2: An ideal traffic management system shares a common modeling framework among three components.

discretization (and a correspondingly small temporal discretization) in a CTM implementation. This sharp queueing and de-queuing can have profound impacts on upstream discharge characteristics, and therefore must be included for an accurate representation of secondary congestion effects. Yet excessive discretization prevents the computational efficiency which makes CTM such an attractive modeling solution.

Second, traffic on arterial roadways does not necessarily follow the principle of *first-in-first-out* or *FIFO*: the first vehicle to arrive on the upstream end of a link will not necessarily be the first vehicle to leave the downstream end of the link because of the presence of parallel movement queues and differences in signalized queue release times. For example, a vehicle intending to turn right may leave a road link before or after a simultaneously-entering vehicle that intends to continue straight. Or in another instance, a left-turning vehicle that is waiting for a permissible gap in oncoming flow may impede the progression of an upstream through-bound vehicle. Yet CTM represents a continuous span of road between two consecutive freeway interchanges as a single stream of flow and makes the assumption that no "overtaking" or violation of FIFO occurs. While it is possible to divide individual movement queues into parallel CTM streams on a numerical representation of a signalized road network, it is difficult to model these types of interactions that realistically occur between distinct movement-based queues.

Hence traffic researchers (and managers) are left without a dominant methodology to analytically study queueing behaviors on signalized road networks. Simplified stochastic queueing models such as those used in the fields of logistics and communications can often

Figure 1.3: Many aspects of the queueing dynamics observed on signalized roadways make accurate modeling difficult. Sharp queueing shockwaves on signalized roadways would require excessive spatial discretization for accurate representation via CTM. Furthermore, a basic CTM is unable to model the interactions between movement-specific queues which form immediately upstream of a signalized intersection.

be used to predict equilibrium queueing conditions, but they are generally unable to re-create the congestion effects of phenomena such as spill-back due to queuing that exceeds a fixed storage capacity. When a more detailed or realistic representation is required, researchers typically use *microscopic* models to simulate arterial dynamics. These representations, often called *car-following models*, depend on detailed stochastic equations dictating the dynamics of individual vehicles. An overview of the modeling equations commonly used in these microscopic simulations is provided in [33]. Because of the level of detail and number of parameters involved in tracking distinct vehicle dynamics, microscopic models require a significant amount of calibration and tuning for accurate results. This in turn implies a large investment from local traffic agencies when building a custom simulation for each corridor that they wish to study. Researchers are also impeded by the fact that no closed-form expression is provided to describe the macroscopic traffic dynamics resulting from the aggregated interactions between the vehicles, and therefore network-level flows cannot be studied or controlled analytically using microscopic traffic simulations.

Therefore while modern innovations in estimation and control of signalized networks can be tested in microsimulation environments, they are largely lacking rigorous analytical guarantees. The work presented in this dissertation attempts to change this trend. Given the lack of a single comprehensive modeling framework, we present a collection of mathematical models and techniques that can be used to analytically derive novel solutions to each of the components illustrated in Figure 1.2. We ultimately propose that the techniques presented here can be used as a basis for a comprehensive arterial management system in an ICM project or in any general mode of operations.

1.2 Organization and overview of the contributions of this work

This dissertation proposes novel contributions to each of the three major components of the arterial management system depicted in Figure 1.2. The remainder of the work is organized as follows:

Chapter 2 provides background information on the challenges associated with each of these system components.

Chapter 3 details two theoretical modeling dynamics which can be used to represent traffic flows on signalized arterials: the well-known *Cell Transmission Model* (CTM), a direct discretization of the physical hydrodynamic laws believed to govern vehicle flows, and a new *Vertical Cell Model* (VCM) which resembles classical "store-and-forward" models with the addition of transit delays and finite buffer capacities. Each of these models is then implemented in a common software framework, which provides an ideal experimental platform for direct comparison of the competing dynamics. Specific contributions of this chapter are as follows:

- A derivation of a coherent, application-ready framework for a discrete-time vertical queuing model with finite link buffers that is compatible (interchangeable) with CTM link dynamics in implementation.
- An introduction of a link-state variable to facilitate the representation of spatial capacity without the need for explicit spatial discretization in link representation.
- A compatible implementation of both CTM and VCM on a shared network representation, and validation of each model against a high-fidelity ground truth data set [10].

Chapter 4 describes a methodology to overcome the technological deficiencies in existing urban roadway detection systems to achieve a real-time estimate of arterial link states/queue lengths. This technique was previously proposed for continuous-flow (freeway) networks, but required novel adaptions to be applied to interrupted-flow networks such as signalized road networks. The following contributions are presented:

- An explicit demonstration of the challenges associated with previously proposed estimation algorithms.
- A formulation of constraints on the set of feasible PDE boundary conditions that must satisfy observed measurements of point-to-point travel-times.
- An adaptation of a PDE-based state estimation procedure which achieves fusion of multiple different types of measurements (including aggregated counts/volumes, densities, trajectories, and point-to-point travel times) and the possibility of near-real-time queue estimation.
- An implementation of this algorithm and validation of resulting estimates against link states extracted from high-fidelity ground truth data [11].

Chapter 5 explores advancements in theoretically-optimizing control algorithms for statistically-modeled signalized queueing networks. In the context of a large body of previous work on flow-impeding control for vertical queueing networks, the practical challenges of traffic signal control are highlighted. Some of these challenges are tackled in the specific case of the *max pressure* controller, an existing algorithm derived from the field of communications networks that has been shown to optimize through-flow in an idealized network model. Specific contributions to this topic include:

- A formulation of a *cycle-based max pressure* (Cb-MP) extension to the max pressure controller that is motivated by practical hardware and safety constraints on realistic traffic signals
- an extension of the guarantee of network stability given applications of the Cn-MP controller [171].

- An implementation of Cb-MP on a calibrated model in the Aimsun micro-simulation platform.
- An experimental finding that, during periods of high congestion, a "naive" cycle-based max pressure controller could out-perform the existing highly-tuned actuated controllers in terms of various delay metrics [12].

Chapter 6 describes an application of statistical learning techniques to improve the operations of an existing *Traffic Responsive Plan Selection* (TRPS) mechanism. Many modern signal controllers are capable of operating in a mode in which the choice of operational signal plan is responsive to detected changes in demands rather then solely dependent on a fixed operation schedule. However this mode is rarely implemented. It is believed that this is largely due to the fact that the existing plan selection mechanism is rigid and complex, and thus unintuitive to calibrate properly. In this chapter we present the following solutions to facilitate the adoption of TRPS operations:

- An analysis of sub-optimal performance due to rigidity in plan switching schedules using real data at an intersection in the I-210 corridor.
- A data-driven methodology for designing the detector weights and plan selection table used by a generic TRPS mechanism.
- A comprehensive calibration procedure for implementing a TRPS system without the need for designing and tuning new signal timing plans.
- A proof-of-concept implementation of a TRPS controller designed using the proposed calibration procedure, and an analysis of the delays resulting from its use (compared to optimal and current scheduled operations).

Ultimately, a comprehensive conclusion to the entire body of work is given in **Chapter 7**.

Chapter 2

Arterial traffic network management technologies: state of practice

We begin by introducing the features and terminology relevant to a signalized road network and reviewing the existing work that has contributed to the current state of practical arterial management.

2.1 Fundamental features of a signalized road network

Intersection design

Classical operations typically assume that traffic signals placed at intersections operate in a cyclical nature, where a signal *cycle* is composed of multiple *phases* of traffic flow. Each phase (ϕ) consists of a set of *movements* or streams of flow that can be simultaneously permitted to flow through an intersection without causing collisions. For example, a typical intersection of two bidirectional roadways has four approaches, and each approach has three movements: left, through, and right (ignoring U-turns for simplicity). These movements are illustrated in Figure 2.1.

Signal controllers

To specify control parameters, the set of signal phases are often visually illustrated in a *ring-and-barrier diagram*, as in Figure 2.2. The organization of this diagram is originally accredited to a standard developed by the National Electrical Manufacturer's Association (NEMA) [154], and is thus the described phases are sometimes referred to as the "NEMA phases".

Each box of this diagram contains a distinct signal phase. By convention, phases 1-4 are assigned to the top row (or *ring*) and 5-8 are on the bottom row (ring). Horizontal *barriers* separate ϕ_1, ϕ_2, ϕ_5, andϕ_6 from ϕ_3, ϕ_4, ϕ_7, and ϕ_8. Within a barrier, a signal may safely actuate one of the two phases from the top ring and one of the two phases from the bottom

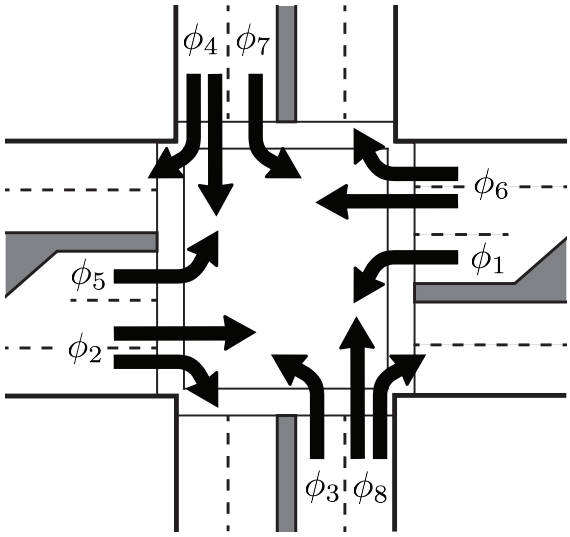

Figure 2.1: An example intersection of two bidirectional roadways has twelve distinct possible movements, which can be divided into eight standard phases $\phi_1 - \phi_8$.

ring simultaneously. For example, ϕ_1 can be shown a green light at the same time as either ϕ_5 or ϕ_6, but not both. While phase transitions on each ring within a barrier can happen asynchronously, transitions across barriers must occur simultaneously on both rings.

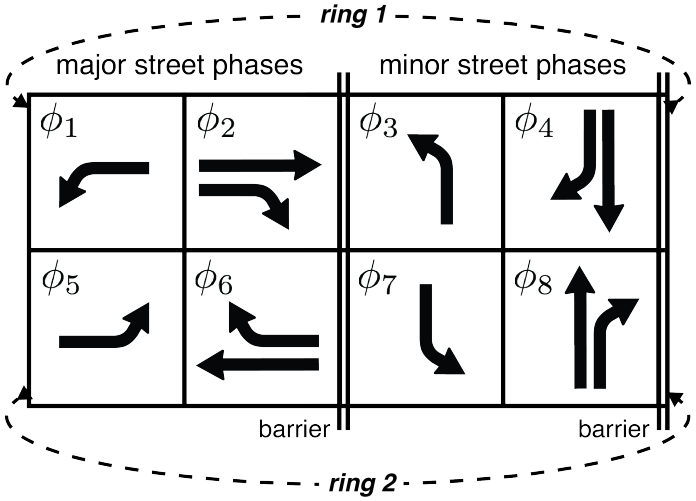

Figure 2.2: Eight phases are placed in a ring-and-barrier diagram to illustrate possible combinations of simultaneously actuated movements. Movements represented with a dashed arrow are potentially *permissive* movements, which must yield to conflicting traffic. By convention, the phases originating from a major street are in the left side barrier, and those involving the minor street are in the right barrier.

Figure 2.2 is only an example of a possible set of phases; the combination and order of movements selected for each phase may vary by design of the traffic operator.

Variables

Define the following variables which will be used throughout this chapter (with subscripts referring to phase-specific or approach-specific values where denoted):

ϕ_k := signal phase k

d := average delay per vehicle (seconds)

C := intersection signal cycle time (seconds)

g_k (r_k) := the green (red) time allocated to ϕ_k

λ_k := green split (proportion of cycle, or g_k/C) for ϕ_k

q := demand volume (vehicles per hour)

s := saturation flow (assume 1800-1900 vehicles per hour)

$x = \frac{q}{\lambda s} \in [0, 1]$:= volume-to-capacity ratio, also called the *degree of saturation* for a phase

n := number of phases in a cycle,

$L = nl + R$:= total *lost time* or non-actuated time per cycle; the sum of startup and yellow time per each phase (nl) plus the all-red time (R, about 10 seconds per cycle)

$y = q/s$:= the ratio of demand flow to saturation flow for an approach or phase (as specified where applicable)

$Y = \sum_k y_k$:= sum of demand-to-saturation-flow ratios for each phase in a cycle

Q := a measure of queue length or number of vehicles waiting for an approach or phase

ν := the number of lanes associated with a specific queue or approach

Terminology

The terminology we will use to describe traffic conditions on signalized networks is as formalized in [169]:

- If no significant queues are observed at the end of a green signal period, the intersection is **uncongested**.
- If significant queues persist at the end of a green period, the intersection is **congested**.

- Congested intersections are then further classified as **saturated** if the observed queues are small such that the excess delay and stoppage is only observed locally on the link in consideration. Typically this implies that the degree of saturation (x) is close to or slightly exceeding 1 for one or more phases in the intersection.

- If queues build to the point where the operation of one or more of the adjacent upstream intersections are affected, the intersection is **oversaturated**. In this case, $x > 1$ for a significant period of time. Oversaturation is also referred to in terms of **cycle failures**, or cycles in which queues for one or more movements were not completely serviced.

2.2 Development of traffic flow models in literature

Many aspects of traffic engineering, from the economics of planning and pricing to the operations of control and traveler information, would benefit from a comprehensive model of road dynamics. Researchers have therefore sought a robust mathematical theory of vehicle traffic flow for over a century. The earliest published work in the field of traffic modeling was done by economists seeking to explain the social costs of road building and usage [170, 108, 211]. While these early efforts produced only idealized models of equilibrium demands, they sparked an interest in producing more detailed models of local and transient traffic dynamics for analysis purposes.

The complicated interrupted-flow dynamics of signalized traffic networks are particularly difficult to model accurately. The following paragraphs trace the academic developments that are specifically related to modeling the characteristics of signalized intersections and that motivate the work described in this dissertation. For a more comprehensive review of many historical and modern approaches to traffic modeling in general and a description of their use cases, see [97].

Stochastic equilibrium delay models

In 1956, Beckmann, McGuire, and Winsten were the first to rigorously formulate the equilibrium flow concepts that had been previously described by economists such as Wardrop with the inclusion of the effects of congestion on journey time [20]. The work in their book is considered a seminal contribution to the theory of traffic flow. It was also among the first to utilize newly developed concepts in optimization and mathematical economics, including the now widely-known nonlinear programming framework of Kuhn and Tucker [111]. A discussions of their specific innovations is available in [32]. In their analysis of signalized roadways in particular, Beckmann, McGuire and Winsten proposed a simple queuing model that could be used to derive a relationship between average delay per vehicle and the average length of an approach queue at the beginning of a red phase [20]. This is believed to be the first quantified model of the expected delay caused by fixed-time signal settings at a traffic intersection.

Beckman's model assumes discrete Bernoulli arrivals: at any fixed time period, a car arrives into the queue with probability α, and this probability is independent of previous arrivals at each time period. Departures are assumed deterministic and constant (one per time period) upon green and disallowed during red. Thus queue length state $Q(t)$ is modeled as a Markov chain with simple transition probabilities corresponding to queue growth during a red period or queue shrinkage during a green period. The length of a queue at approach a just before the first red of the k^{th} cycle can also be written as a Markov equation:

$$Q_a(k+1) = \max\{Q_a(k) + a_a(k) - g_a, 0\} \tag{2.1}$$

where arrivals $a_a(k)$ in a cycle of length $C = r_a + g_a$ has a binomial distribution:

$$P\{a(k) = m\} = \binom{C}{m} (1-\alpha)^{C-m} \alpha^m \tag{2.2}$$

Beckmann et al. then derived a formulation for the average waiting time (delay) per vehicle in an approach a in terms of green splits and the (expected) end-of-green queue length:

$$d_a(k) = \frac{r_a}{C(1-\alpha_a)} \left[\frac{\mathrm{E}(Q_a(k))}{\alpha_a} + \frac{(r_a+1)}{2} \right] \tag{2.3}$$

The usefulness of this and other derivative delay models (such as [138]) is ultimately limited by the presence of the expected overflow queue term and the strong assumption of binomial arrivals. This spawned a search for a more universal formula using more approximate queueing dynamics and incorporating heuristic adjustments. The most well-known of the models generated by this effort is one credited to Webster [212]. According *Webster's delay formula*, the average delay experienced by a vehicle at an intersection approach a is written as:

$$d_a = \frac{C(1-\lambda_a)^2}{2(1-\lambda_k x_a)^2} + \frac{x_a^2}{2q_a(1-x_a)} - 0.65 \left(\frac{C}{q_a^2} \right)^{\frac{1}{3}} x_a^{(2+5\lambda_a)} \tag{2.4}$$

The first term in (2.4) originates from a derivation of delay when traffic arrives at a uniform rate corresponding to degree of saturation x. The second term accounts for randomness in arrivals: it assumes a Poisson arrival distribution and constant departures at a rate equal to the signal capacity. This departure rate is obviously non-realistic, as departures actually only occur upon green and can achieve a rate up to the saturation flow of the approach. Finally, the third term is an empirical adjustment factor that typically evaluates to approximately 5-15% of the value of the first two terms.

Webster also provided a formula to approximate the average queues at the beginning of a green period:

$$Q_a = \max\left\{ \left(\frac{q_a r_a}{2} + q_a d_a \right), q_a r_a \right\} \tag{2.5}$$

Empirical adjustments to this estimate (again of 5-10%) imply a revision to the following expression:

$$Q_a = \max\left\{ q_a \left(\frac{r_a}{2} + d_a \right) \left(1 + \frac{q_a j}{\nu_a v} \right), q_a r_a \left(1 + \frac{q_a j}{\nu_a v} \right) \right\} \tag{2.6}$$

where j is the average spacing between queued vehicles and v is the free-flow speed.

Furthermore, according to these assumed dynamics, he proportion of vehicles which stop at least once is given by:

$$P = \frac{1-\lambda}{1-y_a} \tag{2.7}$$

and the average number of stops and starts per vehicle in each cycle is given by:

$$N = \begin{cases} \frac{Q_a}{q_a C(1-y_a)} & \text{if undersaturated, so } g_a > \frac{Q_a}{s-q_a} \\ \frac{Q_a}{q_a C} + \lambda_a & \text{if saturated/oversaturated, so } g_a < \frac{Q_a}{s-q_a} \end{cases} \tag{2.8}$$

The concept utilized by Webster of dividing delay estimations to uniform and random components persists today in the practical delay-calculation methodology suggested by the *Highway Capacity Manual* (HCM). The 2010 HCM suggests an intersection delay formula with substantially the same uniform delay term as (2.4), but a significantly modified random delay term and a third adjustment term that involves analysis of the impacts of adjustments due to actuation the level of coordination in neighboring signals. The resulting delay equation is ultimately suggested as the basis for determining Level of Service (LOS) of a signal, a primary metric for evaluating intersection performance on existing roadways.

A few years after the work of Webster, Miller developed a competing delay model in which arrivals can be considered any stationary point process with a finite variance for periods of approximately 30 seconds and departures are (as in previous models) considered uniform during green phases [144]. To account for the assumptions of instantaneous acceleration and departure, he also adds additional "lost time" to the beginning of green periods and introduces the concept of an "effective green phase" which includes the remainder of the green and the following yellow period. According to this work, the expected vehicle-delay over a given red phase (of length $C - g_a$) is calculated as:

$$\begin{aligned} \mathbb{E}\{d_a\} &= \mathbb{E}\left\{\int_0^{C-g_a} Q_a(t)dt\right\} = \int_0^{C-g_a} \mathbb{E}\left\{Q_a(t)\right\} dt \\ &= \int_0^{C-g_a} (Q_a(k) + q_a t)dt \\ &= (C-g_a)\left[Q_a(k) + \tfrac{1}{2}q_a(C-g_a)\right] \end{aligned} \tag{2.9}$$

where $Q_a(k)$ represents the number of vehicles in the queue for approach a at the end of the k^{th} effective green and $Q_a(t)$ is the number of vehicles in the queue at time t seconds after the end of an effective green.

To calculate (2.9), Miller uses the same principle as Beckmann et al.: first, calculate the total delay as if the effective green time were infinite and thus the queue is eventually exhausted. Then limit the green phase to end up with just the delay incurred during the (known) finite time g. This results in the following formula for average delay per cycle per arriving vehicle:

$$d_a = \frac{1-g_a/C}{2(s-q_a)}\left\{\frac{2s}{q_a}\mathrm{E}(Q_a(k)) + s(C-g_a) + I_a - 1 + \frac{q_a}{s}\right\} \tag{2.10}$$

where I_a is the variance-to-mean ratio of the distribution of vehicle arrivals. He then uses a simple modeling assumption (similar to (2.1)) to determine an expression for $\mathbb{E}\{Q(k)\}$:

$$Q_a(k+1) = Q_a(k) + a_a(k+1) - sg + \delta_{x+1} \tag{2.11}$$

where δ_{x+1} is a compensating function to ensure that $q(x+1)$ is never negative (such that it is always true that $q_{x+1} \cdot \delta_{x+1} = 0$). This ultimately results in the approximation

$$\mathbb{E}(Q_a(k)) \approx \frac{(2x-1)}{2(1-x)} \cdot I_a \quad \text{and} \quad \mathrm{E}(Q_a(k)) = 0 \text{ when } x \leq \tfrac{1}{2}, \tag{2.12}$$

The expression for average delay per vehicle (on approach a) becomes:

$$d_a = \frac{1-\lambda_a}{2(s-q_a)} \left\{ y_a I_a \frac{(2x-1)}{(1-x)} + s(C-g_a) + I_a - 1 + y_a \right\} \tag{2.13}$$

Subsequent efforts to increase the practical value of the formulation of (2.13) noted that the first two terms typically dominated the valuation, and thus Miller later suggests the following simplification [143]:

$$d_a = \frac{1-\lambda_a}{s(1-y_a)} \left[C(1-\lambda_a) + \frac{2Q_a^0}{q} \right] \tag{2.14}$$

where the initial overflow queue Q_a^0 is calculated assuming Poisson arrivals and fixed green-time service rate:

$$Q_a^0 = \frac{\exp\left[-1.33\sqrt{sg_a(1-x_a)/x_a}\right]}{2(1-x_a)} \tag{2.15}$$

There have been a handful of attempts at validating and comparing delay models such as (2.3), (2.4), and (2.14). Experimental validation is made challenging by the lack of ability to measure actual vehicle delays in practice. Historically, researchers have had to develop their own metrics for comparison against approximate observations [212, 9, 102, 159, 45]. More recent attempts have made use of microsimulation tools to validate delay models [59].

It is universally recognized that most delay models produce effectively identical results in under-saturated conditions ($x < 0.8$). But as demand approaches capacity, steady-state models differ in performance. For example, Webster's delay formula approaches infinity as the degree of saturation (x) approaches 1. This is obviously an undesirable behavior. Hence the primary usefulness of these types of intersection delay models is for planning and efficient estimation of aggregated delays in normal operating conditions; evaluating controller performance in congested conditions requires the use of a time-dependent model.

Time-dependent dynamic queueing models

As queues grow to saturate (and then oversaturate) intersections, the desired objective of a network operator shifts from minimizing local delay and number of stops to maximizing

network throughput by reducing the spatial extent and rate of spread of congestion [169]. Such an objective requires the use of a dynamic intersection model to represent growth and dissipation of large queues.

Fundamental store-and-forward models

The first introduction of a dynamic model for a network of signalized roadways appears to have come from Gazis et al. [74, 73], which introduces a discrete-time *store-and-forward* model of flows between controlled intersections for the purposes of delay-minimizing optimal control. This form of *vertical queueing model* introduces a representation of an intersection as a graphical "node" served by road "links" with time-varying demands. In a vertical queueing model, vehicles waiting to be served by an intersection are stored in an (infinite) queue on the upstream link, and the effects of intra-link congestion on transit times are not considered.

The characteristics of Gazis' original modeling framework were highly limiting: flows were only unidirectional and unconstrained by downstream congestion, transit delays were ignored, and representations of controller switching behaviors were highly simplified. Yet extensions quickly introduced networks of many coupled intersections with more realistic representations of intersections and constraints on flows [72, 189, 52, 166]. Notably, Michalopoulos and Stephanopoulos [141] made the model more applicable to flows in congested regimes by introducing node-to-node transit delays and intersection transmission limitations due to downstream congestion.

Kinematic wave model

A major shift in the traffic modeling community occurred with efforts to explicitly define road capacity. Researchers and practitioners observed that an excessive increase in vehicle concentration leads to a reduction in mean speed, and thus a decrease in the overall vehicle flow rate. Two distinct domains of traffic dynamics were proposed: a *free-flow* domain, where peak velocity is attained over a range of lower densities, and a high-density *congested* domain where velocity (and thus flow) drops with every additional increase in density.

Many researchers have suggested explicit definitions for the relationship between vehicle density and flow that can draw distinctions between these two regimes. Such an equation is commonly known as the *flux function* or *fundamental diagram* of traffic flow. It is typically assumed to be a concave function where flow monotonically increases with density up to a point called the *critical density*, after which flow begins to decrease with increasing density. The critical density therefore defines the boundary between the aforementioned free-flow regime and the higher-density congested regime. If density approaches a maximum *jam density*, flow approaches zero because vehicles are stuck in slow-moving queues.

The earliest documented form for a fundamental diagram was implied in the 1935 work of Greenshields [81], in which the following relationship between observed speed v and vehicle

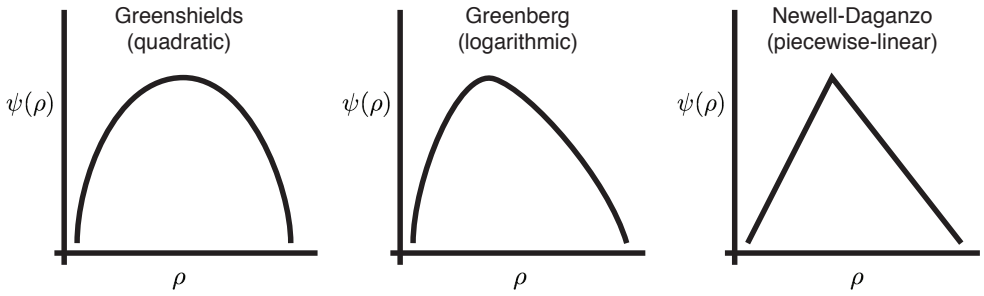

Figure 2.3: Various flux functions or fundamental diagrams for traffic flow have been proposed and justified with experimental data. The linear analytical expression of the Newell-Daganzo fundamental diagram has contributed to its widespread adoption in recent modeling efforts.

density ρ was experimentally postulated:

$$v = \mathbf{V}\left(1 - \frac{\rho}{\rho^{max}}\right) \tag{2.16}$$

for maximum road velocity (speed limit) $\mathbf{V}$ and maximum (jam) density ρ^{max}. Given that flow is equal to density times velocity, the resulting *Greenshields flux function* is parabolic:

$$\psi(\rho) = \mathbf{V}\rho\left(1 - \frac{\rho}{\rho^{max}}\right) \tag{2.17}$$

A later study by Greenberg used principles from fluid dynamics to propose a logarithmic relationship between velocity and density, leading to a logarithmic flow-density relation [80]:

$$v = \mathbf{V}\ln\left(\frac{\rho^{max}}{\rho}\right) \quad \Longrightarrow \quad \psi(\rho) = \mathbf{V}\rho\ln\left(\frac{\rho^{max}}{\rho}\right) \tag{2.18}$$

Recently a simplified piecewise-linear flux model known as the *Newell-Daganzo flux function* has gained popularity in traffic literature. This is due largely to its relative simplicity and beneficial analytical properties [156, 48]. Explicitly, it is written as

$$\psi(\rho) = \min\left\{\mathbf{V}\rho,\ c,\ \mathbf{W}\left(\rho^{\max} - \rho\right)\right\} \tag{2.19}$$

where c is the *capacity* or maximum possible flow rate, ρ^c is the critical density (corresponding to that maximum flow rate), and $\mathbf{W} = \frac{c}{\rho^{max} - \rho^c}$.

A visual comparison of these three fundamental diagrams is portrayed in Figure 2.3.

The work of Lighthill and Whitham [116] (and concurrent work of Richards [174]) was the first to introduce a dynamic traffic model in which traffic flows at freeway bottlenecks are modeled as hydrodynamic waves that were governed (in equilibrium) by one of these flux

functions. Their work used the method of characteristics to derive an explicit solution for the observed *shockwaves* that define the boundaries between free-flow and congested regimes.

Define $\rho(t, x)$ to be the density of vehicles at spatial location x and time t, and $\psi(\rho(t, x))$ to be some chosen convex flux function. The Lighthill-Whitham-Richards (LWR) model is defined by the following flow-conserving *partial differential equation* (PDE):

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial \psi(\rho(t,x))}{\partial x} = 0 \tag{2.20}$$

This first-order model is simply derived by applying the principle of mass conservation to the flow of traffic across a finite region: the change in point density in time is inherently equal to the net vehicle flows at that point in space-time.

While Lighthill and Whitham discussed challenges to the application of his theory to signalized junctions in their original work [116], it was generally believed that this step towards *horizontal queueing models* would lead to improvements in signalized traffic models over vertical queueing variants (such as the store-and-forward technique) because it could represent the backwards-propagating shockwaves caused by dissipation delays after a signal releases a stationary queue. Detailed derivations of the specific shockwave characteristics observed on signalized roadways were later presented in [177, 198, 142].

Cell Transmission Model (CTM)

Hyperbolic conservation laws such as the LWR PDE (2.20) have discontinuous solutions, evidenced in this case by the queue formation and dissipation shockwaves visible in continuous traffic flows. The application of standard finite difference methods to generate numerical solutions to the LWR PDE would generate instabilities or inaccuracies at these shock boundaries. However it has been shown that the Godunov difference scheme [77] provides a stable first-order numerical approximation of the shock propagations in a conservation law with concave flux function $\psi(\cdot)$.

The Godunov scheme is applied by discretizing the temporal variable (t) into short intervals of length Δt and dividing the spatial component (x) into finite-length *cells* within which the system state ρ can be considered uniform. Define a cell $i = [x', x' + \Delta x]$ and time step $k = [t', t' + \Delta t]$. Given a constant initial state $\rho(\tilde{x}, t') = \rho_i(k)$ for all $\tilde{x} \in [x', x' + \Delta x]$, the approximate state $\rho_i(k+1)$ assigned to cell i at time step $(k+1)$ is equal to the spatial average of the explicit solution $\rho(\tilde{x}, t' + \Delta t)$:

$$\rho_i(k+1) = \frac{1}{\Delta x} \int_{x'}^{x'+\Delta x} \rho(y, t' + \Delta t) dy \tag{2.21}$$

$$= \rho_i(k) + \frac{\Delta t}{\Delta x}(f_{i-1}(k) - f_i(k)) \tag{2.22}$$

where

$$f_i(k) = \frac{1}{\Delta t} \int_{t'}^{t'+\Delta t} \psi\big(\rho(x', s)\big) ds \tag{2.23}$$

The application of the Godunov approximation scheme to the LWR PDE with a Newell-Daganzo (triangular) fundamental diagram results in a convergent numerical approximation known as the *Cell Transmission Model* (CTM) [48, 49, 115]. The mathematical details of CTM are further described in Section 3.2.

Field data suggests that CTM closely fits observations of flows on freeways or highways with few interruptions [193, 194], yet validation of CTM on real arterial networks with short signalized intersections is very limited to our knowledge. One existing numerical comparison of CTM to a vertical queueing model on an artificial grid network is presented in [223].

In recent years, CTM has been considered by many researchers to be the standard in macroscopic modeling of traffic flows. It has been used to directly design traffic controllers for freeways [78], and has even been adopted in version 13 of the widely-used traffic optimization package TRANSYT [132]. An analysis of the dynamic properties of CTM for use in control is provided in [79].

More recently, there have been many algorithms proposed for optimal intersection signal control schemes based on the analytical dynamics of CTM [6, 55, 126, 120, 19]. It is still argued, however, that the complexity and high computational requirements of these CTM-based control schemes on detailed urban networks make them impractical for the real-time application for which they were designed [4].

2.3 Arterial monitoring and detection

While it is easy for a driver sitting in a traffic jam to see that the roadway that he is traveling on is congested, remote observation of all of the roadways influencing global congestion patterns on a spatially-distributed road network is technologically challenging. According to a survey conducted for the National Transportation Operations Coalition's *2012 National Signal Report Card*, about half of all traffic management agencies have "little to no regular, ongoing program for performance monitoring to assess operational objectives", and when data is collected there is seldom a methodology in place to assess the quality of that data [150]. Yet most (if not all) of the primary objectives of a traffic manager require reliable estimation of current traffic conditions. Dynamic traffic models (and applications of these models) require an estimate of the initial state of traffic to accurately predict future dynamics, and traffic-responsive or adaptive arterial signalization schemes by definition rely on a sensor feedback mechanism to provide some measure of current congestion.

This section details the current state of traffic monitoring at signalized intersections.

Common detector layouts

At a typical signalized intersection that is equipped for actuated control, sensors are placed at the locations illustrated in Figure 2.4. These sensors are designed with the following objectives:

1. to determine vehicle presence in a queue at a specific movement,

2. to determine flow continuity for phase extensions, and

3. to detect queues beyond a certain threshold for left turn movements.

The selection of sensor type and placement is dependent on the specific objective desired.

There are two general categories of sensors used at signalized intersections: *presence sensors* and *passage sensors*. The names are fairly descriptive of their functionalities: presence sensors indicate the presence of a vehicle when it is slow-moving (such as in a stationary queue at a signal), and passage sensors detect vehicles passing a point in space at a speed of more than 3 to 5 mph [107]. Intuitively, presence sensors are most often placed at or near the stop line, or the front of an expected queue, while passage detectors are placed at a reasonable distance upstream from the stop line.

The primary function of this type of detector layout is local actuated signal control, which will be described in detail in Section 2.4. Measurements are most often only transmitted to the local logic unit that is housed in a nearby controller cabinet, and are not typically returned to a central operations center for network-wide monitoring.

Existing sensor technologies

Inductive loops are the most common type of sensing technology deployed at signalized intersections in the United States. These sensors are composed of a wire loop installed directly into the pavement, which is supplied with power from a detector unit in a local controller cabinet. Current running through the wire creates a magnetic field within the loop. When no vehicles are present on the pavement above the loop, the detector receives a signal at a known baseline resonant frequency. However as a vehicle (or any large metal object) passes over the point of installation, the loop induces a current in this object which in turn increases the resonance of the loop. This ultimately causes a change in the frequency of the signal received by the detector unit [107].

Loop sensors are most often configured to be presence detectors for signal-actuation purposes, but smaller loops can also be used to count passing vehicles. The inductive loop is favored for its robustness to weather conditions and wear (compared to other alternatives) [31].

Image-processing based video detection systems have risen in popularity in recent decades as an alternative to the standard loop detectors. They claim the ability to simultaneously collect information about vehicle counts, vehicle presence, lane occupancy, and speed—and even break down these measurements by vehicle class (i.e. cars, trucks, and motorcycles) [107]. Yet their accuracy has been called into question by multiple studies [31].

Other less-commonly deployed intersection sensor technologies include radar or microwave sensors for speed measurements, magnetometers which serve largely the same purpose as loops, and closed-circuit television (CCTV) networks for incident detection and manual monitoring of operations.

None of these common detectors are particularly good at providing one of the most useful estimates of arterial link-state: the instantaneous *link vehicle-count*, or the number of vehicles

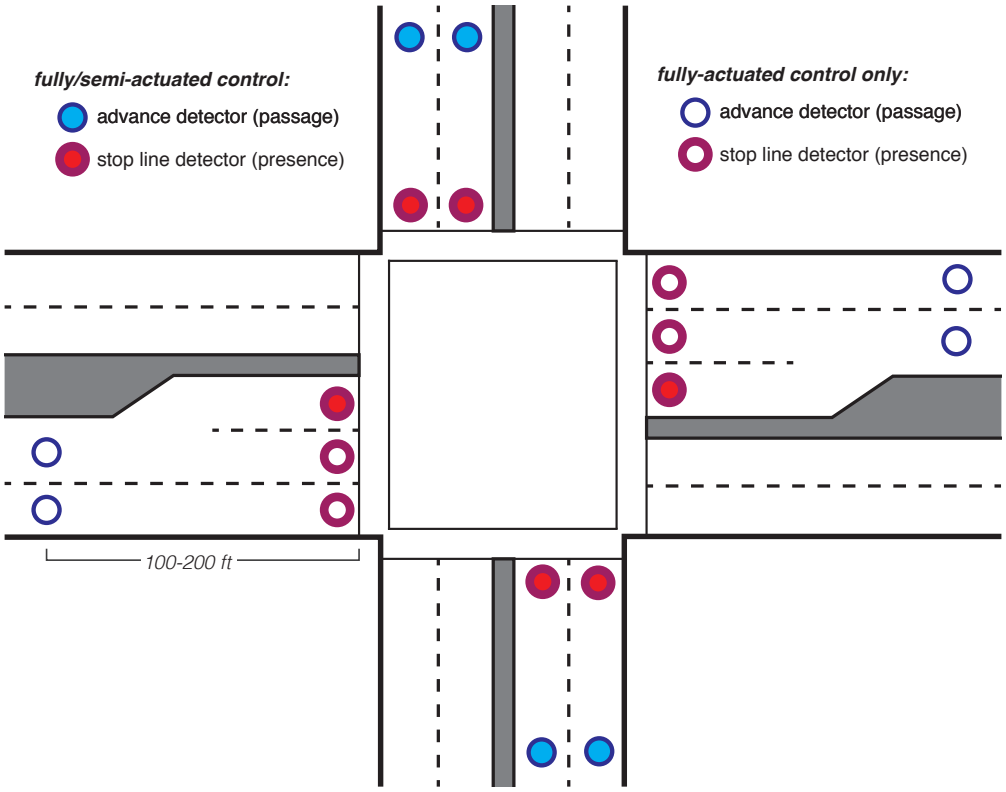

Figure 2.4: *Stop line (or stop bar) detectors* indicate the presence of a vehicle at a movement queue. This information can be used to invoke or prevent the actuation of the relevant movement in a signal cycle. *Advance detectors* are typically passage detectors that are located between 100 and 200 feet upstream of the stop line for a queue. They detect approaching flow profiles to signal the need for green extensions to satisfy immediate demands. All sensor actuations are communicated to local controller hardware that is encoded with the parameters and logic necessary to implement actuated signal control. See Section 2.4 for an overview of actuated signal control logic.

currently waiting for a signal-constrained movement [162]. If both advance and stop line count measurements are centralized for processing, the corresponding arrival and departure flow profiles can also be used to provide a rough estimate of queue length via input-output techniques. Yet these techniques are imprecise and error-prone (see Chapter 4 for more details). Furthermore, long queues that extend beyond the position of the advance detector cannot be estimated. If a video detection system is present, more robust queue estimation procedures may be available. A series of sequential video feeds positioned in each approaching movement can report a precise position of the back of a queue by observing the most upstream stopped vehicle. This queue length estimate can be updated at a rate of once per every 10 seconds, and can then be immediately used to estimate experienced control delays [31]. Yet video detectors are far from universal: the expense and maintenance requirements of video detection systems are barriers for many traffic agencies [162]. Furthermore, the range of video cameras is small and can easily be obscured by environmental obstacles.

The following observations on the state of urban traffic sensing are based on a survey of US state and city traffic agencies in the mid-2000s [31]:

- Surveillance is mostly being conducted using loops, video detection systems, and closed-circuit television (CCTV) networks. Less than half of the intersections governed by the surveyed agencies have loop of video detections that provide real-time data collection capabilities. Only about 4% of the arterial miles governed by these agencies are monitored by CCTV in 2002.

- Agencies use inductive loops or video detection for about 85% of intersection control operations; approximately 75% loops and 10% video detection was documented in 2005.

- Data from passive toll tags (radio-frequency transponders) are being collected and used by less than 1% of large cities.

- Even if sensors are installed, real-time communication capabilities typically do not exist—and hence measured data is very rarely centralized and stored.

Attempts to pursue field data to complete the present work has supported these observations.

Emerging sensors for arterial traffic

Modern innovations are bringing new functionalities to intersection sensing. For example, advanced magnetometers promise the ability to measure presence and counts at a higher reliability than traditional loop detectors [83, 113]. They can also be used to estimate intersection split ratios and delays by matching the magnetic signatures of vehicles at intersection approaches and egresses [112, 181, 106]. Wireless capabilities of these sensors facilitate communication for centralized data archival.

Other new data sources break the traditional concept of sensing infrastructure, most notably the aggregation of trajectory data from the global positioning system (GPS) sensors

on the mobile phones of travelers [222, 17, 91]. Researchers have proposed using mobile phone data for applications such as acquiring travel time estimates [88, 27], calibrating the parameters of dynamic models [93, 25], estimating queue lengths [16, 40], and building network origin-destination demand matrices [197], among others. While most of the proposed procedures require accuracy and sampling rate higher than those that are currently available from cell phones on US roads [51, 41], the concepts have drawn attention from the community developing connected vehicles, which upon deployment will improve the revalence of *probe sensing*.

2.4 Traffic signal controllers: capabilities and challenges

It is estimated that there are 311,000 traffic signals operating in the United States as of 2012, representing an $82.7 billion dollar public investment in the seemingly fundamental task of assigning right-of-way to vehicles and pedestrians at road intersections [150]. The control policies imposed by these signals vary widely in effectiveness. The NTOC's 2012 *National Traffic Signal Report Card* assigns an overall grade of D+ to the observed state of signal control operations, and indicates several areas where improvements could be made in management, monitoring, maintenance, and control [150].

In the following paragraphs we explain the fundamental capabilities of existing traffic signal controllers and provide insight into how control policies are currently designed.

Signal control parameters

Traditional traffic signal operations are generally governed by three control parameters: cycle length, green splits, and offset. *Cycle length* is the length of a signal cycle, or the period of time in which all phases are actuated in sequence. *Green splits* dictate the amount of time in a cycle allocated to actuation of each phase. *Offset* is a parameter governing the relative starting time of the cycles of adjacent intersections.

The simplest control policy is a fixed-time policy in which all three of these parameters are pre-determined and cannot be changed during operations. While there is no universal "standard" concerning how to choose fixed-time signal parameters, a few widely-accepted methodologies from traffic literature are often used as guidelines for manual design of signal timings.

One classical pre-timed policy is called the *Equal Degree of Saturation Policy*, also known as the *Webster Policy* [212]. It is based on minimizing average vehicle delay per phase calculated by some variation of Webster's delay formula (2.4). Recall that $y_a = q_a/s$ is the ratio of demand flow to the saturation flow for an intersection approach. In practice, y_k for a signal phase ϕk is considered to be the maximum of $\{y_a\}$ for all approaches a included in ϕk. A green division proportional to the relative y values of each cycle phase approximates the controller which minimizes d in equation (2.4). Assuming this method is used to allocate

green times within a cycle, the optimal cycle length is defined as

$$c^o = \frac{1.5L + 5}{1 - Y} \tag{2.24}$$

The procedure for setting green splits using the Webster Policy is then as follows:

1. Estimate flow and saturation flow for each approach.
2. Evaluate y (the ratio of flow to saturation flow, or q/s) for each approach. Select the y value for each phase as the maximum of that for each included approach.
3. Decide on all-red periods R (i.e. for pedestrians, turns, etc) and estimate total lost time L.
4. Calculate optimal cycle time from equation (2.24).
5. Subtract the total lost time L from the cycle time giving the available green time and divide this in the ratio of y values:

$$g_k = \frac{y_k}{Y}(c^o - L) \; \forall \;\text{ phases } k \tag{2.25}$$

6. Add l seconds to each effective green time and subtract the amber periods (3 seconds) to give the controller setting of green time.

Many of the signal timing optimization packages used by modern traffic engineers (which will be described in more detail in Section 2.5) are informed by this classical procedure. It is also still used as a heuristic for control design when software tools are not available [89].

Another possible methodology is called the *Greenshields-Poisson Method*, which uses the assumption that vehicles arrive at a queue in a Poisson distribution to derive the following optimal phase time:

$$\lambda_k C = 3.8 + 2.1 Q_k \tag{2.26}$$

where $\lambda_k C$ is the total green time allocated to ϕk and Q_k is the number of vehicles in the queue for the critical movement of ϕk, which is calculated via the measured mean arrival rate [89]. A standard cycle length is assumed, for example 60 seconds for an intersection with two critical phases or 100 seconds for an intersection with four critical phases.

Once cycle length and green splits are fixed by a procedure like (2.25) or (2.26), relative offsets of coordinated controllers are optimized to maximize the amount of time per cycle in which vehicles may travel through a predefined series of signals without encountering a red light. This is achieved by fixing the ending time of a predefined phase (or set of phases) within each signal's cycle to a specific relative time stamp on a synchronized clock. Convention dictates that the through movements of the major roadway are assigned to phases 2 and 6, and thus these are typically chosen as the *coordination phases*. This concept is called

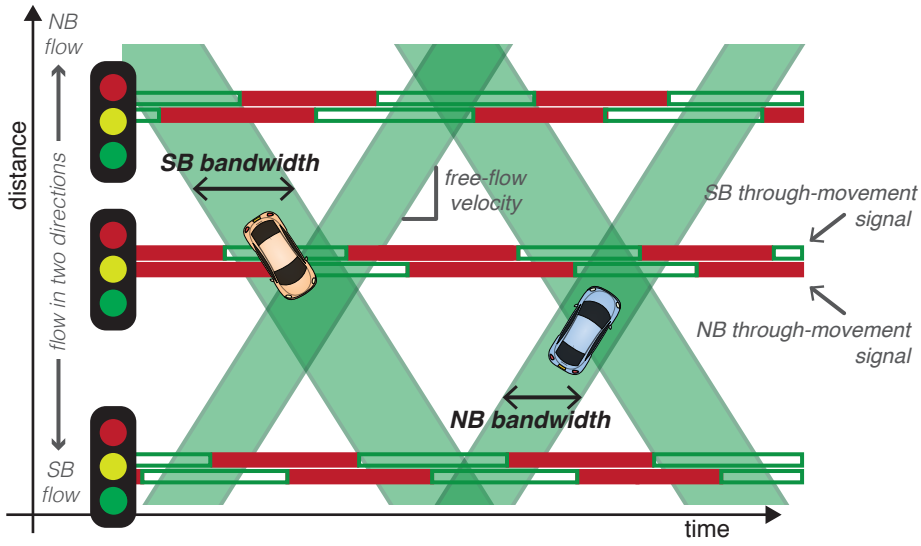

Figure 2.5: The maximization of a progression bandwidth (or green-wave) is achieved via calibration of relative signal cycle offsets to coordinate green signals in the desired movements, usually through-bound flows on a major arterial. Traffic is assumed to travel at a constant rate. Assuming that green splits are previously fixed, the calculation of bandwidth is completely independent of expected demands—it is only a function of expected velocity and the distances between successive signals. Multi-directional bandwidth maximization is still widely discussed in the literature today; traffic managers typically decide on one primary direction to prioritize and adjust timings heuristically to give a secondary direction an increased bandwidth.

bandwidth or *green-wave* maximization. A green-wave which persists over many signals is sometimes also called a *progression*.

Bandwidth optimization has been widely studied in literature. While maximizing green bandwidth in one direction is intuitive, design for simultaneous maximization of bandwidth in both directions of a bidirectional roadway, as illustrated in Figure 2.5, is not as trivial.

The absolute maximization of bi-directional bandwidth is a mixed-integer linear problem (MILP) as first formulated in [121]. Many solution methodologies for this specific problem have since been proposed [139, 122]. In practice, traffic engineers unaided by software typically either choose a single direction to prioritize or use some kind of simplifying heuristic on relative bandwidths to transform the MILP into an efficient linear program or simple geometric exercise [147, 89].

Actuated control

When signals are designated to operate in an *actuated* mode, the timings of specified green-yellow-red sequences are not predetermined. Instead, each interval is "called" on demand and extended in response to the measurements acquired by local intersection detectors, which are typically installed in some configuration similar to that portrayed in Figure 2.4. While the costs of required sensing infrastructure are not insignificant, actuated control greatly reduces delay relative to fixed-time control due to its reduction of wasted green and its capabilities to promote progression "on-the-fly" [109].

An actuated intersection controller that is unconstrained by coordination parameters does not necessarily operate on a fixed cycle length or phase ordering. Timings of phase changes depend fully on detected vehicle presence or passage events, and are constrained only by certain pre-defined maximum or minimum timing parameters. This type of operation can be designed in either a *fully-actuated* sense or a *semi-actuated* sense. In a fully-actuated setting, detection is provided on all approaches. This level of detection is typically used in locations where speed is relatively high or the roads intersecting are both major arterials. Semi-actuated systems are installed where is it clear that one road should be given priority over the other, such as in the case where a major through-way intersects with a more minor road. In this instance, the major through movements are not instrumented with detectors, rather the right-of-way is given to these approaches by default (or when conflicting demand is not detected).

A phase is served as soon as possible after a "call" is placed on that phase by the corresponding vehicle detector (limited of course by the termination of the conflicting phase(s) being served at the time of the call). The phase is then terminated in one of the following manners:

1. The phase has been green for its designed *minimum green time* and no additional vehicles have been detected to be served. Minimum green time is typically equal to the expected amount of time that it would take for a vehicle sensed by the movement detector to pass through the intersection, and can vary greatly by the geometry of the specific sensor installation and the expectations of the traffic engineer who designs the system.

2. When a call is detected on a conflicting phase, a currently green phase can only retain service for an additional amount of time equal to its fixed *maximum green time*. This parameter is typically dependent on the expected demand and the amount of service given to this phase in the previous cycle. It is designed so that any accumulated queue will completely dissipate when possible. Note that because this limitation is only enforced on the green time permitted after a competing call, total green times may (and often do) exceed their so-called maximum green parameter: the name "maximum green" is misleading because it it does not bound the green time of a given phase, but rather serves as an upper-bound on the amount of time that a vehicle detected on a competing phase will have to wait before its movement is serviced.

3. A phase may *gap out* after no activity has been detected on the relevant approach for an amount of time specified by the *gap time* parameter.

4. Where appropriately instrumented, actuated phases can also be terminated due to preemption by transit or emergency vehicles.

An example of phase termination due to a conflicting detector call is illustrated in Figure 2.6. Note that *yellow clearance* and *red clearance* times are often fixed for each phase, and depend mostly on features of intersection geometry such as size, grade, maximum speed, turn movement type, and number of lanes.

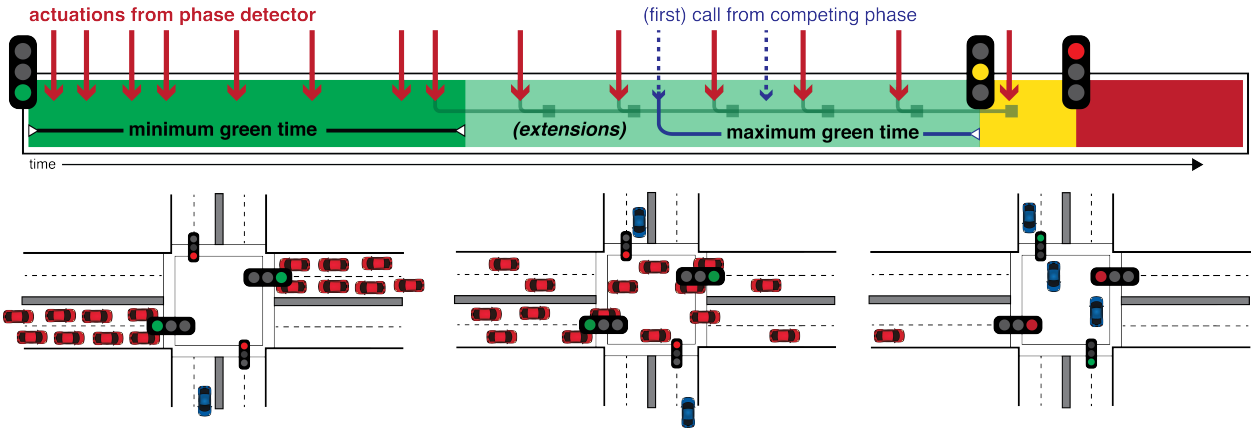

Figure 2.6: After the minimum green time, additional calls on a green phase trigger finite extensions. If a subsequent call on this phase is not detected within the time of the last extension, the phase will gap out. In the case illustrated in this diagram, a call on a competing phase was detected before the current green phase gapped out. The green phase was therefore terminated only after a period of time equal to the phase's maximum green length parameter *after the competing call*.

Semi-actuated deployments are often selected over fully-actuated systems because they are less expensive to install and maintain. The partially instrumented intersections are also appropriate for settings where coordination of subsequent signals along a major arterial is desired. *Actuated-coordinated* control is used to achieve such coordination. The set of coordinated signals are all designed with a common fixed cycle length and a synchronized clock. A traffic engineer designates an offset parameter for each controller to specify the relative timings of the major through phases to enforce the desired progression. Actuation is then only used to dictate the precise distribution of green time within the cycle.

Figure 2.7 depicts an example of an actuated-coordinated cycle. The offset parameter dictates the *yield point* of a controller, which can be considered a "local" 0-time index for the signal's cycle. It explicitly denotes the *force-off* or forced termination time for the coordinated major through phases, usually phases 2 and 6 by convention. After fixed

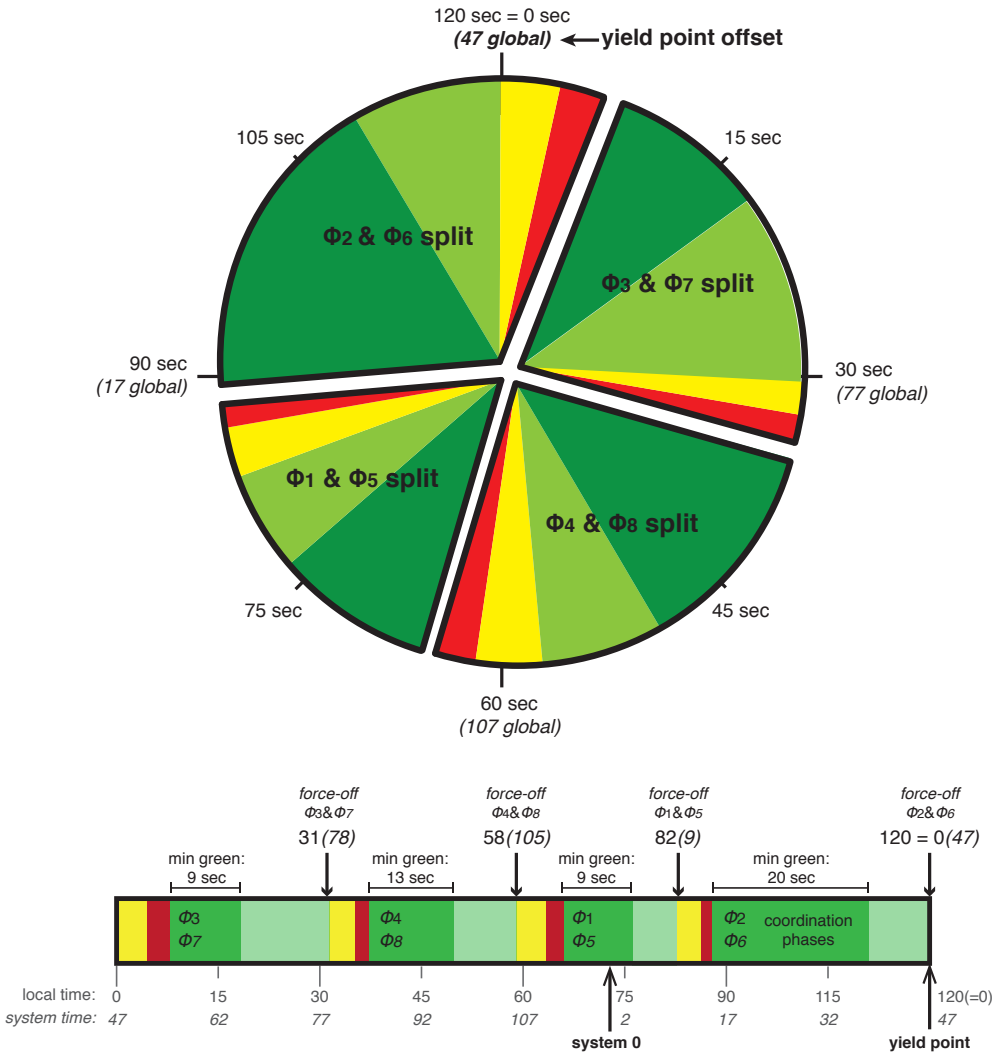

Figure 2.7: This figure illustrates an example timing plan for a actuated-coordinated controller. The yield point corresponds to the force-off of coordination phases (ϕ_2 and ϕ_6). Phase splits are defined as the period of time between the force-offs of adjacent plans, but actualized splits may vary as phases are terminated early due to competing calls or inactivity. Note that simultaneous phases from the two rings do not necessarily share the same parameters. More information can be found in [109].

yellow and red clearance periods, the following phases (3 and 7) are given the green. The termination of these phases may occur any time between their minimum green time and

their specified force-off parameter, as dictated by rules similar to those listed above for non-coordinated actuated controllers. Note that the force-off point preempts a maximum green time parameter when detection is continuous. Each subsequent set of phases follows the same pattern, with phase ordering specified by the controller's ring-barrier diagram. Some systems have the flexibility to skip minor phases such as protected left turns when no demand is present.

The rigid force-off parameters only leave certain *permissive periods* of the cycle where the green phase is not predetermined. These parameters are usually designed by a traffic engineer such that the coordination phases (i.e. 2 and 6) can receive a significant amount of the permissive time (when calls are not detected on the prior phases) to generate maximum progression bandwidth.

For further description of the capabilities of actuated controllers, see [54] or [109].

Plan-switching modes

Typical traffic signals can operate in either of two plan-switching modes:

1. **time-of-day (TOD) mode**, in which signal plans are selected to operate during pre-defined time intervals during the day, and
2. **traffic-responsive (TRPS) mode**, in which signal plans are selected based on feedback from local traffic conditions.

For either of these modes, the plans that are chosen to switch between can theoretically be any combination of fixed-time, actuated, actuated-coordinated, or free plans.

While TRPS mode can potentially provide more optimal performance than rigid TOD plan switches, this mode is highly underutilized in the United States. It has been suggested that this due to a lack of formal guidelines for robust configuration of the many necessary parameters and thresholds [3].

Hence TOD operations are currently implemented almost universally. There are a handful of existing theoretical algorithms to determine optimal TOD switching times based on predicted demand patterns [196, 210], but most often in practice these are chosen heuristically based on observed characteristics of the deployment location. When a predictable prominent peak time occurs, such as during a special event or a road closure leading to reduced capacity, normal operating schedules can be manually adjusted.

Performance metrics

The performance of traffic signals is measured by various metrics (often called Measures of Effectiveness, or MOEs), including:

- experienced travel times
- average vehicle-delay

- Level of Service (LOS), as in Table 2.1
- number of stops
- average speed
- average/maximum queue length

The LOS criteria in Table 2.1, as defined in Chapter 18 of the 2010 Highway Capacity Manual (HCM), provide an indication of the vehicle-delays expected in ideal signal operations [85]. Note that any intersection for which the degree of saturation exceeds 1 is automatically assigned a LOS of F: saturation is considered an unstable and thus undesirable state in all cases.

Table 2.1: HCM Level-of-Service (LOS) criteria [85]

	LOS, by vol-to-cap ratio	
control delay (sec/veh)	$x \leq 1.0$	$x > 1.0$
≤ 10	A	F
$10 - 20$	B	F
$20 - 35$	C	F
$35 - 55$	D	F
$55 - 80$	E	F
> 80	F	F

The HCM2010 also suggests standard methods of calculating vehicle-delay and other MOEs using measurable volumes and knowledge of intended signal timings. The most recent HCM formula for delay at an individual intersection movement consists of three terms:

- *uniform delay* (d_1), which is derived from Webster's delay formula 2.4 and estimates the delay caused by the signal assuming uniform queue arrivals;
- *incremental delay* (d_2), which theoretically accounts for non-uniform arrivals and occasional random instances of temporary saturation (cycle failure) or sustained period of oversaturation; and
- *initial queue delay* (d_3), which accounts for any initial or queue present before the analysis period (and is equal to 0 if no queues are present on the approach).

$$d = d_1 + d_2 + d_3, \quad \text{where}$$

$$d_1 = \frac{0.5C\left(1 - \frac{g}{C}\right)^2}{1 - \left[\min(1, x)\frac{g}{C}\right]} \tag{2.27}$$

$$d_2 = 900T\left[(x-1) + \sqrt{(x-1)^2 + \frac{8kIx}{cT}}\right] \tag{2.28}$$

$$d_3 = \frac{3,600}{qT}\left(t\frac{Q_b + Q_e + Q_{eo}}{2} + \frac{Q_e^2 - Q_{eo}^2}{2c} - \frac{Q_b^2}{2c}\right) \tag{2.29}$$

where

- C is the cycle length,
- x is the volume-to-capacity ratio (degree of saturation) of the movement,
- q is the flow on the movement,
- $c = \frac{\lambda s}{C}$ is the *capacity* of the movement,
- T is the length of the period of analysis over which all other variables are measured (typically 15 minutes or 1 hour),
- t is the amount of time during which unmet demand is observed during time period T,
- k is a *incremental delay calibration factor*, which is a function of x and the green extension parameter of an actuated signal that accounts for the effect of controller type on delay,
- I is an *upstream filtering or metering adjustment factor* to account for the effect of platooned arrivals from coordination with upstream intersections,
- Q_b is the initial movement queue at the beginning of the analysis period,
- $Q_e = Q_b + t(q - c)$ is the number of vehicles present in any queue remaining at the end of the analysis period, and
- Q_{eo} is the number of vehicles present in any queue remaining at the end of the analysis period when $q \geq c$ and $Q_b = 0$: if $q \geq c$, then $Q_{eo} = T(q - c)$ and $t = T$; if $q < c$, then $Q_{eo} = 0$ and $t = \frac{Q_b}{(c-q)} \leq T$.

2.5 Existing signal timing methodologies

The Federal Highway Administration believes that the existing delay at many signals in US cities could be reduced significantly by adjusting or updating timing plans [109]. Indeed, it has been estimated that delays at traffic signals contribute 5-10% of all traffic delay, or

295 million vehicle-hours of delay, just on major roadways alone—and that improving traffic signal management could reduce this delay with a benefit-cost ratio exceeding 40:1 [42]. However, the process of timing (or re-timing) traffic signals is typically undertaken with minimal formal guidelines [187].

The first step in calculating appropriate timings for a traffic signal system is the collection of volume data to predict demands at the relevant intersections. As previously mentioned, existing sensors do a poor job at providing an accurate estimate of turning volume counts or queue lengths. Therefore, technical consultants are typically hired to perform lengthy data collection procedures at each intersection, at a cost of well over $1,000 per intersection [187]. While these studies may show in detail how demands vary significantly throughout the day, practitioners often only make use of one or two highly averaged measurements of morning and evening "peak" volumes. This is because plans are initially designed using out-of-the-box timing optimization software packages that only require these static peak volumes as inputs [187]. Some commonly used packages will be described in the following subsection. Finally, technicians will make heuristic adjustments to the modeled timings when translating them into the necessary control parameters and encoding these parameters onto signal firmware. Such an adjustment process may lead to improved performance, but it is typically very dependent on the experience and expertise of the specific technician who may make decisions with very little oversight [187].

Overall, automating the signal timing process in a verifiable manner would save a great deal of resources and likely significantly improve the performance of arterial traffic networks. A very small number of municipalities have adopted small-scale deployments of advanced *adaptive traffic control systems*, which will also be discussed at the end of this section. A less radical (and less expensive) solution would be to facilitate the implementation of existing *traffic responsive* control functionalities, as we propose in Chapter 6.

Off-line plan design tools

Off-line *bandwidth maximization* packages are "one-shot" design tools that suggest optimal green splits and corresponding progression offsets, or relative cycle-start timings for a continuous series of consecutive traffic signals that maximizes the amount of time in which flow is permitted to travel uninterrupted along the roadway without signal impediment. This concept is explained in Section 2.4. Some commonly used bandwidth maximization tools are listed below.

- **PASSER II (Progression Analysis and Signal System Evaluation Routine)** was developed by the Texas Transportation Institute in 1974 (and has been updated since then). It can select phase sequences and splits, cycle lengths, and offsets for a network of up to 20 intersections. This is often considered the best bandwidth-based tool and continues to be maintained, updated and improved.
- **MAXBAND** suggests cycle lengths, offsets, speeds, and phase sequences to maximize a weighted sum of relevant bandwidths [122]. PASSER IV implements this program,

and can handle a maximum of 20 arteries and 25 intersections [187]. It may have slight performance improvements over PASSER II, but its higher computational requirements have prevented similarly widespread adoption.

- **TSPP/Draft** is a time-space and platoon progression diagram tool which provides visual tools for planners to heuristically choose appropriate signal timings [187].

- **TSDWIN** is another graphical tool for manually "fine-tuning" signal timing plans on a single arterial or small contiguous network [187].

Off-line *model-based optimization* tools take a more comprehensive approach to network-level optimization of many performance criteria including progression but also average delay, number of stops, or HCM LOS. Below we describe the most popular model-based tools.

- **TRANSYT (TRAffic Network StudY Tool)** is an off-line software package which determines the optimal fixed-time traffic signal settings to minimize some chosen balance between total delay and number of stops on the network. It originally used some kind of gradient descent algorithm to reach optimal signal offsets and green splits on networks up to 50 intersections [176]. A newer version (TRANSYT 7-F) uses a genetic algorithm to optimize cycle length, phasing sequence, splits, and offsets at either an arterial or network level [208]. It runs a detailed macroscopic model which simulates platoon dispersion, horizontal queues, and fully-actuated intersection control on larger networks. It can optimize a performance index that is a user-defined combination of delay and stops, fuel consumption, queue length, operating costs, or progression opportunities. TRANSYT 7-F is distributed in the Highway Capacity Software 2010 (HCS2010) package that was developed by the Federal Highway Administration to implement the procedures of the HCM2010 on urban streets and signalized intersections as well as on freeway elements. It features limited modeling of fully-actuated controllers (Synchro is better for this). But it is considered the most flexible and comprehensive modeling package available today.

- **Synchro** is the most widely deployed signal optimization package (as of 2003), and is considered the current "state of the art" [187]. It can be used for generating off-line timing plans for either isolated intersections or coordinated networks. This comprehensive software package can estimate the effects of a fully-actuated NEMA controller, including force-offs and permissive movements, and can even suggest which controllers should be coordinated and which should be actuated independently. It provides an analysis of average approach delay, intersection delay, volume-to-capacity ratio, LOS, 50^{th} and 95^{th} percentile queue lengths, total stops, travel time, emissions, and fuel consumption for each intersection in a network.

- The **Signal Operations Analysis Package (SOAP)** is a macroscopic optimization tool that suggests control plans for individual isolated intersections based on inputs of demands, truck/bus usage, left turn data, saturation flow, and signal constraints. It

can only optimize a single performance criteria, but it handles up to 48 time periods with various input data.

- **PASSER III (Progression Analysis and Signal System Evaluation Routine)** can minimize intersection delays for an isolated intersection or maximize bandwidth of a small, one-way linear network.

A 2000 survey of members of the Institute of Traffic Engineers (ITE) in the United States suggests that Synchro is the most commonly used signal optimization tool (54%), followed by TRANSYT-7F (25%) and Passer II (23%). Other significant responses included Passer III (9%), SOAP (4%), and Passer IV (4%) [187].

Adaptive traffic control systems

While conventional signal control algorithms run with pre-programmed timing parameters (i.e. a fixed cycle length, green splits, and offsets), attention has recently been turned to *adaptive traffic control systems* (ATCS). These technologies often break the traditional notions of cyclical signal operations, and instead allow for more flexible "on the fly" signal switching in response to local sensor feedback.

While they have demonstrated great capability to enhance flow continuity and reduce delay/fuel consumption, ATCSs have not been widely adopted in practice in the United States. In 2010 there were only about 25 ATCS deployments country-wide [199]. This hesitance to adopt ATCS is partly because of unfamiliarity with the new "black-box" systems and uncertainty over their benefits and reliability. But it is also largely due to the perceived costs of the specialized hardware and infrastructure required for their operation. ATCSs can range in cost from $20,000 to $128,000 per intersection and and average of 41 hours of training time per person to maintain operation [54]. ATCS proponents, however, argue that because adaptive algorithms continuously adjust to observed conditions, their infrastructure investments are recovered by preventing the need for lengthy and expensive signal re-timing processes. Cost-to-benefit studies have not yet shown conclusive long-term results on existing deployments [216].

Many algorithms for adaptive signal control have been proposed with various motivations and approaches. While none have gained entirely universal acceptance, a list and description of the systems deployed in the United States is available in Appendix A.

2.6 Data available to study arterial traffic

One of the barriers to improving current arterial management practice is the lack of data available to study the realistic dynamics that exist on signalized roadways. Here we review the data sources that do exist, and explain how they have been used in the work described in this dissertation.

Sensor data

Our experience attempting to acquire data from sensors at signalized intersections reinforces the conclusions we introduced in Section 2.3 regarding the collection of arterial detector data: while small subsets of historical sensor data is occasionally available from local transportation authorities, a persistent and complete observation of a large network is rarely existent.

In Chapter 6 of this dissertation, we make use of volume detector data from existing detectors at a real intersection on Huntington Boulevard, a major arterial the I-210 corridor portrayed in Figure 1.1. This data was made available to us directly by the relevant traffic authority, but was not easily accessed. Because the data was not permanently archived by the local traffic authority, we were required to manually download and store each week's worth of data before it's bi-weekly deletion from the local database.

System documentation revealed that each installed video-based detector was capable of returning count (volume) and occupancy data at five-minute time aggregations. Yet upon analysis of the downloaded data, it was found that only about half of the sensors present on a typical intersection were designated to transmit to the central server. Not all signal movements at an intersection were represented by this data, and thus an accurate estimation of turn ratios was not possible. Some intersections did not even record a minimum of one sensor on each approach. Furthermore, there was no reliable indicator of the functionality of the sensors being recorded; faulty or missing data was extremely common.

It is important to note that there are active efforts to address this lack of data within the traffic community. For example, the recently prototyped SMART-SIGNAL system has been designed to interface with existing actuated signal sensors and cabinet hardware to extract and archive event-based detector data for operational performance monitoring purposes [125].

NGSIM trajectory data

The most complete set of arterial field data known to this author is the vehicle trajectory data acquired by the US Federal Highway Administration's (FHWA) Next Generation SIMulation (NGSIM) program [157]. In an effort to advance algorithms for modeling microscopic-level driver behaviors, NGSIM researchers mounted high-definition video cameras atop a building neighboring a major arterial roadway, Lankershim Boulevard in the Universal City neighborhood of Los Angeles, California (shown in Figure 2.8). Recorded video was processed to transcribe the detailed trajectories of each individual vehicle that traveled along 1,600 feet (5 blocks) of Lankersim Blvd for approximately 30 minutes during a morning peak period (8:28 am to 9:02 am on June 16, 2005).

In total, 2,442 vehicles were detected and tracked during the observation period. Trajectories were recorded at a time resolution of 10 samples per second, and location points are believed to be accurate within a four foot radius. A visual inspection of the data in the form of an animated compilation of vehicle trajectories, available online at `http://youtu.be/jJen2ybNr34`, reveals that flows were typically constrained only by signal con-

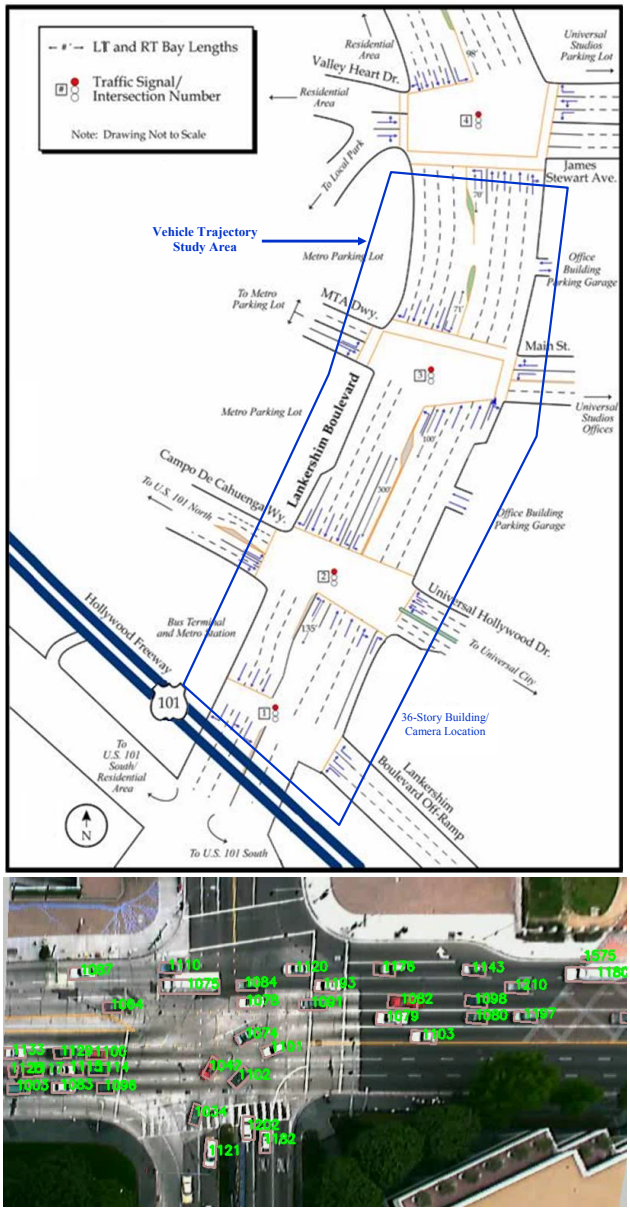

Figure 2.8: High-resolution vehicle trajectories are available for 5 blocks of Lankershim Blvd, capturing queueing behaviors at 3 internal intersections. This diagram and video screenshot was provided as part of the dataset documentation, which is available online at `http://ngsim-community.org/`. [158]

trollers: vehicles did not suffer severe delays due to pedestrians or other uncontrolled obstacles.

The NGSIM data gathering techniques were replicated for an arterial site on Peachtree Street in Atlanta, Georgia—although data from this site is limited to 15 minutes of con-

tinuous data as opposed to 30. These datasets, which are available online at `http://ngsim-community.org/`, have proven to be highly valuable within the traffic community. Not only have they been used for their original purpose of developing microscopic driver behavioral models, but they have also been widely used to simulate and validate macroscopic algorithms such as queue estimation [114], travel time estimation [95, 172], control [110], and even models of the performance of connected vehicles [14]. In this dissertation, we use the Lankershim data set for validation of macroscopic modeling techniques (Chapter 3) and arterial state estimation procedures (Chapter 4).

GPS trajectory data

With the spread of cellular telephones equipped with highly-accurate Global Positioning System (GPS) sensors, it has been increasingly common for various entities to aggregate and make use of this data for monitoring traffic conditions [91, 209]. On urban arterial roadways specifically, GPS data has been used for measuring travel times [87, 25, 26, 27], estimating queue lengths [41, 16, 40], and re-constructing traffic signal patterns [96]. Accurate analysis of cell phone data requires significant pre-processing (such as map-matching and filtering data from parked vehicles) before such estimation algorithms can be achieved. Research on techniques to perform the required pre-processing is ongoing amongst computer scientists and traffic engineers alike [213, 92, 137, 101].

In Section 4.3 of this dissertation, we suggest a method of integrating sampled GPS trajectories into an estimate of arterial link state.

Microsimulation data

Largely due to lack of a more realistic alternative, researchers have largely relied on dynamical data from *microscopic simulation* models (referred to here as *microsim*) for validation of estimation or control algorithms on signalized roadways. Microsim models use detailed representations of individual vehicle dynamics to numerically reconstruct the aggregate behaviors and traffic metrics that are ultimately important to modelers. These equations typically involve a large number of tunable parameters or stochastic random variables to govern the many sub-models of vehicle behaviors such as acceleration/spacing, lane-changing, platoon dispersion, and routing, amongst others. A review of some of the mathematical models commonly used in microsimulation packages is provided in [33].

Popular microsim models include Aimsun [204], CORSIM [43], MATSim [134], VISSIM [65], and Paramics [164]. We use vehicle delay metrics calculated in Aimsun to validate the effectiveness of a novel signal control algorithm in Chapter 5.

The enormous effort required to build and calibrate microsim models is the main barrier to using this type of data. The complicated interactions of these sub-models require a massive amount of parameter tuning to achieve results that correspond to realistic observations with high probability. A study of recent model-building projects revealed that a "small" model of 18 miles of freeway (154 nodes and 174 links) required 540 hours of design and calibration

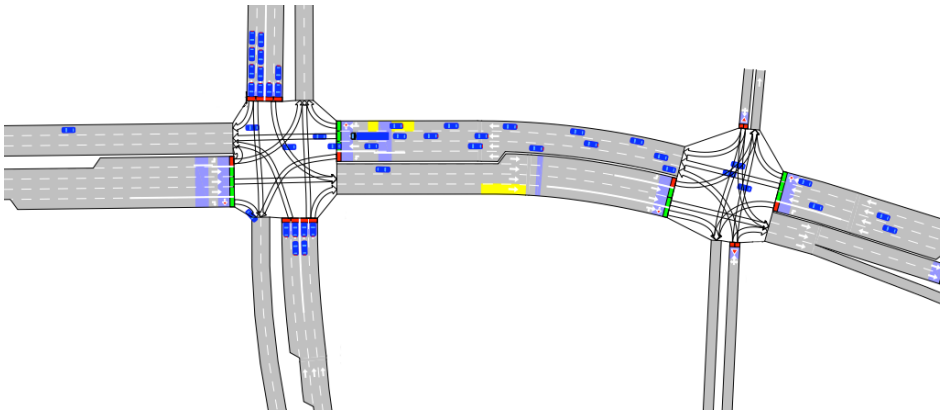

Figure 2.9: A microsimulation like Aimsun (from which this screenshot was generated) generates models of individual vehicles. Macroscopic flows are not explicitly represented, however one can analyze aggregate traffic dynamics after the simulation is complete.

work, while a "large" model on the scale of a 30-mile corridor (659 nodes and 796 links) required 10,080 hours. This effort translates into a significant economic investment for a typical traffic management agency [8].

Chapter 3

Validation of numerical queuing models for signalized traffic networks

Both researchers and traffic managers typically rely on microscopic simulation for describing arterial state dynamics. This is undesirable for two major reasons: first, it requires significant investment in time and expertise for model development and calibration, and second, it fails to provide an analytical expression for the macroscopic behaviors that an operator would desire to analyze and control.

Microsimulation is difficult to use in an ICM project in particular because it cannot be integrated with macroscopic freeway models such as CTM. This deficiency provides motivation to develop an efficient time-discretized model of arterial traffic networks which avoids the pitfalls of CTM, but provides a superior representation of signal-constrained vehicle queueing dynamics than that of existing vertical queueing models. Because such a model could be interfaced with a CTM freeway implementation, it would achieve a crucially-needed representation of the boundaries between mainline and arterial networks where congestion would likely aggregate due to oversaturated freeway approaches or excessive use of freeway ramp meters.

In the following sections we contribute developments towards the implementation of such a model.

3.1 A cell-based modeling framework for signalized traffic networks

Consider a set of short urban roads separated by signalized intersections. As previously described, we model this network as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ where the set of links $\mathcal{L}$ correspond unidirectional roadways and the set of nodes $\mathcal{N}$ represent intersections or general points of flow division.

More specifically, each individual link $l \in \mathcal{L}$ represents a unidirectional path between two nodes in the network with a physical capacity for vehicles to be stored. Define $In(l)$ to be

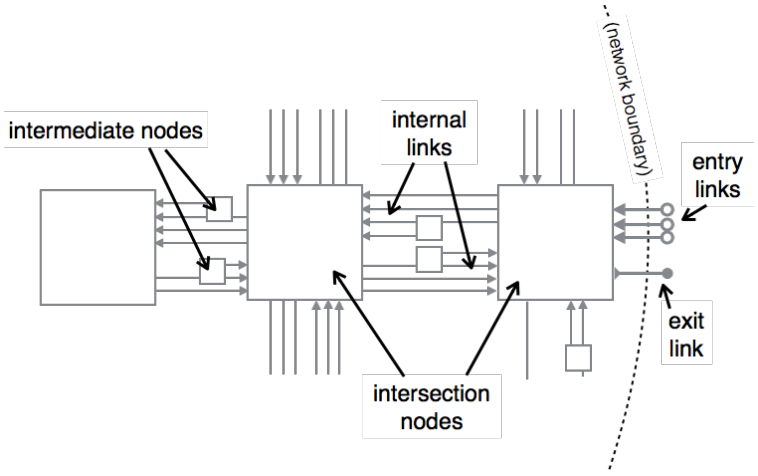

Figure 3.1: A road network is represented by a graph containing internal links, entry links, exit links, intersection nodes, and intermediate nodes.

the set of all links immediately upstream of l in the network such that they can pass flow directly into l (through the node between them), and define $Out(l)$ as the set of all links which receive a non-trivial flow directly from l.

Physical roads are typically divided into a set of *movements* corresponding to each immediate downstream destination. A movement contains to all vehicular flow on a single roadway that intends to subsequently enter the same downstream roadway at the next intersection. In this work, we consider each movement to be represented as a distinct link that is parallel to the links representing all other movements on the same roadway. These parallel links can span the entire block from upstream intersection to downstream intersection, or they can originate somewhere in between two subsequent intersections at a position corresponding to the location at which a road forks into a turn pocket. This mid-link split allows for modeling of shared lanes or turn bays which can result in partial output blocking for one or more movements sharing common upstream resources.

The set of all links $\mathcal{L}$ is divided into three subsets: entry links $\mathcal{L}_{\text{entry}}$, exit links $\mathcal{L}_{\text{exit}}$, and internal links $\mathcal{L}_{\text{int}}$. Internal links must be bounded on either side by a network node connecting them to a non-trivial set of neighboring network links. They each have finite length L_l with corresponding finite storage capacity. Therefore, the inflows and outflows of an internal link are inherently constrained by the number of vehicles on the link at any given time. Entry links originate from outside of the network and terminate at an internal network node. Because $In(l) = \emptyset \quad \forall\ l \in \mathcal{L}_{\text{entry}}$, demand on these links is exogenous to the network – entry links are in fact the only point of entry for external demand. For the purposes of this work, consider all time-dependent exogenous demands to be known. While there is a rate limitation on the flow exiting entry links, there is no bound on link storage; any expected demand can be stored on these links indefinitely until service is available. In the same manner, exit links serve as an infinite repository for flow exiting the network. These

links have no departing flow, and by definition $Out(l) = \emptyset \ \forall \ l \in \mathcal{L}_{\text{exit}}$. Yet due to an infinite storage capacity, the presence of congestion on exit links will never limit overall network outflow.

Nodes are storage-less "gateways" that govern flow between neighboring links. A node $n \in \mathcal{N}$ is defined geometrically by its set of incoming links I_n and its set of outgoing links O_n. It is parameterized by a split ratio matrix β_n of dimension $|I_n| \times |O_n|$ that defines the proportion of vehicles in each incoming link that are waiting to enter each outgoing link $m \in Out(l)$. Elements $\beta_n^{l,m}$ of feasible split matrices must obey the following characteristics:

$$0 \leq \beta_n^{l,m} \leq 1 \ \forall \ l \in I_n, \ m \in O_n, \ n \in \mathcal{N} \tag{3.1}$$

$$\sum_{m \in O_n} \beta_n^{l,m} = 1 \ \forall \ l \in I_n, \ n \in \mathcal{N} \tag{3.2}$$

Each node in the network represents either a simple splitting of a single upstream link into many "movement" links or a point of flow exchange between two or more intersecting roadways. The topological differences between *intermediate* nodes and *intersection* nodes are illustrated in Figure 3.1. The operation of the intermediate nodes is straightforward: a single demand flow is consistently divided into many downstream supply flows as specified by a fixed split ratio parameter. Intersection nodes similarly split flows according to dictated split ratios, but furthermore must resolve conflicts between multiple input flows which require the use of shared physical resources.

We therefore define a *phase* as a set of incoming links which can flow simultaneously through the node without causing resource conflicts. Each phase ψ for a node n is encoded as a sparse binary matrix of dimension $|I_n| \times |O_n|$, where element $\psi_{l,m} = 1$ if link l is permitted to flow into link m as part of that phase (and otherwise 0). The set of all possible feasible phases for node n is denoted Ψ_n.

A flow-impeding signal controller is placed on each intersection node to ensure safe operation of the modeled junction by restricting concurrent flows across the node to those input links encoded in an element of $\mathbf{G}_n \subset \Psi_n$. Note that $\mathbf{G}_n$ is generally limited to some subset of Ψ_n because practical signal controllers typically only actuate a limited number of phases due to hardware limitations or safety regulations. A controller on node n must alternate between actions $G \in \mathbf{G}_n$ at an update rate dictated by management constraints or objectives.

Assume for now that all turn directions of a "shared movement" link are actuated simultaneously: if a node's input link l is permitted to flow into any one of its downstream neighboring links $m \in Out(l)$, it is also permitted to flow into any of its other downstream neighbors. Then define $G_l \in \{0, 1\}$ to be the indicator that link l is permitted to discharge according to its fixed expected split ratios.

Because links are defined to have finite storage capacity, nodes must also enforce physical limitations in flow due to congestion on their neighboring links. Flow through a node is therefore furthermore limited by two additional factors: the *sending constraints* of its upstream links, and the *receiving constraints* of its downstream links. Intuitively, the set $\{S_l(t), \ l \in I_n\}$ of upstream sending constraints imposed on node n considers the number of

vehicles currently available to be serviced by the node on each of it's incoming links. The relevant downstream receiving constraints $\{R_m(t),\ m \in O_n\}$ are limitations in the service rate of the node due to lack of space downstream to receive the transmitted flow.

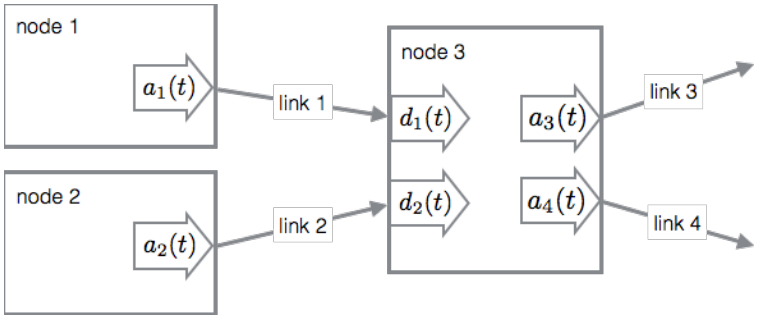

Figure 3.2: A node enforces downstream space limitations on flows departing upstream links $\big(d_l(t),\ l \in \{1,2\}\big)$ and distributes flows arriving into downstream links $\big(a_l(t),\ l \in \{3,4\}\big)$ according to pre-defined split ratios.

The specific forms of $S(t)$ and $R(t)$ will vary according to the link dynamics being modeled. But in terms of these generalized constraints, the flow departing each upstream link $l \in I_n$ can be defined as follows:

$$d_l(t) = G_l(t) \min\left\{S_l(t),\ \min_{z \in Out(l)}\left\{\frac{1}{\beta_n^{l,z}} R_z(t)\right\}\right\} \tag{3.3}$$

where n is the terminal node of link l. Notice that this is designed to enforce that vehicles follow a *first in, first out* (FIFO) principle: queue discharge is limited by the *most restrictive* downstream demand function so that downstream queue capacities are not exceeded while discharge remains consistent with the specified static split ratios. The flow arriving into downstream queues $m \in O_n$ must then balance this departing flow:

$$a_m(t) = \sum_{k \in I_n} \beta_n^{k,m} d_k(t) \tag{3.4}$$

3.2 The cell transmission model

As introduced in Section 2.2, the cell transmission model (CTM) is a stable numerical approximation of the LWR conservation model of traffic flow (2.20). While it was first derived independently by Daganzo in [48, 49], it was later shown by Lebacque [115] to be exactly equivalent to a Godunov difference scheme of the LWR PDE.

CTM specifically defines a piecewise-linear relationship between flow q and density ρ, as illustrated in Figure 3.3. For a modeled link l, this *fundamental diagram* relationship is

parameterized by free-flow velocity $\mathbf{V}_l$, maximum (capacity) flow c_l, queue dissipation speed $\mathbf{W}_l$, and maximum (jam) density $\rho_l^{\max}$:

$$q_l(\rho_l(x,t)) = \min\left\{\mathbf{V}_l\rho(x,t), c_l, \mathbf{W}\left(\rho_l^{\max} - \rho_l(x,t)\right)\right\} \tag{3.5}$$

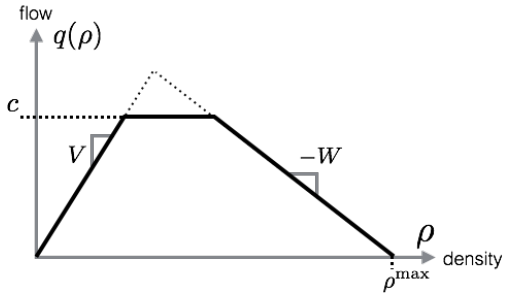

Figure 3.3: The discretization scheme of CTM enforces a piecewise-linear relationship between the spatial density of vehicles on a link and the flow of vehicles through the link. A triangular variety in which $\rho^{\max} = c\left(\frac{1}{\mathbf{V}} + \frac{1}{\mathbf{W}}\right)$ is also widely employed.

Consider a temporal discretization with uniform time steps of length that is small relative to typical actuation times for a signal control phases in the network being modeled (on the order of 1-5 seconds). In CTM, a road is spatially discretized into a series of homogenous *cells* of uniform length which is equal to the distance traveled by free-flowing traffic in a single model time interval Δt. Each network link l is therefore divided into a series of $\tau_l = \lfloor \frac{L_l}{\mathbf{V}_l \cdot \Delta t} \rfloor$ cells. An illustration of this cell division is presented in Figure 3.4.

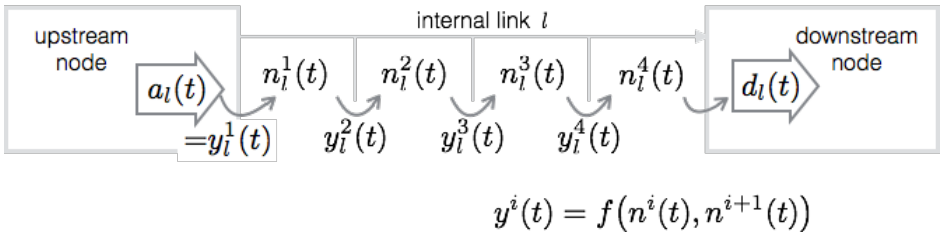

Figure 3.4: At each CTM time step, cell state $n^i(t)$ is increased by inflow $y^i(t)$ and decreased by outflow $y^{i+1}(t)$. The state update equation for this cell is therefore a function of both the state of cell i and the state of the downstream cell $i+1$. A link's receiving constraint $R_l(t)$ is a function of only the first cell's state $n_l^1(t)$, and its sending constraint $S_l(t)$ is a function of only the last cell's state $n_l^{\tau_l}(t)$.

Also define the following *cell-normalized* fundamental diagram parameters:

$$\tilde{\mathbf{V}}_l = \mathbf{V}_l \cdot \frac{\Delta t \tau_l}{L_l}$$
$$\tilde{\mathbf{W}}_l = \mathbf{W}_l \cdot \frac{\Delta t \tau_l}{L_l}$$
$$\tilde{c}_l = c_l \Delta t$$
$$\tilde{N}_l = \rho_l^{\max} \frac{L_l}{\tau_l}$$

The state of each network link l can then be represented by a vector n_l with elements $\{n_l^i,\ i = 1, \ldots, \tau_l\}$ representing the vehicle-count in each of these sequential cells. These scaled density states evolve according to the vehicle conservation relationship

$$n_l^i(t+1) = n_l^i(t) + y_l^i(t) - y_l^{i+1}(t) \tag{3.6}$$

where $y^i(t)$ represents the number of vehicles entering cell i during each time interval, defined as follows:

$$y^i(t) = \begin{cases} a_l(t), & i = 1 \\ \min\left\{\tilde{\mathbf{V}}_l n^{i-1}(t),\ c_l,\ \tilde{\mathbf{W}}_l\big[\tilde{N}_l - n_l^i(t)\big]\right\}\nu_l \cdot \Delta t, & i = 2, \ldots, \tau_l \\ d_l(t), & i = \tau_l + 1 \end{cases} \tag{3.7}$$

with ν_l equal to the number of lanes on link l. The receiving constraint put on the upstream node of a CTM link is a function of the state of only the first cell in a link, $n_l^1(t)$:

$$R_l(t) = \nu_l \min\left\{c_l,\ \tilde{\mathbf{W}}_l\big[\tilde{N}_l - n_l^1(t)\big]\right\}(\Delta t) \tag{3.8}$$

Similarly, the sending constraint put on a downstream node is a function of only the last cell's state $n_l^\tau(t)$:

$$S_l(t) = \nu_l \min\left\{c_l,\ \tilde{\mathbf{V}}_l n_l^\tau(t)\right\}(\Delta t) \tag{3.9}$$

3.3 The vertical cell model

Concern over the impracticality for large-scale implementation of CTM has spawned a second look at using less detailed vertical queueing dynamics to analyze global flow patterns and design model-based control [105, 53, 161].

Bolstered by similar research in the field of communications and mechanical service networks, modern vertical queueing models have introduced a variety of transit delay and congestion propagation improvements which have made them more widely applicable than those

of the 1960's and 1970's. Recent work yields promising results for practical network-wide control algorithms based on vertical queueing models [215, 57, 22, 118, 5, 218, 205].

The vertical cell model (VCM) is a new approach to a vertical queueing model which largely approximates the behavior of a point-queue model in the same cell-based context as CTM. Unlike most vertical queueing model formulations, it maintains a well-defined representation of both link travel delay and finite queue capacity. In this way, VCM may be classified as what some call a *spatial queueing model*.

Upon entering a VCM link, flows propagate at each step (without constraint) through a sequence of $(\tau_l - 1)$ *transit cells*, after which they enter a terminal *queueing cell*. Therefore $\tau_l = \lfloor \frac{L_l}{\mathbf{v}_l \cdot \Delta t} \rfloor$ state variables are required to represent the state of each link l in VCM (as in CTM):

- $v_l^i(t),\ i = 1, \ldots, (\tau_l - 1)$ are non-negative values representing the amount of vehicle-flow that has entered the link at time $t - i$, but has not yet traveled the length of the link to become eligible to exit, and
- $q_l(t)$ represents the amount of vehicle-flow which has traversed the entire link length and is therefore queued to immediately exit the link.

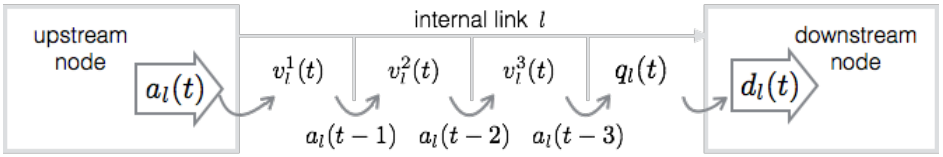

Figure 3.5: VCM passes link input flows a constant rate between non-physical "transit cells" until they reach a vertical "exit queue". To enforce finite queue storage capacity, link receiving constants are a function of the number of vehicles in all "cells" of the network. Sending constraints, however, only depend on the number of vehicles that are explicitly in the exit queue.

Explicitly, the transit queue states $v_l^i(t)$ and exit queue state $q_l(t)$ evolve as follows:

$$\begin{bmatrix} v_l^1(t+1) \\ v_l^2(t+1) \\ v_l^3(t+1) \\ \vdots \\ v_l^{\tau_l-2}(t+1) \\ v_l^{\tau_l-1}(t+1) \\ q_l(t+1) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 1 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} v_l^1(t) \\ v_l^2(t) \\ v_l^3(t) \\ \vdots \\ v_l^{\tau_l-2}(t) \\ v_l^{\tau_l-1}(t) \\ q_l(t) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} a_l(t) \\ d_l(t) \end{bmatrix} \tag{3.10}$$

This progression of vehicles across a VCM link is also illustrated in Figure 3.5.

Each VCM link is parameterized by a free flow travel velocity $\mathbf{V}_l$, a maximum flow rate (capacity) c_l, and a fixed queue capacity (per lane) κ_l. Notably, however, VCM links do not enforce a fundamental relationship between spatial density and flow rate. Unlike the independent cell supply limitation of CTM, receiving constraints in VCM are a function of the state of all link cells:

$$R_l(t) = \nu_l \min\Big\{c_l \cdot \Delta t,\ \kappa_l - q_l(t) - \sum_{i=1\ldots(\tau_l-1)} v_l^i(t)\Big\} \tag{3.11}$$

VCM sending constraints, however, depend only on vehicles that are explicitly in the exit queue:

$$S_l(t) = \nu_l \min\big\{c_l \cdot \Delta t,\ q_l(t)\big\} \tag{3.12}$$

3.4 Physical interpretation of the differences in vertical and horizontal cell models

The independent storage capacity constraints of cells within a CTM link propagates the spatial location of high-density flows within a road link. This characterizes a *horizontal* queueing model. Such spatial differentiation is a theoretically desirable characteristic, especially for representing freeways or major highways with long spans where flow is not artificially interrupted. While CTM should be able to propagate the effects of the theoretical "stop and go" shockwaves created by frequent stopping due to signal controllers, accurate representation of these types of behaviors would necessitate extremely high spatial resolution and therefore high temporal discretization in a CTM implementation (recall that $\tau_l = \lfloor \frac{L_l}{\mathbf{V}_l \cdot \Delta t} \rfloor$).

Meanwhile, network cells on an arterial road corresponding to a model time step of 1-5 seconds typically represent very small portions of roadway where the relationship between flow and density is not always as well-defined. As can be seen in a visualization of observed vehicle trajectories on an urban network (available online at `http://youtu.be/jJen2ybNr34`), queue aggregation and dissipation behaviors can vary significantly from link to link. Hence tuning the required fundamental diagram parameters to precisely represent observed conditions is a challenge—especially the backwards shockwave speed $\mathbf{W}$, which is even difficult to measure will full knowledge of system state. Furthermore, queues do not always dissipate as would be predicted by a rigid fundamental diagram relationship. This is perhaps because un-modeled factors such as varying driver response and vehicle acceleration times are dominant in practical urban traffic dynamics.

Unlike the horizontal queue of CTM, VCM models a *vertical queue* or "stack" which is not assigned to any physical distance along the link. It can be interpreted as modeling vehicles that can only either travel at maximum velocity ($\mathbf{V}$) or be completely stopped. These vehicles traverse the entire distance of a link, unconstrained by downstream congestion, before stopping in a queue. Hence flow across a VCM link is less constrained by the presence of congestion. Furthermore, queue dissipation is not limited by shockwave speed $\mathbf{W}$ in VCM

as it is in CTM. Instead, link flow capacity is the only constraint on queue discharge. It is therefore expected that link queues may empty more quickly in VCM than in CTM.

Another effect of the vertical queueing approach is the difference in upstream receiving (or supply) constraints. A downstream departure will immediately yield effects on link receiving constraints in VCM, while the effect of departures will take at least τ time steps to impact receiving constraints in CTM. The physical representativeness of this behavior remains to be examined, but it will depend on the synchronization of a link's service periods to demand patterns caused by upstream signals.

These two simplifications of VCM (as compared to CTM) result in the linear model for link dynamics shown in equation (3.10). In fact, the only non-linearities in VCM are contained within the network node model. The linear link model yields potential benefits for the derivation of link-state estimation and model-based control procedures.

3.5 Validation and comparison of model implementations

We built an experimental network to simulate the flows on five blocks of Lankershim Blvd. in Los Angeles, California, USA. The selection of this simulation area was motivated by the availability of a set of high resolution ground-truth vehicle trajectory data that was collected at this location by the *Next Generation Simulation Community* (NGSIM). See section 2.6 for a detailed description of this data set.

To specifically evaluate both VCM and CTM, we compared the output densities and flows output by each model to the observed density and flow patterns generated by aggregating the positions documented in the tracked trajectories over time and space. Both models were implemented using the *Berkeley Advanced Traffic Simulation* (BeATS) platform with a model discretization of $\Delta t = 1$ second. They shared a common graphical network (shown in Figure 3.6) and were initialized with the same geometric information, input flows, split ratios, and signal timings.

Geometry data such as link length and lane count was compiled from a satellite map image and various NGSIM documentation. The network graph used in this procedure is illustrated in Figure 3.6. Incoming boundary flows were collected by aggregating the initial appearances of tracked vehicles at each entry link with a five second resolution. Split ratios were considered static over the entire simulation period. They were estimated by comparing the vehicle counts corresponding to each intersection approach aggregated over the entire observation period. Signal parameters were documented in the NGSIM data package, but we synchronized the precise offsets to correspond to trajectory data timestamps via the initiation of observed outflows in links corresponding to the major approaches for each intersection.

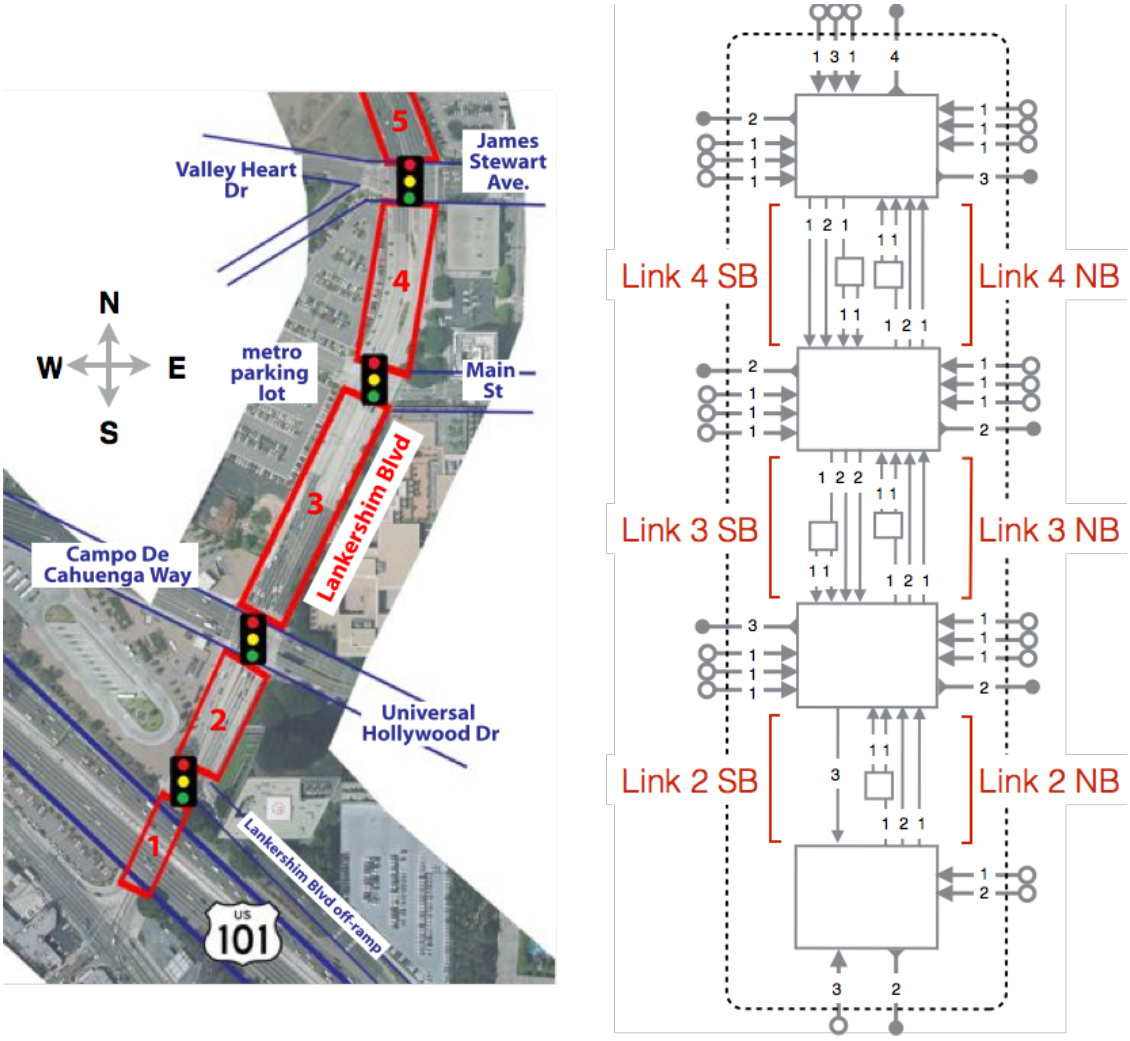

Figure 3.6: High-resolution vehicle trajectories are available for five blocks of Lankershim Blvd. The graphical representation to the right (not drawn to scale) was used to model link flows for independent movements. The four larger intersection nodes represent the four major signalized intersections on Lankershim Blvd. Each internal road is represented by 1-4 parallel links in each direction indicating the independently actuated movement queues. Five smaller intermediate nodes represent locations where spill-back from turning bays which could potentially cause partial blocking of the neighboring through movement. Network links are labeled with the corresponding lane counts used in simulation.

Comparison of VCM and CTM

Both VCM and CTM were able to accurately predict the general levels of congestion observed on the Lankershim Blvd network during the 30 minute observation period with minimal parameter tuning. In fact, we found only minute differences in the modeled link outflows

for the entire period of available data. Both models seemed to smooth "spikes" in the true link outflows, which were most likely physically caused by unpredictable variations in driver acceleration behaviors. When differences in modeled outflows did exist, the most common observation was that CTM seemed to attenuate the observed outflows at a slightly higher magnitude, as seen in the example in Figure 3.7. This could be caused by the delay on queue dissipation imposed by the dissipation wave speed **W** in CTM's flow-density relation: modeled flows are constrained in a queue slightly longer and at a location further upstream in CTM than in VCM, where the exit queue is only limited by link flow capacity and it is assumed that the travel time of de-queuing vehicles has already been incurred.

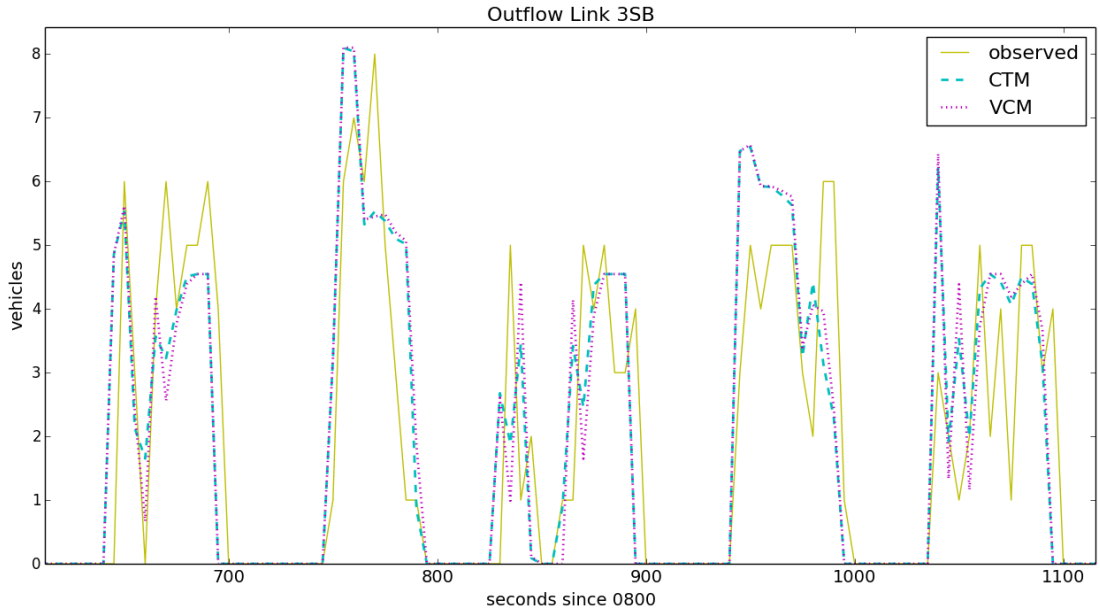

Figure 3.7: Modeled and observed flows exiting the through movement of Link 3 in the southbound direction illustrate typical outflow variations in this analysis. Both models seemed to smooth the true outflows, but reached approximated capacity flows at similar times.

The errors observed in all modeled link outflows, as shown in Figure 3.8, reveal little variation between the errors in outflows predicted by the two link dynamics. But by analyzing cumulative modeled link outflows, it becomes apparent that both models have a slight tendency to overestimate links outflows. The percent of cumulative error in outflow estimates for all through-movements on Lankershim Blvd. links over the entire simulation period are tabulated in Table 3.1. The similarity between the cumulative outflow estimation errors of the two models suggests that this error was more likely caused by unpredictable flow impediments, parameter mistuning, or a common misrepresentation in the network geometry than by differences in fundamental modeling assumptions.

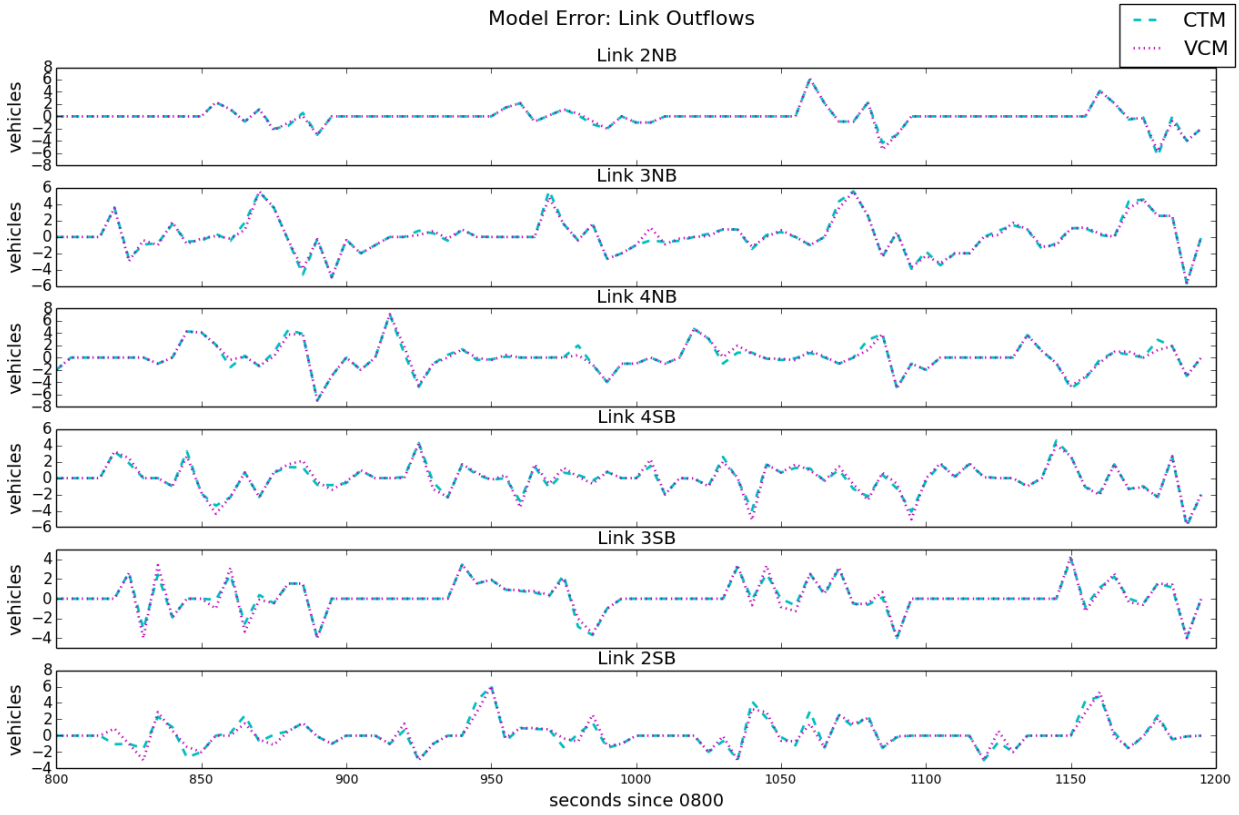

Figure 3.8: Differences in link outflow error between CTM and VCM (relative to observation) were minimal.

Table 3.1: Cumulative outflow model error.

Model Type	Link 2NB	Link 3NB	Link 4NB	Link 4SB	Link 3SB	Link 2SB
CTM	0.93%	4.53%	1.69%	0.79%	9.25%	6.74%
VCM	0.84%	4.29%	1.27%	0.76%	9.09%	6.53%

As expected given the similarities between link flow transfers, CTM and VCM yielded very similar link-vehicle counts aggregated over a 5-second time period. The modeled and observed states of all internal network links representing through movements are depicted in Figure 3.9. While VCM typically resulted in slightly larger link densities than CTM, it was not necessarily more representative of true link state. Neither model seemed to have a clear advantage over the other in correctly estimating link state.

Our results serve to validate the use of a vertical queueing model such as VCM in the case of the studied network. In fact, in our analysis (which was limited to a single model run by the lack of appropriate data), VCM appears to have performed slightly better than CTM in terms of cumulative link flow error. This work may help justify the relative analytical simplicity of

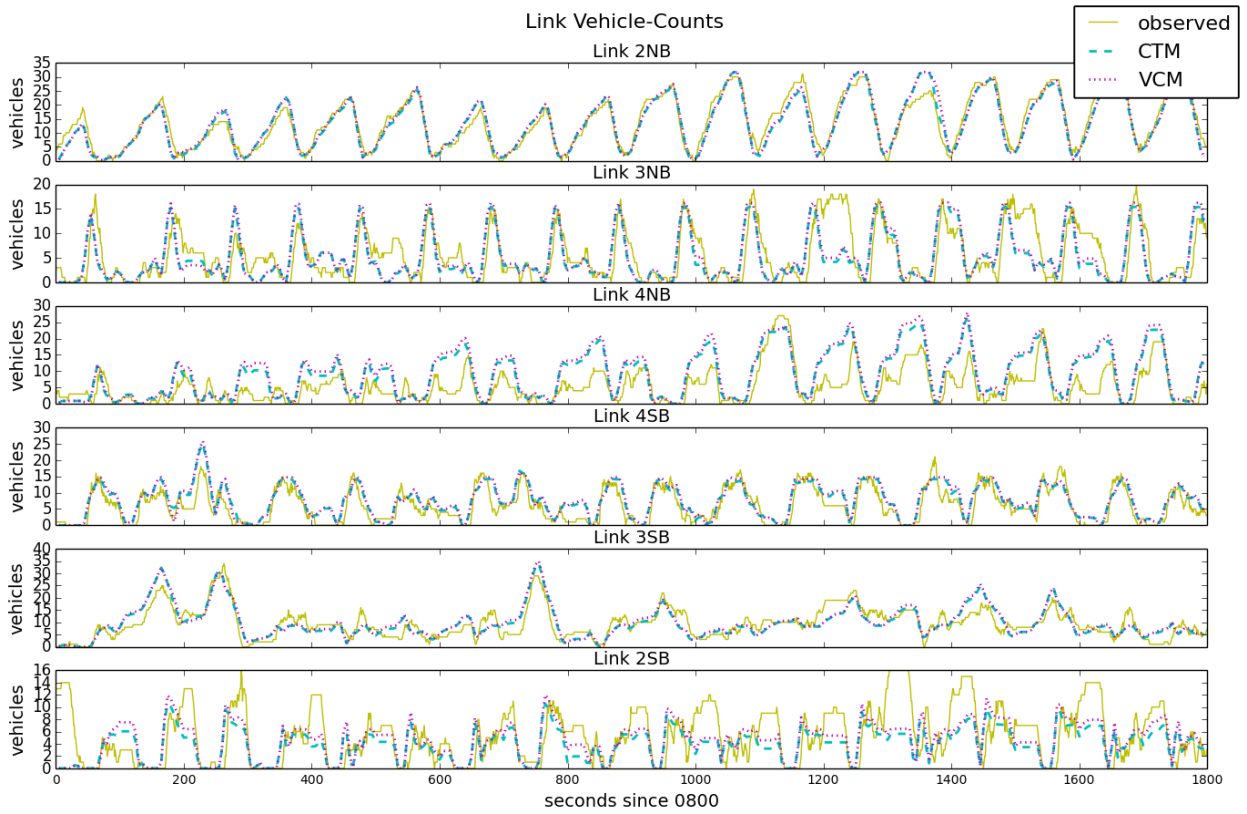

Figure 3.9: The modeled flow differences resulted in minor variations in link density states in the internal links, but neither model consistently resulted in a better representation of link state.

using a similar point queue dynamics for applications of model-based arterial estimation and signal control—especially in cases where the measurements available for controller feedback is of very low resolution.

Because the links on Lankershim Blvd remained in the undersaturated regime for the entire observation period, we were notably not able to investigate model performance in the high-congestion conditions where CTM is expected to out-perform a vertical model. Likely due to this lack of over-saturation, we also found very little output sensitivity to backwards shockwave speed **W** in the CTM model—which is the critical difference in the assumptions of CTM and VCM. This unfortunately prevented conclusive results on the relative benefits of a horizontal queueing model. Yet further analysis is limited by the lack of quality ground truth data on an appropriate urban network.

While VCM is analytically simpler than a horizontal queueing model and results in different queue dissipation behaviors, it does not provide any computational benefit over a CTM that is run at the same time discretization. Future work could include the development of a generalized analytical point queue model which could be implemented on a simplified

arterial graphical network. For example, great benefit could be derived from a link dynamics which eliminates the need for explicit representation of parallel "movement" links in the underlying network structure. Instead, separate capacities on co-located movement queues could be tracked within the mathematical structure of the vertical queuing model. This would greatly simplify the process of building an appropriate representative network structure– which is a significant obstacle to a practical implementation of such a queueing model.

Chapter 4

Estimation of predictable queueing behaviors using existing measurements

Many of the tools that are theoretically available to arterial traffic managers (including, by definition, any type of "traffic-responsive" arterial signalization scheme) rely on the accurate estimation of real-time or near-real-time vehicle queue lengths and turning ratios [163, 100, 205]. Yet as discussed in Chapter 2, there is currently no reliable and cost-effective method of measuring the most useful indicator of congestion on signalized roadways: the instantaneous length of vehicle queues. Furthermore, there are significant mathematical challenges of performing classical estimation techniques to derive queue state from the limited sensor measurements that do exist.

In this chapter we introduce a novel technique which can fuse multiple existing sources of data into a single arterial link-state estimate in a manner that is consistent with the kinematic-wave type models introduced in Chapters 2-3.

4.1 Background: previous approaches to arterial state estimation

The traffic engineering community does not currently possess a reliable and cost-effective methodology to estimate or predict the queue length on an arterial roadway or highway on-ramp. The following paragraphs explain the successes and limitations of the approaches commonly explored in relevant literature.

Input-output methods

By simple mass conservation, the number of vehicles currently in a queue must be the difference between entered vehicles and exited vehicles (plus the initial queue). Therefore several

proposed queue estimation schemes rely on counting the vehicles entering the queue (or upstream end of the link) and those leaving the queue (or downstream end of the link). These algorithms vary according to how and where they propose the counting to take place, but are generally referred to as *Input-Output* methods. They are known to introduce significant errors into queue estimates primarily for two reasons: an inability to correct for offsets introduced by vehicle miscounts, and incorrect estimates of the queue's initial length [123, 221, 206].

Work has been done to improve the performance of this queue estimation approach by incorporating occupancy measurements [206, 220], by introducing heuristic volume adjustment mechanisms [123, 221] or by using statistical analyses to correct for drifting count errors [207]. Multiple lines of research have suggested using statistical *Kalman filtering* techniques on a simple input-output queueing model to correct accumulating count errors and fuse data from existing types of arterial link detectors [44, 206]. While these methods have demonstrated improvements over simpler input-output techniques, such improvements often require more sensors than typically exist on a single arterial link. An experimental comparison of the Kalman filtering algorithm proposed in [206] and a simpler conservation-based procedure suggests that the relative benefits of the previous approach are often outweighed by the costs of the additional calibration procedures required for the underlying dynamical model [221].

Improvements in sensing technologies

Advanced magnetic sensors with vehicle re-identification capabilities can improve the results of input-output methods by occasionally "resetting" the initial queue estimates via tracking individual vehicles at each end of a queue [112, 180]. The re-identification algorithm is subsequently modified in order to improve the re-identification accuracy of slowly-moving vehicles, and was experimentally shown to provide adequate on-ramp queues estimates [179]. Unfortunately, most successful re-identification techniques rely on the use of specialized magnetometer arrays that are not yet widely available.

Comprehensive video detection systems provide queue lengths estimates via video processing techniques. But accuracy degrades with increasing distance from the video camera. Presence can only be detected within 400 feet, so longer queues are difficult to distinguish. Furthermore, these systems are also expensive to install and maintain—especially given their known sensitivity to weather conditions and various physical obstacles that could easily disrupt line-of-sight.

Reconstructing kinematic wave behaviors

After initial introduction of the LWR model of traffic flow, the study of kinematic waves on signalized roadways produced various theoretical estimates of queue length as a function of demand [177, 198, 142]. Yet all of these models assumed that input demand was uniform or otherwise predictably distributed. More recent work has applied these principles to real-

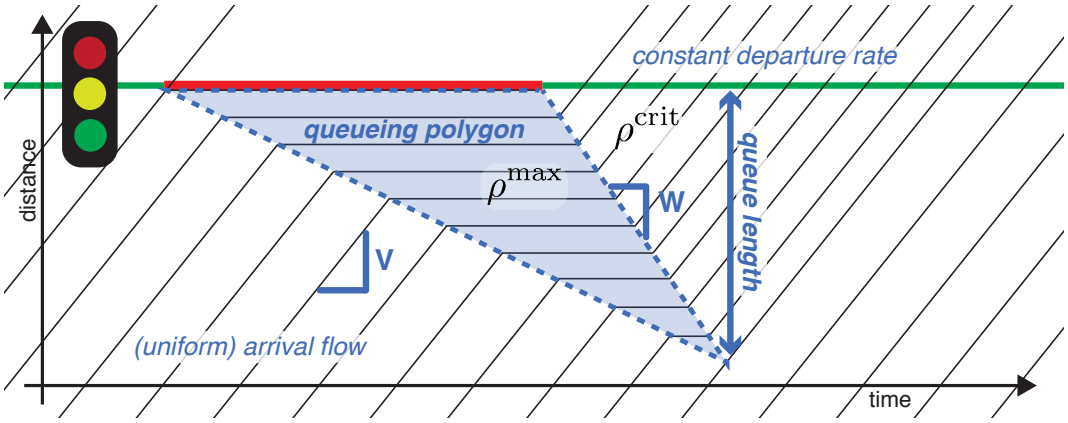

Figure 4.1: The temporal-spatial presence of queued vehicles forms a polygon with boundaries that can be estimated using knowledge of signal timings and arrival flows. Geometric principles can be used to calculate the presumed spatial extent of a vehicle queue. While the rate of queue formation depends on arrival rate, queue dissipation is assumed to be constant. A uniform arrival rate would create a triangular queueing polygon like the one shown here.

time sensor measurements in an attempt to reconstruct the back-of-queue behavior based on observed inflow profiles[191, 124].

With precise knowledge of signal timing switches, saturation headways and outflow counts from a stop line detector, and inflow measurements from advance detector actuations, it is possible to reconstruct the "queueing polygon" experienced on a movement approach at each signal cycle (see Figure 4.1). This provides an estimation of the position of the "back of the queue" with high spatial and temporal precision.

To further improve results, automated processes to estimate the maximum discharge rates (and thus dissipation shockwave speeds) used in the kinematic wave model have been proposed in [68, 185].

Limitations of this technique include the inability to estimate queues that extend beyond the advance detector in real-time (though it is possible to estimate maximum queue length after complete discharge [124, 192]), and more fundamentally, the lack of access to the high-frequency event-based (or even cycle-by-cycle) data required to implement the procedures.

Recent work has suggested the incorporation of information gathered from probe vehicles to estimate the shockwaves defining the queueing polygon. In [16], sparse travel time estimates are combined with signal information to reconstruct a probable queueing pattern that would have caused the observed delay. The algorithm presented in [173] uses GPS trajectories alone (without any knowledge of signal timings) to estimate the position of queueing shockwaves. This algorithm may become applicable once probe positioning data is more accurate (as to detect lane positioning), higher frequency (to receive multiple data points per link), and less sparse (to detect multiple vehicles per queue)—which may happen in the near future with the advent of connected vehicle systems.

4.2 An alternate representation of hydrodynamic traffic flows: cumulative number of vehicles

In section 2.2 we introduced the *Lighthill-Whitham-Richards* (LWR) model relating traffic density $\rho(t,x)$ and flow $f(t,x)$ [116, 174] . We restate it here for convenience:

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial \psi(\rho(t,x))}{\partial x} = 0 \tag{4.1}$$

This PDE has been widely used to predict arterial traffic; see for example [192, 75]. However, it is difficult to use this model for state estimation on short arterial links because its representation of density as a continuous aggregated quantity presents mathematical barriers for assimilating individual measurements of internal link flows or artificially-impeded vehicle trajectories. In addition, the non-smoothness of the solution of this PDE creates mathematical challenges for estimation [219, 28].

An alternate description of traffic state dynamics can be generated using an intuitive transformation of the conservations of vehicles principle [155, 46, 47]. Consider a function $\mathbf{M}(t,x)$ defined such that its spatial derivative is equal to the negative of the equation defining spatial density on a road link, and its temporal derivative is equivalent to the equation describing the resulting traffic flow:

$$\frac{\partial \mathbf{M}(t,x)}{\partial x} = -\rho(t,x) \tag{4.2}$$

$$\frac{\partial \mathbf{M}(t,x)}{\partial t} = f(t,x) = \psi(\rho(t,x)) \tag{4.3}$$

Equation (4.1) can be rewritten in terms of $\mathbf{M}(t,x)$ to formulate what we refer to as the *Moskowitz PDE* [39, 15]:

$$\frac{\partial \mathbf{M}(t,x)}{\partial t} + \psi\left(-\frac{\partial \mathbf{M}(t,x)}{\partial x}\right) = 0 \tag{4.4}$$

The solution $\mathbf{M}(t,x)$ of the Moskowitz PDE is called the *cumulative number-of-vehicles* function. Physically, it is an absolute measure of the total mass (number of vehicles) to have passed point x by time t. Another interpretation can be to consider assigning consecutive integer labels to vehicles entering a link at $x = \xi$ and tracing the trajectory of those vehicles over time, then if the vehicle labeled n is at location x' at time t', $\lfloor \mathbf{M}(t',x') \rfloor = n$. This concept is illustrated in Figure 4.2.

An additional benefit of this representation is the ease of calculating delay and queue length metrics. Consider plotting $\mathbf{M}(t,x')$ for multiple values of x' as a function of time on a single graph, as in Figure 4.3. The vertical distance between two neighboring curves is equal to the number of vehicles (queue length) present between the two corresponding x-values, and the horizontal distance between these curves is the travel time experienced by a vehicle traveling between the points.

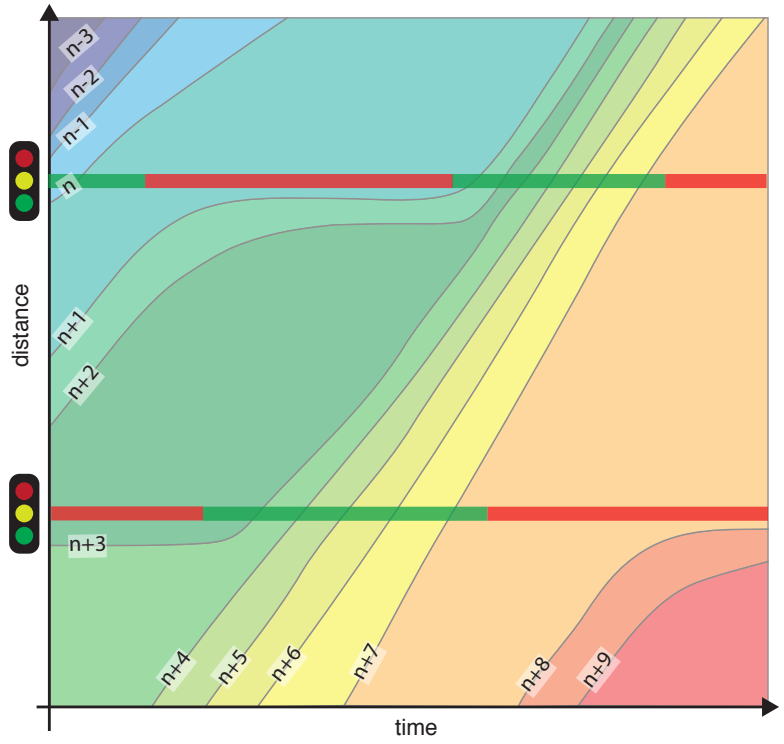

Figure 4.2: The level sets of the the cumulative number of vehicles function $\mathbf{M}(t, x)$ could be constructed by assigning integer labels to consecutive vehicles and tracing the trajectories of these vehicles through the entire domain being modeled, as illustrated in this example.

4.3 Developing boundary conditions from heterogeneous data sources on a signalized network

In this section we present a new method for queue length estimation on arterial links based on a macroscopic horizontal queuing model. Because it only depends on aggregate measurements and a macroscopic flow model to determine a "best fit" of link state given a known bound on measurement error, it is less sensitive to imprecise or erroneous measurements than some other queue estimation techniques. While our method can utilize measurements from traditional in-road sensors, it can also integrate measurements from advanced sensing systems such as re-identification or travel time monitors when they are available. In related work, the same techniques have been used with trajectory or position data on freeways, for example from GPS-enabled smartphones [37].

Our desired function is obtained by solving the Moskowitz model (4.4). Importantly, observe that (4.4) takes the form of a *Hamilton-Jacobi partial differential equation* (HJ-

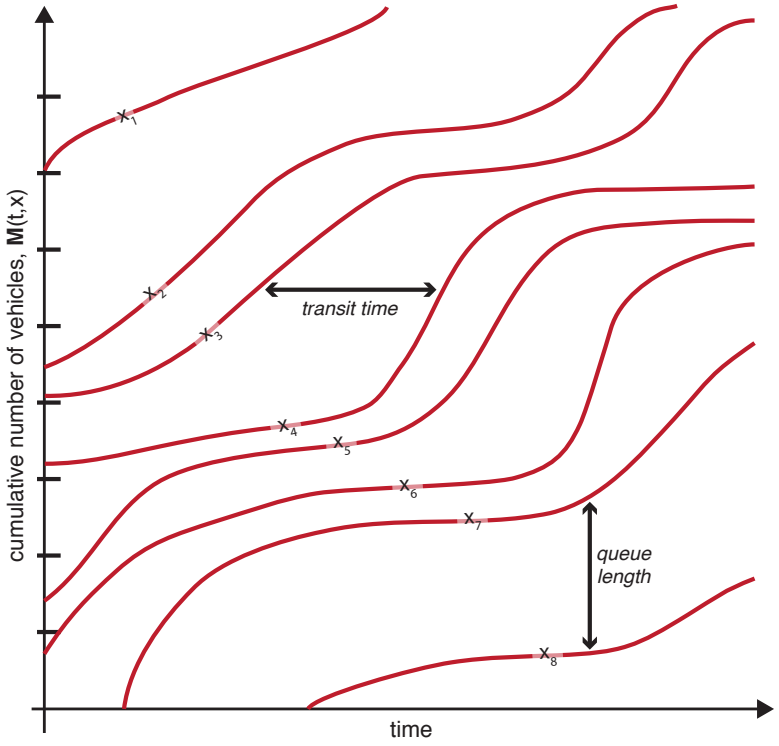

Figure 4.3: In this graph, $\mathbf{M}(t, x')$ is plotted for various fixed values of x'. Experienced transit time is portrayed as the horizontal distance between two consecutive curves. The vertical distance between two lines is a measure of instantaneous queue length, or the number of vehicles between two points on the road.

PDE) with the flux function $\psi(\rho)$ serving as the Hamiltonian. This specific model has been well-studied in the traffic community, in part because there are several known methodologies for finding explicit weak solutions to HJ-PDEs. In the present work we chose to use a class of weak solutions known as the *Barron-Jensen/Frankowska* (B-J/F) solutions [18, 67]. Because we are able to find an explicit analytical solution for traffic state, we can operate on any spatial or temporal resolution of sensor data without the need for mapping measurements to a discretized grid.

Measurements are incorporated into our solutions via choice of initial and/or boundary conditions to be input into the HJ-PDE. As previously presented in the context of freeways in [37], we choose boundary flows that optimize some desired convex function of the unknown value conditions within constraints imposed both by the kinematic dynamics of the LWR PDE and the available measurements. The objective of this algorithm is therefore to generate a realistic estimation of the aggregate traffic flow behaviors over the measured time horizon which could feasibly generate the included observations–fulfilling both an estimation and

data reconciliation functionality. Yet our work varies from classical approaches to estimation such as the Kalman filter: instead of iteratively finding the state estimate that minimizes least square measurement error, we seek a one-shot solution which does satisfy all available measurements but primarily optimizes an objective function designed to represent the most likely link dynamics that are "unknown" or left unconstrained by existing measurements.

We would like to emphasize that our objective is to reconstruct a general "averaged" measure of queuing behaviors and demands for the purposes of immediate estimation and control actuation. While others have studied means of adjusting macroscopic modeling to account for behavioral and higher-order dynamic effects [201], we do not attempt to reconstruct microscopic or even lane-specific behaviors.

Mathematical formulation

Consider an arterial road link defined between spatial locations ξ and χ. This link is unidirectional, has a constant number of lanes l along its entire domain, and traffic can only enter or exit at the upstream and downstream link boundaries (respectively).

We define the state of this link $\rho(t, x)$ to be the evolution of spatial density of the link for all locations $x \in [\xi, \chi]$ at all times $t \in [t_{\min}, t_{\max}]$. For known link parameters freeflow velocity v, shockwave (or queue dissipation) speed w, and critical density ρ_c, the flow $f(t, x)$ of vehicles across a single point x is described by the following piece-wise linear flux function:

$$f(t,x) = \psi(\rho(t,x) = \begin{cases} v\rho & \text{if } \rho \leq \rho_c \\ w(\rho - \rho_c) & \text{otherwise} \end{cases} \tag{4.5}$$

The road segment is bounded downstream (at $x = \chi$) by a traffic signal which can influence link state by impeding link outflow for fixed time durations. The time t_{red} at which downstream flow is artificially restricted by a red signal is known ($f(\chi, t_{\text{red}}) = 0$).

As with any PDE, a specific solution $\mathbf{M}(t, x)$ to (4.4) requires a pre-defined set of initial and/or boundary conditions to satisfy. Here we define the concept of a *value condition* to encompass the common notions of initial, boundary, and internal conditions. A value condition $\mathbf{c}(\cdot, \cdot)$ is defined as a lower semicontinuous function defined on some subset of domain $[0, t_{\max}] \times [\xi, \chi]$. Any solution to the PDE being investigated must satisfy all associated value conditions on their respective domains.

In this work, the value conditions $\mathbf{c}_j$ are not known specifically but rather must be estimated before attempting to solve for $\mathbf{M}(t, x)$. Chosen conditions must not only satisfy the physical limitations imposed by the model, but also permit the feasibility of any available measurements of network state. We therefore use the following framework to develop constraints on the set of feasible value conditions.

Format of initial and boundary conditions

This work specifically employs a class of weak solutions to HJ-PDEs known as the *Barron-Jensen/Frankowska* (B-J/F) solutions [18, 67]. These solutions are represented by the *Lax-*

Hopf formula [15, 38].

Definition 4.1 (**Lax-Hopf formula**). For value condition $\mathbf{c}_j(\cdot,\cdot)$,

$$\mathbf{M}_{\mathbf{c}}(t,x) = \inf_{(u,T)\in \mathrm{Dom}\ (\varphi^*)\times\mathbb{R}^+} (\mathbf{c}(t-T, x+Tu) + T\varphi^*(u)) \tag{4.6}$$

where $\varphi^*(\cdot)$ is the *Legendre-Fenchel transform* of Hamiltonian $\psi(\cdot)$.

Definition 4.2 (**Legendre-Fenchel transform**).

$$\varphi^*(u) := \sup_{p\in \mathrm{Dom}(\psi)} [p\cdot u + \psi(p)] \tag{4.7}$$

Assume the following affine (generalized) initial and upstream/downstream boundary conditions, defined for discrete spatial blocks k of length X and discrete time blocks n of length T:

initial condition:

$$M_k(t,x) = \begin{cases} -\sum_{i=0}^{k-1}\rho(i)X - \rho(k)(x-kX) & \text{if } t=0 \text{ and } x\in[kX,(k+1)X] \\ +\infty & \text{otherwise} \end{cases} \tag{4.8}$$

upstream condition:

$$\gamma_n(t,x) = \begin{cases} \sum_{i=0}^{n-1} f_{\mathrm{in}}(i)T + f_{\mathrm{in}}(n)(t-nT) & \text{if } x=\xi \text{ and } t\in[nT,(n+1)T] \\ +\infty & \text{otherwise} \end{cases} \tag{4.9}$$

downstream condition:

$$\beta(t,x) = \begin{cases} \sum_{i=0}^{n-1} f_{\mathrm{out}}(i)T + f_{\mathrm{out}}(n)(t-nT) & \\ \qquad\qquad - \sum_{k=0}^{k_{\max}}\rho(k)(x-kX) & \text{if } x=\chi \text{ and } t\in[nT,(n+1)T] \\ +\infty & \text{otherwise} \end{cases} \tag{4.10}$$

A direct application of (4.6) on (4.8)-(4.10) yields the following solutions (via the approach

of [38]):

$$\mathbf{M}_{M_k} = \begin{cases} +\infty & \text{if } x \le kX + wt \text{ or } x \ge (k+1)X + vt \\ -\sum_{i=0}^{k-1} \rho(i)X + \rho_c(tv + kX - x) & \text{if } kX + tw \le x \le kX + tv \\ -\sum_{i=0}^{k-1} \rho(i)X + \rho(k)(tv + kX - x) & \text{if } kX + tv \le x \le (k+1)X + tv \\ -\sum_{i=0}^{k} \rho(i)X + \rho_c(tw + (k+1)X - x) - \rho_m tw & \text{if } (k+1)X + tw \le x \le (k+1)X + tw \\ -\sum_{i=0}^{k-1} \rho(i)X + \rho(k)(tw + kX - x) - \rho_m tw & \text{if } kX + tw \le x \le (k+1)X + tv \end{cases} \tag{4.11}$$

$$\mathbf{M}_{\gamma_n} = \begin{cases} +\infty & \text{if } t \le nT + \frac{x-\xi}{v} \\ \sum_{i=0}^{n-1} f_{\text{in}}(i)T + f_{\text{in}}(n)(t - \frac{x-\xi}{v} - nT) & \text{if } nT + \frac{x-\xi}{v} \le t \le (n+1)T + \frac{x-\xi}{v} \\ \sum_{i=0}^{n-1} f_{\text{in}}(i)T + \rho_c v(t - (n+1)T - \frac{x-\xi}{v}) & \text{otherwise } \left(\text{if } t > (n+1)T + \frac{x-\xi}{v}\right) \end{cases} \tag{4.12}$$

$$\mathbf{M}_{\beta_n} = \begin{cases} +\infty & \text{if } t \le nT + \frac{x-\chi}{w} \\ \sum_{i=0}^{n-1} f_{\text{out}}(i)T + f_{\text{out}}(n)(t - \frac{x-\chi}{w} - nT) & \\ \qquad - \sum_{k=0}^{k_{\max}} \rho(k)X - \rho_m(x - \chi) & \text{if } nT + \frac{x-\chi}{w} \le t \le (n+1)T + \frac{x-\chi}{w} \\ \sum_{i=0}^{n} f_{\text{out}}(i)T - \sum_{k=0}^{k_{\max}} \rho(k)X & \\ \qquad + \rho_c v\left(t - (n+1)T - \frac{x-\chi}{w}\right) & \text{otherwise } \left(\text{if } t > (n+1)T + \frac{x-\chi}{w}\right) \end{cases} \tag{4.13}$$

Model constraints

Note that while (4.6) implies that $\mathbf{M}_{\mathbf{c}}(\cdot,\cdot)$ exists for any $\mathbf{c}$, this solution is not guaranteed to be compatible with the corresponding value condition. In other words, it is not necessarily true that $\forall (t,x) \in \text{Dom}(\mathbf{c}),\ \mathbf{M}_{\mathbf{c}}(t,x) = \mathbf{c}(t,x)$.

To account for this, note that the structure of equation (4.6) implies the *inf-morphism* property:

Definition 4.3 (**Inf-Morphism Property**)**.** Let $\mathbf{c}(\cdot,\cdot)$ be a minimum of a finite number of lower semicontinuous functions,

$$\forall (t,x) \in [0, t_{\max}] \times [\xi, \chi],\ \mathbf{c}(t,x) := \min_{j \in J} \mathbf{c}_j(t,x) \tag{4.14}$$

Then $\mathbf{M}_{\mathbf{c}}$ can be decomposed as

$$\forall (t,x) \in [0, t_{\max}] \times [\xi, \chi],\ \mathbf{M}_{\mathbf{c}}(t,x) = \min_{j \in J} \mathbf{M}_{\mathbf{c}_j}(t,x) \tag{4.15}$$

For reference on the inf-morphism property, see [38].

Define decision variable associated with the value conditions in (4.8), (4.9), and (4.10) as

$$y := (\rho(1), \ldots, \rho(k_{\max}), f_{\text{in}}(1), \ldots, f_{\text{in}}(n_{\max}), f_{\text{out}}(1), \ldots, f_{\text{out}}(n_{\max})) \tag{4.16}$$

and denote by $\mathcal{Y}$ the vector space of these decision variables y. In this work, we use the inf-morphism property to ensure that all value conditions used to find a final solution for $\mathbf{M}(t,x)$ will apply in the strong sense by defining a set of physical constraints on $\mathcal{Y}$ that are implied by the explicit solutions (4.11), (4.12), and (4.13): value condition $\mathbf{c}(\cdot,\cdot) = \min_{j\in J} \mathbf{c}_j(\cdot,\cdot)$ satisfies $\forall (t,x) \in \text{Dom}(\mathbf{c})$, $\mathbf{M}_{\mathbf{c}}(t,x) = \mathbf{c}(t,x)$ if and only if

$$\mathbf{M}_{\mathbf{c}_j}(t,x) \geq \mathbf{c}_i(t,x)\ \forall (t,x) \in \text{Dom}(\mathbf{c}_i),\ \forall (i,j) \in J^2 \tag{4.17}$$

Explicitly, (4.17) implies the following *model constraints* on y:

$$\begin{cases} \mathbf{M}_{M_k}(0,x_p) \geq M_p(0,x_p) & \forall (k,p) \in \mathbb{K}^2 \\ \mathbf{M}_{M_k}(pT,\chi) \geq \beta_p(pT,\chi) & \forall k \in \mathbb{K}, \forall p \in \mathbb{N} \\ \mathbf{M}_{M_k}(\frac{\chi - x_{k+1}}{v},\chi) \geq \beta_p(\frac{\chi - x_{k+1}}{v},\chi) & \forall k \in \mathbb{K}, \forall p \in \mathbb{N} \text{ s.t } \frac{\chi - x_{k+1}}{v} \in [pT,(p+1)T] \\ \mathbf{M}_{M_k}(pT,\xi) \geq \gamma_p(pT,\xi) & \forall k \in \mathbb{K}, \forall p \in \mathbb{N} \\ \mathbf{M}_{M_k}(\frac{\xi - x_k}{w},\xi) \geq \gamma_p(\frac{\xi - x_k}{w},\xi) & \forall k \in \mathbb{K}, \forall p \in \mathbb{N} \text{ s.t } \frac{\xi - x_k}{w} \in [pT,(p+1)T] \end{cases} \tag{4.18}$$

$$\begin{cases} \mathbf{M}_{\gamma_n}(pT,\xi) \geq \gamma_p(pT,\xi) & \forall (n,p) \in \mathbb{N}^2 \\ \mathbf{M}_{\gamma_n}(pT,\chi) \geq \beta_p(pT,\chi) & \forall (n,p) \in \mathbb{N}^2 \\ \mathbf{M}_{\gamma_n}(nT + \frac{\chi - \xi}{v},\chi) \geq \beta_p(nT + \frac{\chi - \xi}{v},\chi) & \forall (n,p) \in \mathbb{N}^2 \text{ s.t } nT + \frac{\chi - \xi}{v} \in [pT,(p+1)T] \end{cases} \tag{4.19}$$

$$\begin{cases} \mathbf{M}_{\beta_n}(pT,\xi) \geq \gamma_p(pT,\xi) & \forall (n,p) \in \mathbb{N}^2 \\ \mathbf{M}_{\beta_n}(nT + \frac{\chi - \xi}{w},\xi) \geq \gamma_p(nT + \frac{\chi - \xi}{w},\xi) & \forall (n,p) \in \mathbb{N}^2 \text{ s.t } nT + \frac{\chi - \xi}{w} \in [pT,(p+1)T] \\ \mathbf{M}_{\beta_n}(pT,\chi) \geq \beta_p(pT,\chi) & \forall (n,p) \in \mathbb{N}^2 \end{cases} \tag{4.20}$$

For a full derivation of these inequalities, refer to [37].

Notice that because the solutions described by (4.8)-(4.10) associated with the given value conditions are all linear in y, all of these constraints described by (4.18)-(4.20) are also linear in y. We can therefore represent the model constraints in the matrix form

$$A_{\text{model}} y \leq b_{\text{model}} \tag{4.21}$$

Data constraints

While the model constraints (4.18)-(4.20) encode the limitations due to the physics of traffic flow, they do not add any new information about the existing state of a system. To estimate boundary conditions such that all known measurements will be satisfied by the derived solution, we must define a separate set of *data constraints*. This requires explicit formulation of the sensor data in terms of decision variable y.

To preserve convexity in the resulting optimization problem, data constraints can often be represented as convex inequalities which account for errors inherent in practical measurement techniques. Here we will furthermore assume that all data constraints are linear; they are therefore represented in general as

$$C_{\text{data}} y \leq d_{\text{data}} \tag{4.22}$$

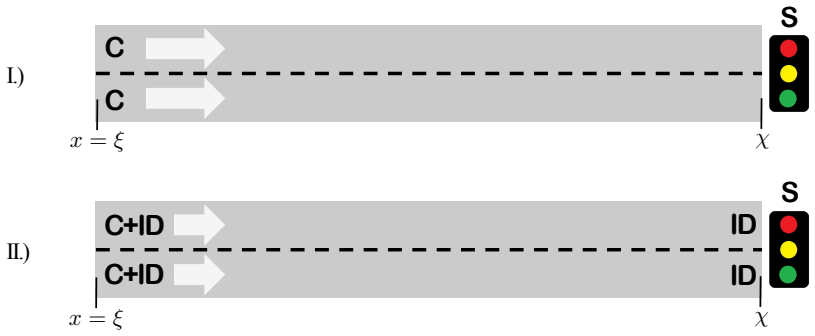

Figure 4.4: Sensor configurations investigated in this work. **C** denotes an available count measurement; **ID** denotes a sensor with vehicle identification capabilities for travel time measurements.

For example, we specifically investigate the following realistic infrastructural sensor configuration scenarios that are illustrated in Figure 4.4:

Scenario I: Detectors providing vehicle counts are placed at the upstream boundary of the link, providing vehicle count measurements that can be aggregated into flow estimates $\bar{f}^k(T, \xi)$ for a fixed time step T. These measured flows have known error percentage $\bar{e_f}$. Because signal timings are known, partial information about link outflow is also available.

Relevant Data Constraints:

- $f_{\text{out}}(t_{\text{red}}) = 0$
- $(1 - \bar{e_f})\bar{f}^k(t_k, \xi) \leq f_{\text{in}}(k) \leq (1 + \bar{e_f})\bar{f}^k(t_k, \xi) \quad \forall\, t_k \in [k \cdot T, (k+1) \cdot T]$

Scenario II: Flow measurements as in Scenario I are given. Additionally, re-identification sensors placed at the upstream and downstream ends of the link provide point-to-point travel times $\bar{t}$ with maximum error $\bar{e_t}$, corresponding to exit time stamps $\bar{t_f}$ for 5-15% of the vehicles traveling across the link.

Relevant Data Constraints:

- $f_{\text{out}}(t_{\text{red}}) = 0$
- $(1 - \bar{e_f})\hat{f}^k(t_k, \xi) \leq f_{\text{in}}(k) \leq (1 + \bar{e_f})\bar{f}^k(t_k, \xi) \quad \forall\, t_k \in [k \cdot T, (k+1) \cdot T]$
- $M(\bar{t_f} - \bar{t} - \bar{e_t}, \xi) \leq M(t_f, \chi) \leq M(\bar{t_f} - \bar{t} + \bar{e_t}, \xi)$ for $\bar{t}$, $\bar{t_f}$ sampled from 5-15% of exiting vehicles

Estimation of unobservable boundary flows

To estimate the unknown or uncertain boundary conditions, we formulate an objective problem over space $\mathcal{Y}$ with the model constraints (4.18)-(4.20) and data constraints corresponding to any available measurements:

$$\begin{aligned} \text{minimize:} \quad & g(y) \\ \text{subject to:} \quad & \begin{cases} A_{\text{model}} y \leq b_{\text{model}} \\ C_{\text{data}} y \leq d_{\text{data}} \end{cases} \end{aligned} \tag{4.23}$$

The objective $g(y)$ can be any convex piecewise affine function of the decision variable. However if limited availability of data suggests a highly underdetermined problem, the objective should be crafted to ensure realism in the resulting solution.

For example, the scenarios investigated in this work do not include full constraints on outflows via measurements; they only assume zero outflow when impeded by a signal. Therefore many feasible solutions y with various exiting flow profiles can satisfy the existing constraints. But because drivers usually act to maximize their velocity whenever possible, we should prefer solutions where rapid outflow is encouraged. This is achieved by maximizing the sum of outflows weighted by a small, decreasing function $\mu(n)$:

$$\begin{aligned} \max_{y \in \mathcal{Y}} \quad & \sum_n \mu(n) f_{\text{out}}(n) \\ \text{subject to:} \quad & \begin{cases} A_{\text{model}} y \leq b_{\text{model}} \\ f_{\text{out}}(t_{\text{red}}) = 0 \\ \text{[other data constraints]} \end{cases} \end{aligned} \tag{4.24}$$

Queue calculation

Ultimately, the optimal initial/boundary conditions $y^* = \arg\max(g(y))$ are used to determine the solution of the Moskowitz function explicitly via equations (4.11) - (4.13). The integer level-sets of the resulting piecewise linear function $\mathbf{M}(t, x)$ represents "modeled" vehicle trajectories. Multiple link performance metrics can be estimated from this result, including queue lengths. Two criteria will identify queues:

- Point density is maximized: $\rho(t, x) = \left| \frac{\partial \mathbf{M}(t,x)}{\partial x} \right| = \rho_j \pm \delta$

- Point flow is zero: $f(t,x) = \frac{\partial \mathbf{M}(t,x)}{\partial t} \approx 0$

The first criterion (maximum density) is a more reliable indicator of queued state than zero flow, as the latter may also occur when the link is entirely empty. We therefore define the instantaneous queue length as the location of the boundary between areas of jam density and areas of lesser density at each time t.

4.4 Experimental results

This algorithm was validated on the NGSIM vehicle trajectory data set captured on Lankershim Blvd. in Los Angeles, California. A detailed description of this data set is provided in Section 2.6 of this work.

We simulated count sensors at the immediate upstream end of each link by extracting all timestamps at which a vehicle enters the link from the adjacent intersection. Flow measurements were then estimated by aggregating these "counts" within every five-second time period. We extracted travel time measurements from randomly sampled trajectories, where entry and exit times correspond to the timestamps at which the sampled vehicles were first and last detected on the relevant link. Note that the time a vehicle spent within the surrounding intersections is not included in the travel time samples. Red signal times were extracted from the signal timing plans provided in the NGSIM database.

For demonstration of our algorithm, we chose to analyze data from the four links highlighted in Figure 4.5:

- link 2 southbound, a 3-lane link between a busy cross-street and a signalized intersection with no possible turn movements;
- link 2 northbound, a 3-lane link that expands to 5 lanes downstream with one designated left-turn lane and two permissive right-turn lanes;
- link 3 southbound, a link with three through-lanes, two left-turn lanes, and a right-turn pocket;
- link 4 northbound, a 4-lane link with an intermediate entry-exit point and a small left-turn pocket at the downstream end.

These links were chosen to be representative of a variety of physical features, such as both specialized and shared turn lanes, and intermediate entry/exit points.

Calibration of the fundamental diagram parameters v, w, and f_c was performed via visual inspection of the trajectories. The following values were used for all links:

- free-flow velocity, v = 15.64 meters/sec (35 miles/hour)
- shockwave velocity, w = -6.70 meters/sec (-15 miles/hour)

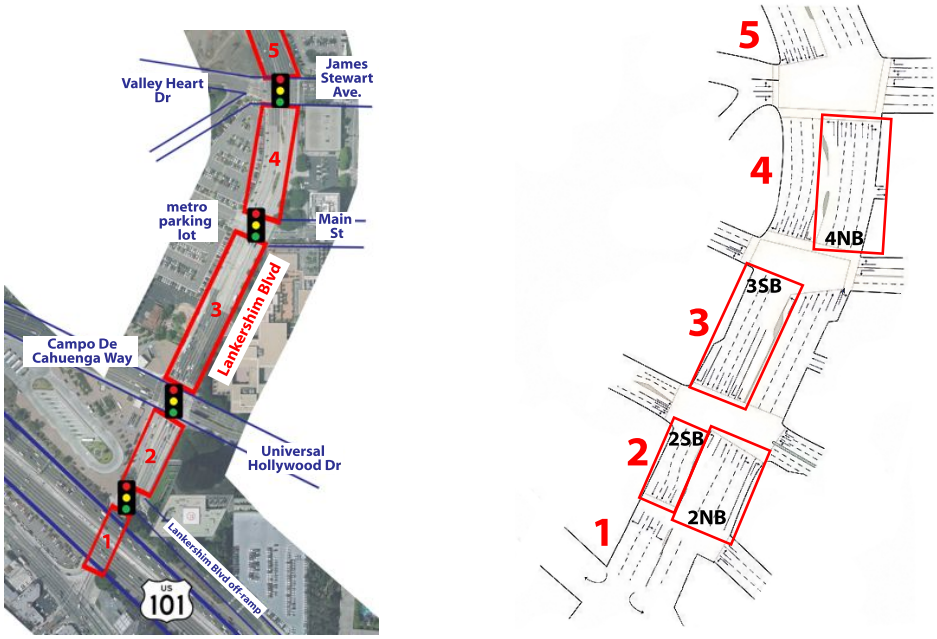

Figure 4.5: High-resolution vehicle trajectories are available for 5 blocks of Lankershim Blvd. We tested our queue estimation algorithm on the highlighted links, which are representative of multiple typical link geometries: 2 northbound, 2 southbound, 3 southbound, and 4 northbound.

- critical density, $\rho_c = 0.375$ vehs/meter (60 vehs/mile)
- jam density, $\rho_j = \rho_c(1 + \frac{v}{w}) = 0.125$ vehs/meter (200 vehs/mile)

These parameters correspond to the dynamics of a single lane. To ensure that the "measured" input flows are treated consistently, the calculated flows were scaled by the inverse of the number of lanes at link entry. Results are therefore intended to represent an "average" queuing behavior on each of the links, and not expected to exactly match the behavior observed on any one lane.
We also chose common values for measurement error:

- Count sensors are accurate within 5%.
- Travel time estimates have a maximum error of 0.5 seconds.

We solved the relevant linear programs using a MATLAB-based optimization software package. We then used a separate MATLAB toolbox to generate the desired B-J/F solutions to the Moskowitz HJ-PDE [136]. This LWR toolbox is available at `http://traffic.berkeley.edu/project/downloads/lwrsolver`.

We ran this code on each of the four link-directions shown in Figure 4.5 for all sensor configuration scenarios. Specifically, we compared the time-resolved queue length estimates

generated from the calculated Moskowitz solutions to those detected in the data. In the data, we define the back of a queue as the location of the car with the highest entry index that is stopped on a link at a given time. Because vehicles tended to "drift" slightly when in a queue, this detection method was not always accurate; discontinuities in detected queue lengths sometimes caused unrealistic noise in the resulting queue length error calculations.

Scenario I: observing upstream flows and signals

Figure 4.6 illustrates a sample of the results of our estimation algorithm on Link 2 SB, a 3-lane link with simple geometry with no possible downstream turn movements.

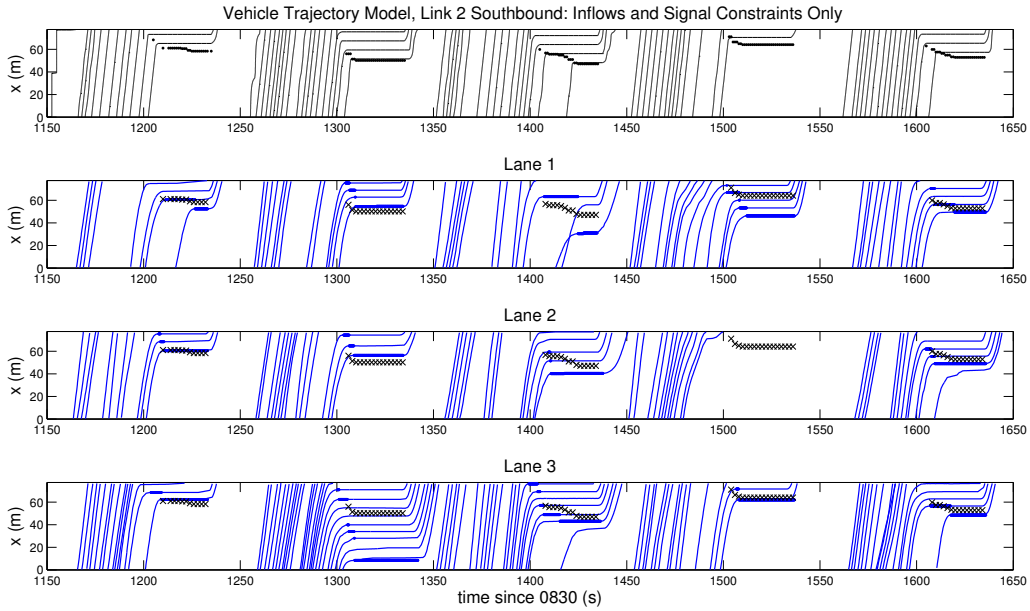

Figure 4.6: Modeled and ground-truth trajectories for a sample time period. Inflows and constrained (zero) outflows were imposed. For visual comparison, the queue lengths estimated by the PDE model are shown on all data plots.

With this basic lane geometry and low demands, we see that the modeled trajectory behavior closely follows an "average" of the three exiting lanes. However it fails to replicate the excessive queueing (and possible spillover) seen on lane 3 at 1300 seconds. Replication of such extreme queuing occurring only in a single lane is not expected given the lane-averaged flows input into the model. To achieve a more accurate representation of true behavior, one may need to access lane-specific flow sensors and run this model on each lane independently. This procedure, however, would likely be sensitive to lane-changing behaviors and inaccuracies in turn ratio estimation.

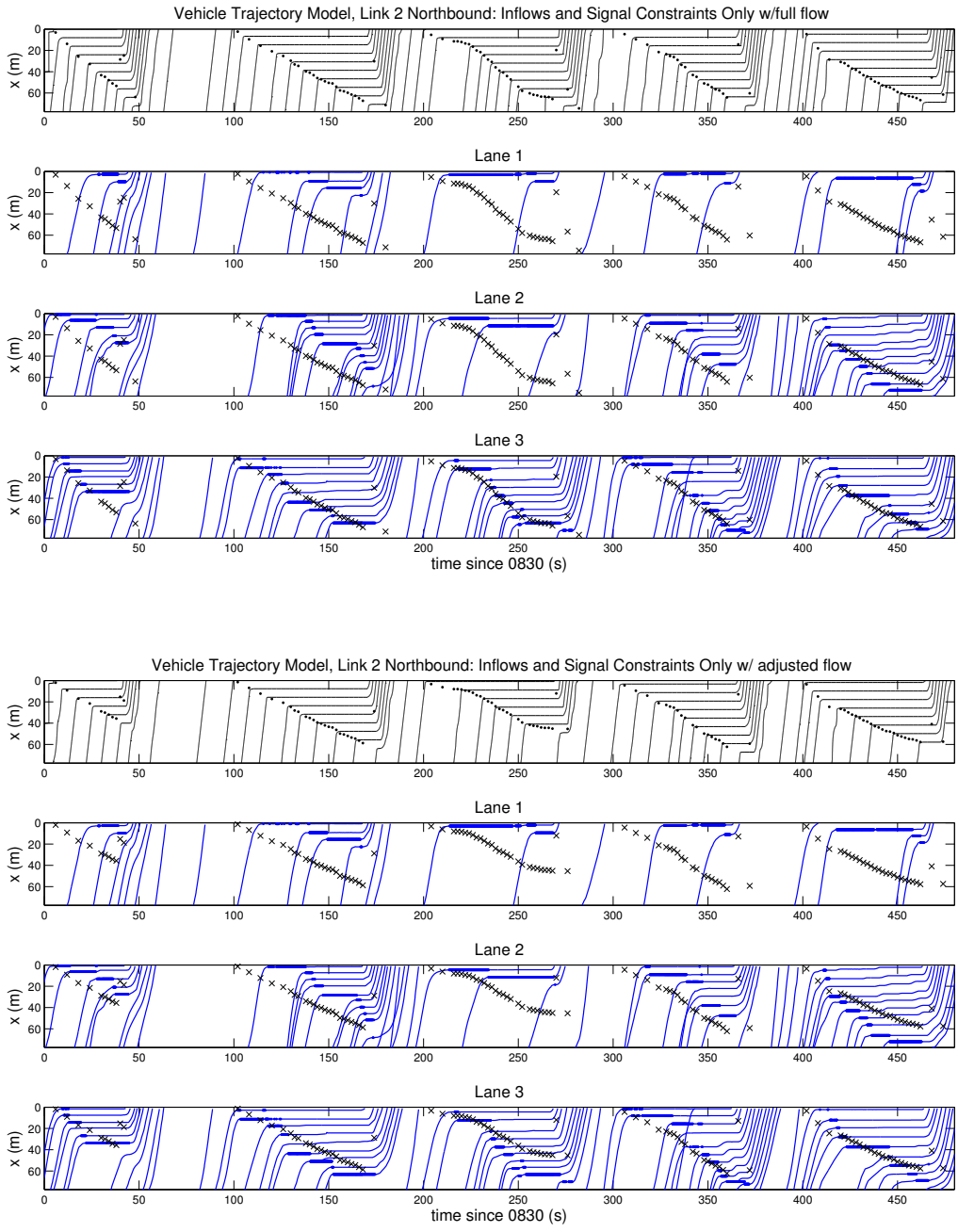

Figure 4.7: Top: Algorithm demonstrated on Link 2NB with turning lanes present and no inflow adjustments. Bottom: Model performance on Link 2NB with inflow reduced by 24%, to be representative of through movements only.

Achieving similarly accurate queue length results on links with complicated geometries and downstream turn movements requires additional processing. Link 2 NB, for example, is a link with two through-only lanes, one shared through-right turn lane, one right turn only lane, and one left turn only lane, as seen in Figure 4.5. We used aggregate inflows from the three lanes present at the upstream end of the link, and restricted modeled outflows according to signal timings which restricted through movement at the downstream end. We also assumed that flows were evenly divided between the three through movement lanes, and therefore divided inflows by three before modeling.

However the resulting trajectories, such as those shown in Figure 4.7, tended to overestimate the queues in all of the through lanes. These results suggested that it was necessary to further reduce the inflow measurements used in the modeling process to account for the turning flows, which not only entered downstream queues disproportionately but also were restricted by different signal timings than those of the through-flows. We therefore reduced the measured inflows by the estimated percentage of turning vehicles before processing data constraints.

For example, in the case of link 2 NB we determined that over the entire 30-minute study period, approximately 4% of vehicles exiting the link in this direction turned left and 20% turned right. Hence we reduced inflows by 24%, and re-ran the optimization and PDE solution procedures. The trajectories modeled using the lesser inflows were more representative of the average behavior seen on all through-only lanes, as can be seen in Figure 4.7. Note that while we were able to "predict" turn ratios fairly accurately in this work via analysis of our detailed data set, similar procedure can be followed in practice using turning ratio estimates determined by previous local surveying or OD-estimation techniques.

Figure 4.8 demonstrates similarly successful results on the through-only lanes of Link 3 southbound, a block with two dedicated left-turn lanes and a third dedicated right-turn lane.

The estimation error function, illustrated in Figure 4.9, reveals that while the accuracy of the queue length estimate varies significantly between lanes, link-average error remains very low–within ± 16 meters, or two car lengths at maximum density $\rho_j = 0.125$ vehicles/meter. These results were typical of instances where there was no abnormal disturbances such as a long truck or turn lane spillover on any lanes of the link. There is no evidence that queues are systematically over-estimated or underestimated, or that the estimated lengths are consistently less accurate at either the beginning or end of a queueing cycle using this technique.

In our study of Link 4 NB, we expected that the intermediate entry/exit point would cause error in both modeled queue lengths because this comprised an obvious violation in the mass conservation assumption of the underlying LWR model. Yet the level of flows exiting and entering the link did not constitute a significant percentage of link flows in the samples we studied, and thus the results were not notably affected by such violation.

Scenario II: observing upstream flows, signals, and travel times

The additional constraints due to travel time measurements as described in Scenario II

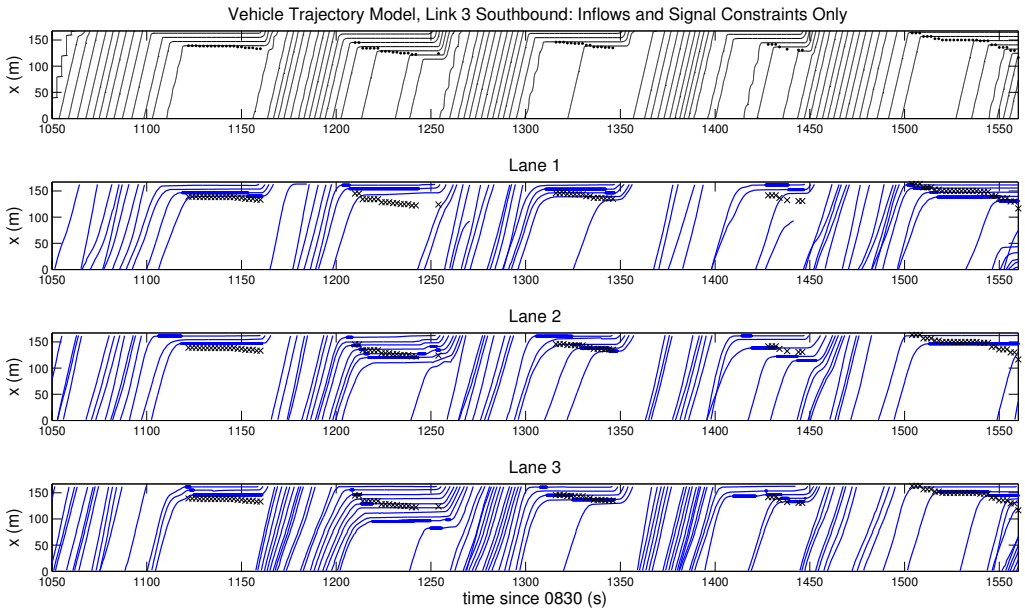

Figure 4.8: Model performance on through lanes of Link 3 SB. Inflow was reduced by 36% to account for turning vehicles. Modeled queues are represented in bold lines on all plots.

initially caused the boundary condition optimization function to become over-constrained and thus infeasible. This is due to the flow and queue aggregation assumed in our implementation. For example, Figure 4.6 illustrates a situation where a vehicle entering Lane 3 of Link 2 at 1300 seconds would encounter a dramatically different queue (and thus experience a significantly different travel time) than a vehicle concurrently entering Lane 2 of the same link. If conflicting travel times were sampled, the corresponding conflicting constraints would cause the problem to become unsolvable.

Without studying individual lane behaviors, we were therefore constrained to using very small penetration rates which did not contain samples which conflicted outside the range of permissible error. We also made a further adjustment in the solution procedure: in addition to the 0.5 second error permitted in travel time measurements, we added a 0.25-vehicle error on the solution of the Moskowitz function directly. This effectively modified a travel time constraint to the following:

$$M(\bar{t_f} - \bar{t} - \bar{e_t}, \xi) - 0.25 \leq M(\bar{t_f}, \chi) \leq M(\bar{t_f} - \bar{t} + \bar{e_t}, \xi) + 0.25 \tag{4.25}$$

While these adjustments to the boundary condition algorithm allowed for the identification of feasible solutions, they also minimized the impact of travel time measurements on the resulting trajectories and queue lengths. We found that with realistic penetration rates of 5-15%, the addition of travel time estimates did little to improve the accuracy of

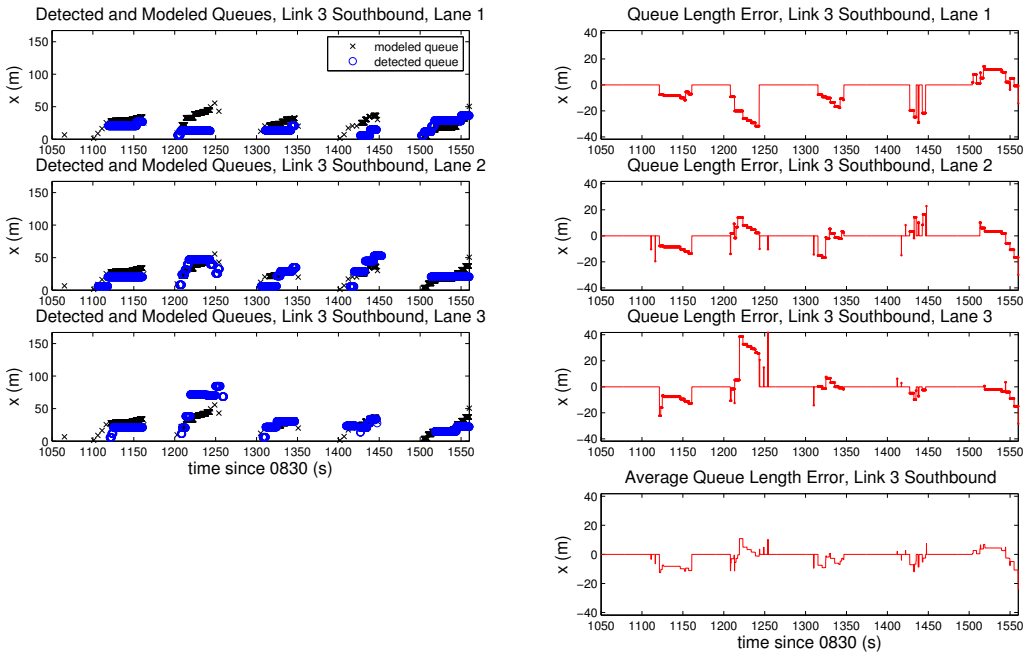

Figure 4.9: A comparison of modeled queues to detected queues. Average queue length error remains under 16 meters, or < 2 vehicles at jam density $\rho_j = 0.125$ vehicles/meter.

modeled queue lengths. In most common congestion patterns with well-distributed flow, the additional constraints were already satisfied by the solution found in Scenario I and thus did not have any impact on the resulting trajectories. When additional active constraints were imposed by travel time samples, they did not typically improve the lane-averaged error in queue length estimates. See for example the trajectories on Link 3SB shown in Figures 4.10 - 4.11.

From the comparison of the error resulting from estimates of the two sensor scenarios in Table 4.1 below, it is clear that travel time measurements do not consistently provide useful information beyond that available with just inflow and signal timing information.

Table 4.1: Average Absolute Error in Queue Length Estimates

Link	Scenario 1	Scenario 2 (15%)
2 SB	9.88 m	9.88 m
2 NB	14.73 m	19.30 m* (w/ 5%)
3 SB	13.69 m	15.53 m
4 NB	11.67 m	11.67 m

The best results for both scenario were observed in Link 2SB, the link with no turn

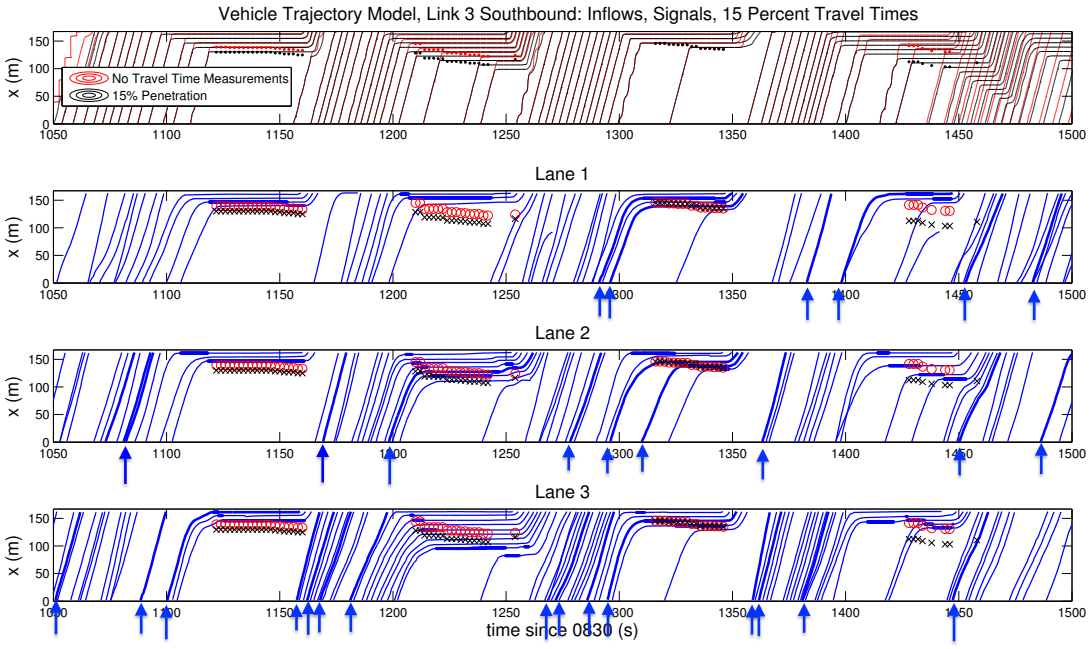

Figure 4.10: The trajectories from which travel times were sampled are highlighted with an arrow at time of entry. Several "outlier" trajectories from the end of the queues on Lanes 2 and 3 caused an adjustment in the modeled queue for the first, second, and fourth light cycles.

movements to cause differentiation in lane behaviors. In contrast, Link 2NB is a short link with both left and right turn movements; it is likely that the rapid lane changing and queue blocking of the turning vehicles cause the exaggerated error seen in our model results. Note that because of this significant variance between queues on the three through lanes of this link, we were unable to find ten queue cycles where it was feasible to satisfy the constraints of a 15% travel time sample. Hence the value listed in Table 4.1 represents results for a 5% travel time penetration rate.

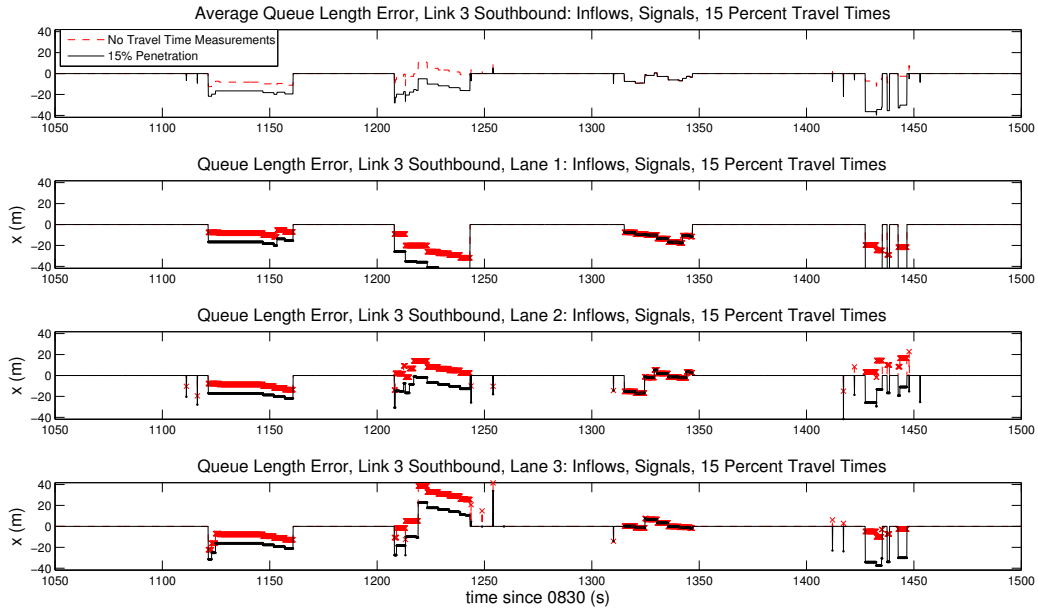

Figure 4.11: Travel time samples decreased the error for some lanes as they "tuned" the output to those lanes for specific queue cycles. However, the estimates were often made worse for the other lanes, causing increased error in the average length estimates for the affected cycles.

Chapter 5

Towards a practical signal controller with global flow objectives

Fixed-time signal controllers are typically designed with the objective of optimal local performance, considering coordination of at most a handful of neighboring intersections. Advanced adaptive systems promise more globally-beneficial performance, but almost always require the installation and upkeep of expensive communications networks to transmit high-frequency measurements and control commands back and forth between a distributed network of signals and a centralized processing unit. An ideal responsive controller would provide improvements in network-wide performance metrics without the need for central control—a distributed controller with global guarantees.

Recent theoretical research on the control of networks in various fields has derived such a controller, which goes by many names including *backpressure*, *MaxWeight*, or *max pressure*. Tassiulas and Ephremides [203] introduced this algorithm and used the theory of Lyapunov drift to prove that it stabilized a vertical queueing system in the context of controlling a multi-hop communications network. It has since been applied to various fields, including traffic signal control in [218] and [205]. While the theoretical benefits of a backpressure controller are appealing, practical signal controllers have hardware and safety limitations that prevent ideal application.

In this chapter, we introduce the historical advances in theoretical network-wide signal control from the traffic community. The modeling assumptions and control constraints commonly adopted by this community differ significantly from the requirements of a theoretical max pressure signal controller. We aim to begin rectifying these differences by extending the max pressure control algorithm to a cycle-based implementation which can take into account the green time and service rate constraints of practical arterial traffic signals.

5.1 Background: flow-impeding control on a traffic network

Section 2.2 explores the literature surrounding historical development of queueing models to represent signalized traffic flows, and mentions that many of these models were developed in the context of advancing signal control algorithms. A more detailed investigation into the relationship between theoretical control algorithms and traffic model development follows.

Control of an congested intersections

Webster's intuitive equal degree of saturation policy [212] was inherently developed with the assumption that queues at signaled intersection could be adequately serviced in their entirety with finite green periods within a reasonable cycle length. Researchers quickly recognized a need for addressing the optimal control of *saturated* intersections, where the fulfillment of delay objectives under cycle length and other service constraints were much less intuitive.

The work of Gazis [74, 73] appears to be the first discussion of the optimal control an intersection during a period of excessive congestion. A simple intersection model is derived with two competing streams governed by a signal which must split its fixed cycle length between serving each of the two streams with constraints on the minimum (and maximum) service time for each. It is concluded that the control policy which minimizes the total duration of congestion is to operate cycles which allocate maximum time to the higher demand stream at the beginning of the rush period, up to an explicitly derived "switching point" at which the lower demand stream should be given maximum in the signal cycles.

The author later expands his approach by discretizing a network of coupled (congested) intersections in a store-and-forward model [72, 52]. He formulates a problem to solve for the optimal allocation of time between "policies" representing different extremum (min/max green) service rates for each competing stream given a set of known (constant) rush-period demand. The objective is to minimize the duration of the rush period, with the ordering of the policies then selected to minimize vehicle-delay (effectively a generalization of the single-intersection case described above).

Many subsequent extensions improved the realism of Gazis' model. Singh and Tamura [189] and Lim et al. [117] extended the model to the case of oversaturated networks, where queues are constrained by a maximum storage capacity and the majority of the queues in the network are expected to be non-empty at the end of their respective service periods. Optimal signalization of this model is formulated as a linear-quadratic problem. Park et al. [166] presented optimization of a network model that is appropriate for both oversaturated and undersaturated conditions appearing on the same network.

Michalopoulos and Stephanopoulos [140, 141] additionally include constant transit delays between the intersections. As opposed to the previous analytical approach, this work presents a numerical algorithm for practically computing optimal policies—but ultimately suggests

that if too many intersections are considered, or if queues on all approaches reach or exceed the physical capacity constraint, minimization of total delay becomes infeasible.

All of the theoretical formulations mentioned up to this point were derived for single-directional links and two phase signals. They also assume a uniform "average" demand flow, and do not address the need to deal with"bursty" step-like demands that are observed at a real intersection due to the influences of upstream controllers. Furthermore, they generate pre-timed (fixed-time) signal timing plans that do not make use of the actuation and networking equipment that were becoming available during the 1960's and 1970's. In fact, the introduction of networked controllers opened a new approach to designing "on-line" coordinated signal controllers which could respond to realistic demands using closed-loop feedback control.

Dunne and Potts [60] therefore provided an early framework for actuated signal switching which enforces minimum and maximum green time constraints and minimizes some objective function which is a linear combination of intersection queue lengths. They prove that this controller is stabilizing for undersaturated intersections and provide a bound on queue length under this control (assuming approximately uniform demand).

In response, Longley [129] suggested that while the work of Gazis or Dunne and Potts is adequate for the end of a period of congestion (where demand will soon return to a undersaturated rate), it is not appropriate to deal with consistent congestion of unknown duration. He identifies two types of congestion caused by congestion:

1. *primary congestion* caused by queues at the immediate junctions being controlled
2. *secondary congestion* due to blockage of upstream junctions by primary queues

Primary congestion is unavoidable in saturated conditions, so priority should be given to minimizing the onset and effects of secondary congestion. But because the impacts of secondary congestion are beyond the jurisdiction governed by the originally congested controller, a network-aware formation of control algorithm is required. Longley then derives a controller to enforce a desired ratio of approach queues that is dependent on the geometry of potential upstream blockages. The objective of this algorithm is to minimize "queue un-balance" at a given intersection using a centralized controller where neighboring signals are coupled to minimize instabilities due to untimely platoon arrivals. A comparison of the expected average delay suggests that Longley's algorithm will outperform an "optimal" fixed time policy derived by the approach of Gazis at high degrees of saturation (in the sense of Webster) [169]. However even today, the requirement for on-line measurement of queues restricts practical applicability.

Control of networked queues

More detailed models of capacitated networks of queues arose from the mathematical and early computational sciences fields in the 1960's and 1970's [66, 24, 99]. Rather than minimizing congested periods, this literature focuses on solving for ways for maximize steady-state

flow through intersections in highly coupled networks where demands could be predicted from upstream service rates to better anticipate and compensate for the effects of secondary congestion. To apply these principles to practical signal control where frequent stop-and-go behaviors influence instantaneous link flows, the traffic community pressed for the inclusion of the kinematic wave principles of Lighthill and Whitham [116] into the network flow models being optimized [198]. It appears that little advancement was made in this direction until the end of the century, when a rise in computational power and methodologies to solve discrete integer programming problems prompted a renewed interest in model-based networkwide optimization of signals [6, 55, 215, 214].

Lo et al. [127, 126, 128] explicitly formulate a delay-minimizing signal optimization problem as an mixed-integer linear program to control a CTM representation of signalized arterials with simplified 2-phase intersections. Lin and Wang [120] optimized a signalized CTM that allows variable lengths for cycles and splits. Beard and Ziliaskopoulos contributes an explicit representation of turn movements (with permissive phases) and even more flexible timing parameters [19].

While these recent CTM-based algorithms provide solutions that are consistent with the principles of kinematic wave models and are usually tractable for off-line analysis of controllers, they still are not sufficiently efficient for informing traffic operations in real-time. Thus some researcher in the traffic community have returned to "oversimplified" vertical queueing representations, which promise optimization and control with (optimistically) polynomial-time complexity (in the number of links modeled).

A research effort led by Papageorgiou and Diakaki returned to the simplified store-and-forward models of Gazis to derive a linear-quadratic optimal feedback controller strategy for urban traffic signals known as TUC (traffic-responsive urban control) [161, 58, 57]. Later work by this group produced a model-predictive controller based on a similar simplified model [4].

Newer *spatial-queue* models such as [23] and [223] explicitly incorporate transit delays and capacitated queues. The use of these models for model-predictive control has also been proposed [22, 119].

Because communications typically limit the practical deployment of model-predictive algorithms, decentralized network-optimizing approaches to control based on the dynamics of vertical queueing models have also recently been considered [53, 218, 205]. We will discuss and extend one of these controllers, known as the *MaxWeight* or *Max Pressure* controller [205], in the remainder of this chapter.

5.2 Max pressure: a theoretical flow-maximizing controller for simple vertical queuing networks

Max pressure is a distributed network control policy derived from the concept of a "back-pressure" or "MaxWeight" controller, which was first studied in the context of routing packets

through a multi-hop communications network [203]. It has since been introduced to many other networked applications including process scheduling [50], manufacturing [202], wireless networks [13] and general stochastic networks [200]. The idea was applied to road traffic management more recently by [205] as well as [218].

The concept is intuitive: at each intersection, priority is given to the signal phase which will be able to service the most traffic given knowledge of both available upstream demand and the subsequent feasibility of downstream queues. It is a particularly attractive concept for control of a signalized urban traffic network because it can be operated in a distributed manner on local controller hardware but still provides theoretical guarantees on network-wide performance. Therefore unlike existing adaptive signal control systems such as SCOOT [100] or SCATS [188], max pressure does not require centralized communications or operations. Also, max pressure is a universal algorithm which does not require site-specific tuning for a given network geometry or expected demand set. In fact it operates with no a-priori knowledge of demand beyond a basic requirement of serviceability. This presents a huge benefit over most existing traffic control systems which require a timely and expensive re-timing process in the event of changes in demand patterns. Max pressure is also attractive to the academic community because of it's theoretical guarantee of ensuring stability in a network with simple vertical queueing dynamics. This property will be discussed in detail below.

An infinite-capacity vertical queueing framework

The properties of a theoretical max pressure traffic signal controller were originally derived in [205] on a simplified vertical queueing model in which a finite set of non-conflicting turning movements (or *phases*) can be permitted to flow simultaneously across each network node.

Consider a network of arterial roads with infinite storage capacity, modeled topologically as a graph with road links being edges and intersections being vertices. An individual link $l \in \mathcal{L}$ can be either at the entry of the network ($l \in \mathcal{L}_{\text{ent}}$) or in the interior of the network ($l \in \mathcal{L} \backslash \mathcal{L}_{\text{ent}}$). The inflow on entry links is a defined entirely by a random demand d_l, while the input flows of all other links depend on queues on upstream links and the relevant set of physical flow constraints are defined within the network. We require that each link has an exit path, or a continuous set of connected links on which vehicles can travel from the original link to eventually exit the network. Each link in the network model can have multiple *queues* corresponding to individual *movements*: all vehicles in a given queue on any link are intending to advance onto the same subsequent link (though not necessarily the same subsequent queue). We describe the dynamics of these queues as a discrete time dynamical model using the following notation:

- A *movement* (l, m) distinguishes an intention to travel from link l to link m such that $m \in Out(l)$ where $Out(l)$ is the set of links immediately downstream of l;
- A *queue* $x(l, m)(t)$ is the number of vehicles on link l waiting to enter link m at timestep

t, and $X(t)$ is the set (vector or matrix) of all the queue lengths on the network at timestep t;

- A *capacity* $c(l, m)$ is the expected number of vehicles that can travel from link l to link m per time step given maximum demand for the queue $x(l, m)$, and $C(l, m)(t)$ is the *realized saturation flow* at time t;
- The *turn ratio* $r(l, m)$ is the expected proportion of vehicles that are leaving l which are intending to enter m, and $R(l, m)(t)$ is the *realized turn proportion* at t;
- The *demand vector* d of dimension $|\mathcal{L}_{\text{ent}}|$ specifies demands at network entry links;
- The *flow vector* f of dimension $|\mathcal{L}|$ denotes flows on all links of the network such that f_l is the flow in link l.

Note that flow on an entry link is dependent only on demand d_l, and flow on internal links is dependent on routing proportions and the flow on upstream links:

$$f_l = \begin{cases} d_l & l \in \mathcal{L}_{\text{ent}} \\ \sum_m f_m r(m, l) & l \in \mathcal{L} \backslash \mathcal{L}_{\text{ent}} \end{cases} \tag{5.1}$$

Hence there is necessarily a linear relationship between the observed link flows f and the boundary demand d:

$$f = dP \tag{5.2}$$

where matrix P depends only on observed routing proportions within the network (but is not necessarily unique).

In this framework, a road intersection is modeled as a node. Controllers (traffic signals) are placed at every node to limit the set of queues permitted to discharge at any given time. A set of movements that can be simultaneously actuated without flow conflicts is called a *phase*. Each permissible phase for a given intersection can be represented as a binary control matrix S that is defined as follows:

$$S(l, m) = \begin{cases} 1 & \text{if movement } (l, m) \text{ is activated} \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

Denote U_n as the known finite set of permissible control matrices for node n. (Note that the subscript n is often dropped for ease of notation.)

Practically, only one phase can be actuated at any point in time: at each model time step t, a single control matrix $S(t)$ encodes which set of queues approaching the intersection are permitted to discharge during that time step. The selection of such a controller can be based on feedback representing the queue state of the network queues at a previous time step.

The evolution of network queue lengths $X(t)$ contained at the nodes of this network can be seen as a Markov chain: the state at time $(t+1)$ is a function of only the state at time t and external demand d,

$$X(t+1) = F(X(t), d) \tag{5.4}$$

Define $[\, a \wedge b \,] := \min\{a, b\}$. To describe queue dynamics explicitly, we must make a distinction between entry links and internal links:
if $l \in \mathcal{L}_{\text{ent}}$,

$$x(l,m)(t+1) = x(l,m)(t) + d_l(t+1) - [C(l,m)(t+1)S(l,m)(t+1) \wedge x(l,m)(t)] \tag{5.5}$$

and if $l \in \mathcal{L} \backslash \mathcal{L}_{\text{ent}}$,

$$x(l,m)(t+1) = x(l,m)(t) + \sum_k [C(k,l)(t+1)S(k,l)(t+1) \wedge x(k,l)(t)]R(l,m)(t+1) \\ - [C(l,m)(t+1)S(l,m)(t+1) \wedge x(l,m)(t)] \tag{5.6}$$

We focus on networks for which the boundary inflow demands $d = (d_l)_{(l \in \mathcal{L}_e)}$ are *feasible*—that is, the network is servicing a distribution of inflows for which it is possible to find a controller that allows *in average* more departures than arrivals at each link.

Define $conv(U)$ to be the convex hull of the set of permissible control matrices U. The following properties are then shown in [205]:

Property 5.1. *A matrix M is in $conv(U)$ iff $\exists$ a sequence of control matrices*

$$\overline{S} = \{S(1), S(2), ..., S(t), ... | S(\cdot) \in U\}$$

such that $\forall (l,m)$

$$M(l,m) = \liminf_T \frac{1}{T} \sum_{t=1}^{T} S(l,m)(t) \tag{5.7}$$

The element $M(l,m)$ in (5.7) can be interpreted as the long-term average proportion of intersection capacity given to movement (l,m). Then define $M_{\overline{S}}$ to be the specific long-term control proportion matrix constructed as in (5.7) using the specific control sequence $\overline{S} = \{S(1), S(2), ..., S(t), ...\}$.

Property 5.2. *A demand d is* feasible *if and only if $\exists\ M_{\overline{S}} \in conv(U)$ and $\varepsilon > 0$ such that*

$$c(l,m)M_{\overline{S}}(l,m) > f_l r(l,m) + \varepsilon. \tag{5.8}$$

where $f = dP$ as in (5.2).

Define D^0 to be the set of all average demand vectors $d = \{d_l\}$ that satisfy (5.8) and are therefore feasible.

Mathematical formulation of the max pressure controller

Consider a weight assigned to each queue (l, m) as a function of all network queue lengths X:

$$w(l,m)(X(t)) = x(l,m)(t) - \sum_{p \in Out(m)} r(m,p)x(m,p)(t) \tag{5.9}$$

where $Out(m)$ is the set of all links receiving flow from link m. The *pressure* $\gamma(S)$ that is potentially alleviated by a control action S at time step t is defined as follows:

$$\gamma(S)(X(t)) = \sum_{l,m} c(l,m)w(l,m)(X(t))S(l,m)(t) \tag{5.10}$$

At each time step t, the standard max pressure controller $u^*(X(t))$ explicitly choses the phase $S^* \in U$ that maximizes $\gamma(S)(X(t))$:

$$S^*(t) = u^*(X(t)) = \arg\max\{\gamma(S)(X(t)) | S \in U\} \tag{5.11}$$

Network stability

A network controller is defined to be *stabilizing* if its application ensures that the mean length of all queues in the network remain bounded for any arbitrary time horizon T.

Definition 5.1 (**Network stability**). A network is *stable* if the following quantity is bounded:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big\{|X(t)|_1\Big\} \tag{5.12}$$

where $|X|_1 = \sum_{l,m} |x(l,m)|$.

This concept is relevant to many applications such as communications networks [76, 152, 160] or industrial systems [62, 36, 103]. In traffic networks, excessive or rapidly accumulating queues imply oversaturation and the associated impacts of secondary congestion.

The following stability result for the standard max pressure controller has been proven in [205]:

Theorem 5.3. *The max pressure control u^* is stabilizing in a network with state dynamics described by* (5.5)-(5.6) *whenever the average demand vector $d = \{d_l\}$ is within the set of feasible demands D^0.*

This theoretical guarantee is one of the many attractive qualities of max pressure for controlling vehicular traffic in urban road networks. A proof of Theorem 5.3 was originally published in [205]; it is re-written in Appendix B of this thesis for reference.

5.3 The cycle-based max pressure controller (Cb-MP)

The max pressure controller as originally formulated is not practical for application on a signalized traffic network for three reasons:

1) it does not account for capacity reductions (lost time) due to excessive signal switching,

2) it cannot enforce coordination between subsequent intersections for purposes of maximizing flow continuity, and

3) it does not provide guarantees that low-demand queues will be served within a finite time period.

These limitations motivate a new extension of the max pressure algorithm which bounds signal switches and can maintain timed cyclical behaviors for signal coordination and queue service equity. While a similar concept was suggested in [205], the current work further extends a simple proportional phase controller to allow model dynamics to explicitly act at a faster rate than the controller update period. We then extend the stability proof of [205] to prove that our *cycle-based max pressure* (Cb-MP) controller still provides the desired guarantee of queue stability with a penalty to the theoretical bound on queue lengths due to the decreased rate of controller update.

In the following section, we define a new *cycle-based max pressure* (Cb-MP) controller which bounds the number of signal switches per fixed time period, provides capacity for standard signal coordination methods, and can easily guarantee a minimum service rate for all intersection phases.

For safety reasons, an intersection controller cannot switch signal phase actuation immediately. Instead, it must incorporate a pause of $R \approx 2.5$ seconds in which all signal phases have a red light. This *clearance time* allows all vehicles in the previously actuated phase to clear the intersection before a conflicting phase can be permitted to use the intersection. In the standard formulation of max pressure, the controller chooses an appropriate action based feedback received at every time step of the modeled dynamics. To accurately capture queuing behaviors observed on arterial roadways, a model would need to operate with a time discretization of $\Delta t < 10$ seconds. A signal switch at every time step could therefore result in more than 25% loss of intersection service capacity, which is not considered in the theoretical examination presented in [205].

Selection of cycle length

In this new formulation we explicitly specify the number of signal switches that occur in a fixed number of model time steps using the familiar concept of a *signal cycle*. As typical with modern traffic lights, a signal operating the Cb-MP algorithm rotates through all available signal phases within a small fixed time period. We define *cycle time* τ as a predefined number of model time steps and require that each controller phase S must be green for some

proportion $\lambda_S \geq \kappa_S$ of the τ steps, where the *minimum green splits* $\kappa_S \in (0,1)\ \forall\ S \in U$ are parameters selected by a network manager to enforce equity in movement actuation.

The selection of a cycle length τ intuitively affects intersection capacity. Our proof of network stability in the following sections relies on the fact that road links are *undersaturated*: that is, the expected demand is served (on average) within a signal cycle. To avoid link saturation, we pose the following convex optimization problem (extended from that in [9]) to determine minimum constrained feasible actuation time Λ^*:

$$\begin{aligned} \Lambda^* = \min_{\lambda=\{\lambda_S\}} \quad & \sum_{S\in U} \lambda_S \\ \text{subject to} \quad & \lambda_S \geq \kappa_S\ \forall\ S \in U \\ & f_l r(l,m) < \sum_S \lambda_S c(l,m) S(l,m) \end{aligned} \tag{5.13}$$

where $\kappa_S \in [0,1]\ \forall S \in U$ and $\sum_S \kappa_S < 1$. If $\Lambda^* > 1$, the demand is not feasible under the set of $\{\kappa_S\}$ for any cycle length. If $\Lambda^* < 1$, then we can define a cycle length for which flow is admissible without link saturation. However, this cycle length τ must be significantly greater than Λ^* to account for clearance times. If we define $L = \lceil (\frac{R}{\Delta t} \cdot |U|) \rceil$ to be the total number of *lost time steps* per cycle, a feasible cycle length τ must satisfy the following condition:

$$\tau > \frac{L}{1-\Lambda^*} \tag{5.14}$$

A relaxed signal controller

Furthermore, define a *relaxed controller* as a matrix S^r with each element $S^r(l,m)$ representing the fraction of the operational time steps that are allocated to the movement $\{l,m\}$:

$$S^r(l,m) = \lambda_{l,m} \in [0,1] \tag{5.15}$$

Such a relaxed controller can be seen as a convex combination of all possible control matrices, $S^r = \sum_{S\in U} \lambda_S S$.

Given an appropriate τ which satisfies (5.14), the cycle-based max pressure controller is a relaxed control matrix that is constructed as follows:

$$S^{r*}(t) = u^{c*}(X(t)) = \sum_{S\in U} \lambda_S^* S, \quad \text{where} \tag{5.16}$$

$$\{\lambda_S^*\} = \arg\max_{\lambda_1,\ldots,\lambda_{|U|}} \sum_{S\in U} \lambda_S \gamma(S)\big(X(\lfloor t/\tau \rfloor)\big) \tag{5.17}$$

$$\text{subject to} \quad \lambda_S \geq \kappa_s, \quad \sum_S \lambda_S \leq 1 - \tfrac{L}{\tau}$$

At time step $t = n\tau$ for integer n, the controller u^{c*} uses feedback measurements $x(t)$ to select a relaxed control matrix S^{r*} with components λ_S^* that satisfy (5.17). This relaxed

controller is then applied for the subsequent τ time steps $\{t, t+1, \ldots, t+\tau-1\}$ before the controller is updated.

Note that this controller is modeled such that all phases in an intersection are simultaneously actuated at some proportion of their maximum flow capacity. This is not possible in practice, as many phases will have to make conflicting use of the same intersection resources. Hence individual phases S will have to be actuated in series, with each having a duration corresponding to a number of "time units" that are equal to cycle proportions ($\lambda_S \tau \cdot \Delta t$). Feedback measurements will then be a measure of "average" cycle queue length acquired over a set of measurements spanning the previous cycle. If the solution for (5.17) is not unique, one of the optimal solutions is chosen at random using a uniform probability distribution (or according to some practical actuation priority criteria chosen by the network manager). Because it can be implemented such that phases occur in a predictable order, a controller running Cb-MP can be synchronized with neighboring controllers to enforce a "green-wave progression" as is standard practice in existing traffic signal control design.

Stability of Cb-MP

The Cb-MP controller formulated in (5.16)-(5.17) is fundamentally different from the standard max pressure formulation in [205] in two ways: first, it only updates the controller once every signal cycle (or τ model time steps); second, it applies a relaxed phase actuation (which is some convex combination of standard phase actuations). In this section, we address how each of these modifications impact the resulting network dynamics, and ultimately show that the application of Cb-MP yields a similar stability guarantee to that shown by Varaiya for the standard max pressure controller given slightly weaker conditions on demand flow.

Define $conv_\kappa$ as the set of convex combinations of control matrices with coefficients larger than κ:

$$conv_\kappa = \Big\{ \sum_S \lambda_S S \Big| \ \lambda_S > \kappa_S \ \forall S \in U \Big\} \tag{5.18}$$

Also define a set of *undersaturated* admissible demands D_κ with elements d such that $f = dP$ and

$$f_l r(l,m) < c(l,m) S^r(l,m) \tag{5.19}$$

This condition (also seen in (5.13)) ensures that a demand $d \in D_\kappa$ can in average be served *within a single cycle* by a relaxed control matrix that maintains a specified minimum time allocation for each phase.

Theorem 5.4. *The cycle-based max pressure controller defined in* (5.16)-(5.17) *stabilizes a network whenever the demand is within a set of feasible undersaturated demands* D^κ.

The remainder of this section proves Theorem 5.4 by finding a bound on (5.12) given a cycle-based max pressure controller. The structure of this proof is as follows:

1) First, we introduce the concept of a τ-*updated* controller and we show that switching control only once every τ time steps does not impact summarize the mathematical structure used in [205] to derive a bound for the expected network state (5.12).

2) Next we define an intermediate "relaxed max pressure" formulation to demonstrate the impacts of expanding the domain of control actions to relaxed controllers which are convex combinations of allowable phase matrices.

3) We then demonstrate the intra-cycle queue dynamics given a τ-updated relaxed controller.

4) We combine the above steps to show that queue stability holds given a Cb-MP controller with both relaxed actuation and τ-updating.

5) Finally, we compare the our Cb-MP queue length bounds to those originally derived in [205] to illustrate the increase due to cycle-updating.

Properties of a τ-updated controller

Suppose that we impose that the control actuation $S^*(t)$ can only be updated every τ model time steps. A resulting τ-*updated control sequence* is composed of a single control matrix repeated for τ time steps of the model dynamics:

$$S(n\tau + 1) = S(n\tau + 2) = \ \ldots \ = S\big((n+1)\tau\big) \tag{5.20}$$

First, we prove that the set of demands that can be accommodated using τ-updated control sequences is the same set of feasible flows as in (5.8). This equivalence becomes intuitive when one considers that our definition of feasible flows considers only the long-term (more precisely, infinite-term) average of both demand and service rates, and any infinite control sequence with limited admissible phases can be re-arranged to form a τ-updated sequence for some τ.

Lemma 5.5. *All flows which satisfy Property 5.2 given a controller u updated at every model time step will also satisfy Property 5.2 with a τ-updated controller for some τ.*

Proof. Given the set admissible phases U, define:

- $\mathcal{U}$ is the set of control sequences with distinct elements:

$$\mathcal{U} = \{S(1), S(2) \ldots S(t) \ldots | S(\cdot) \in U\}$$

- $\mathcal{U}_\tau$ is the set of τ-updated control sequences:

$$\mathcal{U} = \{S(1), S(1), \ldots, S(\tau+1), S(\tau+1), \ldots, S(n\tau+1), S(n\tau+1), \ldots | S(\cdot) \in U\}$$

Also define the following sets of *long-term control proportion matrices*, which are similar to the formulation in (5.7):

$$M_{\mathcal{U}} = \left\{ \liminf_T \frac{1}{T} \sum_{t=1}^{T} S(t) \middle| \{S(1), S(2), \ldots, S(t), \ldots\} \in \mathcal{U} \right\}$$

$$M_{\mathcal{U}_\tau} = \left\{ \liminf_T \frac{1}{T} \sum_{t=1}^{T} S(t) \ \cdot \middle| \{S(1), S(1), \ldots, S(\tau+1), S(\tau+1), \ldots\} \in \mathcal{U}_\tau \right\}$$

By Property 5.2, a demand d is only feasible if there exists a control sequence $\overline{S}$ such that the corresponding long-term control proportion matrix $M_{\overline{S}}$ satisfies (5.8). Here we show $M_{\mathcal{U}} = M_{\mathcal{U}_\tau}$, and therefore any flows that are admissible given an unrestricted controller in $\mathcal{U}$ can also be accommodated using a τ-updated controller in $\mathcal{U}_\tau$.

Obviously, $M_{\mathcal{U}_\tau} \subset M_{\mathcal{U}}$. To show equality, we must also demonstrate that $M_{\mathcal{U}} \subset M_{\mathcal{U}_\tau}$. Suppose there exists a control sequence $\hat{S} = \{S(1), S(2), \ldots\} \in \mathcal{U}$. By definition,

$$\begin{aligned} M_{\hat{S}} &= \liminf_T \frac{1}{T} \sum_{t=1}^{T} S(t) = \liminf_T \frac{1}{\tau T} \sum_{t=1}^{\tau T} \tilde{S}(t) \quad \text{where } \tilde{S} = \{S(1), S(1), \ldots, S(t), S(t), \ldots\} \\ &= \liminf_T \frac{1}{T} \sum_{t=1}^{T} \tilde{S}(t) \in M_{U_\tau} \\ &\implies M_{\mathcal{U}} \subset M_{\mathcal{U}_\tau} \end{aligned}$$

□

As will be shown in the subsequent proof, occasional updating will also lead to an increased bound on queue lengths relative to the standard max pressure setting.

Formulating a queue bound

Our ultimate goal is to derive a bound for the average expected queue state (5.12). The approach taken in this work follows that of [205]: we bound the incremental model-step queue increase $|X(t+1) - X(t)|$ and then recursively compute a bound on average queue lengths $\sum_{t \in [0,T]} \mathbb{E}\{|X(t)\}$ for an arbitrary time horizon T.

Begin by considering the expectation of the following function of queue state perturbation conditioned on the past queue state:

$$\begin{aligned} |X(t+1)|^2 - |X(t)|^2 &= |X(t) + \delta(t)|^2 - |X(t)|^2 = 2X(t)^T \delta(t) + |\delta(t)|^2 \\ &= 2\alpha(t) + \beta(t) \end{aligned} \tag{5.21}$$

with $\delta(t) = X(t+1) - X(t)$, $\alpha(t) = X(t)^T \delta(t)$, and $\beta(t) = |\delta(t)|^2$. We continue by addressing bounds on $\beta(t)$ and $\alpha(t)$ separately.

First we consider $\beta(t) = |\delta(t)|^2$.

Lemma 5.6.

$$\beta(t) = \left|\delta(t)\right|^2 \leq NB^2 \tag{5.22}$$

where $B = \max\left\{\overline{C}(l,m), \sum_k \overline{C}(k,l), \overline{d}(l,m)\right\}$, *$N$ is the number of queues in the network,* $\overline{C}(l,m)$ *is the maximum value of the random service rate* $C(l,m)(t)$, *and* $\overline{d}(l,m)$ *is the maximum value of random demand* $d(l,m)$.

The proof of Lemma 5.6 is exactly the same as the bound on $\beta(t)$ presented in Appendix B and will therefore not be repeated here. Note that because these bounds hold for any arbitrary $S(l,m)(t) \in [0,1]$, the original bound on $\mathbb{E}\Big\{\beta(t)\Big\}$ is trivially extended to any convex combination of control matrices; hence it is still valid in our extension.

Now we examine a bound on $\alpha(t) = X(t)^T\delta(t)$. Again following [205], we define additional sub-terms:

$$\begin{aligned}
&\mathbb{E}\{\alpha(t)|X(t)\} \qquad (5.23)\\
&\quad = \sum_{l\in\mathcal{L},m} w(l,m)(t)\left[f_l r(l,m) - \mathbb{E}\Big\{\big[C(l,m)(t+1)S(l,m)(t) \wedge x(l,m)(t)\big]\big|X(t)\Big\}\right]\\
&\quad = \alpha_1(t) + \alpha_2(t)
\end{aligned}$$

with

$$\alpha_1(t) = \sum_{l\in\mathcal{L},m} \left[f_l r(l,m) - c(l,m)S(l,m)(t)\right] w(l,m)(t) \tag{5.24}$$

and

$$\alpha_2(t) = \sum_{l\in\mathcal{L},m} S(l,m)(t)w(l,m)(t)\left[c(l,m) - \mathbb{E}\Big\{\big[C(l,m)(t+1) \wedge x(l,m)(t)\big]\big|X(t)\Big\}\right] \tag{5.25}$$

Lemma 5.7. *For all l, m, t,*

$$\alpha_2(t) \leq \sum_{l\in\mathcal{L},m} c(l,m)\overline{C}(l,m) \tag{5.26}$$

The proof of Lemma 5.7 again directly follows that presented in Appendix B; an extension from a binary controller $S \in \{0,1\}$ to a relaxed controller $S^r \in [0,1]$ is trivial.

In fact, the extension made here only affects the $\alpha_1(t)$ term. To demonstrate a bound on $\alpha_1(t)$ given application of a cycle-based max pressure controller u^{c*}, we first examine the stability of a standard max pressure controller using relaxed controllers with minimum phase proportion constraints and a stricter limitation on network demands. We will then show that a τ-updated cycle-based max pressure controller also stabilizes a network, but results in an increase in queue length bounds that is proportional to cycle length τ.

Impact of a relaxed controller

Define an intermediate "relaxed max pressure" policy in which relaxed controllers are applied at the standard max pressure update rate (once per time step of the model dynamics). This situation was suggested in [205] to simulate enforcing minimum phase proportions in a cycle formulation of max pressure. Yet this proposal unrealistically models "cycle" updates at the same rate as the model of queueing and discharging behaviors (hence the introduction of the τ-updated formulation in this work). Nonetheless, we use this intermediate formulation to demonstrate that queue stability is still achieved upon use of a relaxed controller.

Lemma 5.8. *If a "relaxed" max pressure control policy S^{r*} is updated and applied at each time step t and the demand d is in the set of feasible undersaturated demands D^{κ}, then there exists an $\varepsilon > 0$, $\eta > 0$ such that*

$$\alpha_1(t) \leq -\varepsilon\eta\big|X(t)\big| \tag{5.27}$$

Proof. Consider the relaxed max pressure control matrix S^{r*} defined in (5.16) for $\tau = 1$. By construction, $\forall\ S^r \in conv_{\kappa}$

$$\sum_{l,m} c(l,m)w(l,m)(X(t))S^r(l,m) \leq \sum_{l,m} c(l,m)w(l,m)(X(t))S^{r*}(l,m) \tag{5.28}$$

with equality only if $S^r = S^{r*}$. Thus $\forall\ (S^r \in conv_{\kappa}) \neq S^{r*}$,

$$\sum_{l,m} \big[f_l r(l,m) - c(l,m)S^{r*}(l,m)(t)\big]w(l,m)(X(t)) < \sum_{l,m} \big[f_l r(l,m) - c(l,m)S^{r}(l,m)\big]w(l,m)(X(t)) \tag{5.29}$$

If the demand flow is admissible according to (5.19), then $\exists\ \hat{S} \in conv_{\kappa}$ and some small $\varepsilon > 0$ such that

$$c(l,m)\hat{S}(l,m) = \begin{cases} f_l r(l,m) + \varepsilon & \text{if } w(l,m)(X(t)) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned} \alpha_1(t) &= \sum_{l,m} [f_l r(l,m) - c(l,m)S^{r*}(l,m)(t)]w(l,m)(X(t)) \\ &< \sum_{l,m} \big[f_l r(l,m) - c(l,m)\hat{S}(l,m)(t)\big]w(l,m)(X(t)) \\ &= -\varepsilon \sum_{l\in\mathcal{L},m} \max\{w(l,m)(X(t)),0\} + \sum_{l\in\mathcal{L},m} f_l r(l,m)\min\{w(l,m)(X(t)),0\} \end{aligned} \tag{5.30}$$

We assume that by our choice of $\hat{S}$, $f_l r(l,m) > \varepsilon$. Hence $\alpha_1(t) < -\varepsilon\sum_{l,m} w(l,m)(t)$. Given the linearity of (5.9) and the known properties of $r(l,m)(t)$, it can be show that $\sum_{l,m} w(l,m)(t) \geq \eta|X(t)|$ for some $\eta > 0$. This completes the derivation of (5.27). □

For ease of notation, now combine (5.22), (5.26) and (5.27) to obtain the following expression given application of the "relaxed max pressure" controller:

$$\mathbb{E}\Big\{|X(t+1)|^2 - |X(t)|^2 | X(t)\Big\} = \mathbb{E}\Big\{2\alpha(t) + \beta(t)\Big\}$$
$$< -2\varepsilon\eta\,|X(t)| + 2\sum_{l\in\mathcal{L},m}[c(l,m)\overline{C}(l,m)] + NB^2 \qquad (5.31)$$

where N and B are as in (5.22).

Intra-cycle queue bound

Next we establish a bound on queue growth in a single time step between controller updates.

Lemma 5.9. *Assuming a cycle-based max pressure controller with an cycle steps τ beginning at time t, the following bound on state perturbation holds for all* steps since update $p \in [0, \tau - 1]$*:*

$$\mathbb{E}\Big\{|X(t+p+1)|^2 - |X(t+p)|^2 \big| X(t)\dots X(t+p)\Big\} < Y + h(p) - 2\varepsilon\eta|X(t+p)| \qquad (5.32)$$

$$\textit{for} \quad Y = 2\sum_{l,m} c(l,m)\overline{C}(l,m) + NB^2 \quad \textit{and} \qquad (5.33)$$

$$h(p) = 2pNB\Big(\varepsilon\eta + \sum_{l,m}\big[f_l r(l,m) + c(l,m)\big]\Big) \qquad (5.34)$$

Proof. As in Lemmas 5.6-5.8 above, begin by dividing the argument of (5.32) into three parts: $|X(t+p+1)|^2 - |X(t+p)|^2 = 2(\alpha_1(t+p) + \alpha_2(t+p)) + \beta(t+p)$, where β, α_1 and α_2 are quantities that depend on the controller applied at $(t+p)$:

$$\beta(t+p) = |X(t+p+1) - X(t+p)|^2 \qquad (5.35)$$

$$\alpha_1(t+p) = w(l,m)(X(t+p)) \cdot \sum_{l,m}\Big(f_l r(l,m) - c(l,m)S(l,m)(t)\Big) \qquad (5.36)$$

$$\alpha_2(t+p) = w(l,m)(X(t+p))\cdot \qquad (5.37)$$
$$\sum_{l,m}\Big(c(l,m)S(l,m)(t) - \mathbb{E}\Big\{\big[C(l,m)(t+p+1) \wedge x(l,m)(t+p)\big]\big|X(t+p)\Big\}\Big)$$

Bounds on the expectations of $\beta(\cdot)$ and $\alpha_2(\cdot)$ were previously established for any binary or relaxed control matrix in (5.22) and (5.26), respectively. Thus we already know that:

$$\mathbb{E}\Big\{|X(t+p+1)|^2 - |X(t+p)|^2 \big| X(t)\dots X(t+p-1)\Big\} \qquad (5.38)$$
$$< 2\sum_{l,m} c(l,m)\overline{C}(l,m) + NB^2 + \mathbb{E}\Big\{2\alpha_1(t+p)\Big\}$$

The remainder of the bound proposed in (5.32) originates from the $2\alpha_1(t+p)$ term, which is directly dependent on the explicit form of the controller S. Rewrite $2 \cdot \alpha_1$ from (5.36) as follows:

$$\begin{aligned}
&2\sum_{l,m} w(l,m)(X(t+p))[f_l r(l,m) - c(l,m)S(l,m)(t)] \\
&\qquad = 2\sum_{l,m} w(l,m)(X(t))[f_l r(l,m) - c(l,m)S(l,m)(t)] \qquad (5.39)\\
&\qquad\qquad + 2\sum_{l,m} \Big\{ w(l,m)\Big(X(t+p) - X(t)\Big) \cdot [f_l r(l,m) - c(l,m)S(l,m)(t)] \Big\} \\
&\qquad = \xi_1(t,p,S) + \xi_2(t,p,S)
\end{aligned}$$

$$\text{for} \quad \xi_1(t,S) = 2\sum_{l,m} w(l,m)(X(t))[f_l r(l,m) - c(l,m)S(l,m)(t)] \qquad (5.40)$$

$$\text{and} \quad \xi_2(t,p,S) = 2\sum_{l,m} \Big\{ w(l,m)\Big(X(t+p) - X(t)\Big) \cdot [f_l r(l,m) - c(l,m)S(l,m)(t)] \Big\} \qquad (5.41)$$

By Lemma 5.8 we know that $\xi_1(t,S) < -2\varepsilon\eta|X(t)|$. Because $|X(t)| = |X(t+p) - (X(t+p) - X(t))| > |X(t+p)| - |X(t+p) - X(t)|$, we find that

$$\begin{aligned}
\xi_1(t,p,S) &< -2\varepsilon\eta\big(|X(t+p)| - |X(t+p) - X(t)|\big) \\
&< -2\varepsilon\eta|X(t+p)| + 2\varepsilon\eta\sum_{i=1}^{p} |X(t+i) - X(t+i-1)| \\
&= -2\varepsilon\eta|X(t+p)| + 2\varepsilon\eta\sum_{i=1}^{p} |\delta(t+i-1)| \qquad (5.42)
\end{aligned}$$

So by (5.42) and (5.22),

$$\begin{aligned}
\xi_1(t,S) &< -2\varepsilon\eta|X(t+p)| + 2\varepsilon\eta p\sum_{l,m} \max\left\{ \overline{C}(l,m), \sum_k \overline{C}(k,l), \overline{d}(l,m) \right\} \\
&= 2\varepsilon\eta \cdot \Big(pNB - |X(t+p)|\Big) \qquad (5.43)
\end{aligned}$$

To bound ξ_2, we study the term

$$\begin{aligned} w(l,m)(X(t+p)) - w(l,m)(X(t)) &= \sum_{n=1}^{p} w(l,m)(X(t+n)) - w(l,m)(X(t+n-1)) \\ &= \sum_{n=1}^{p} \Big\{ x(l,m)(t+n) - x(l,m)(t+n-1) \\ &\qquad - \sum_{s \in Out(m)} [x(m,s)(t+n) - x(m,s)(t+n-1)] r(m,s) \Big\} \\ &= \sum_{n=1}^{p} w(l,m)(\delta(t+n-1)) \end{aligned} \tag{5.44}$$

By (5.22) and the fact that $w(\cdot)$ is linear,

$$|w(l,m)(\delta(t+n-1))| < NB \tag{5.45}$$

Plugging (5.45) back into the definition of ξ_2, we obtain

$$\begin{aligned} \xi_2(t,p,S) &= 2\Big(\sum_{l,m} [f_l r(l,m) - c(l,m) S(l,m)(t)] \cdot \sum_{n=1}^{p} w(l,m)(\delta(t+n-1)) \Big) \\ &< 2 \sum_{n=1}^{p} \sum_{l,m} [f_l r(l,m) - c(l,m) S(l,m)(t)] \cdot \sum_{u,v} \max \Big\{ \overline{C}(u,v), \sum_{k} \overline{C}(k,u), \overline{d}(u,v) \Big\} \\ &= 2NB \sum_{n=1}^{p} \sum_{l,m} [f_l r(l,m) - c(l,m) S(l,m)(t)] \end{aligned} \tag{5.46}$$

Also note that

$$\Big| \sum_{n=1}^{p} \sum_{l,m} [f_l r(l,m) - c(l,m) S(l,m)(t)] \Big| < p \sum_{l,m} [f_l r(l,m) + c(l,m)] \tag{5.47}$$

so (5.46) becomes

$$\xi_2(t,p,S) < 2NBp \cdot \Big(\sum_{l,m} [f_l r(l,m) + c(l,m)] \Big) \tag{5.48}$$

Substituting (5.43) and (5.48) into (5.38) yields (5.32). □

Long-term network queue bound

Given Lemmas 5.6-5.9, we show that for a time t within any number K of τ-updated cycles, the following quantity is bounded:

$$\begin{aligned}\sum_{t=1}^{K\tau}\mathbb{E}\Big\{|X(t+1)|^2-|X(t)|^2|X(t)\Big\} &= \sum_{t=1}^{K-1}\sum_{p=0}^{\tau-1}\mathbb{E}\Big\{|X(t+p+1)|^2-|X(t+p)|^2|X(t+p)\Big\}\\ &< \sum_{t=1}^{K-1}\sum_{p=0}^{\tau-1}(Y+h(p)-2\varepsilon\eta|X(t+p)|)\\ &< -2\varepsilon\eta\sum_{t=1}^{K\tau}|X(t)|+(K-1)\Big(\tau Y+\sum_{p=0}^{\tau-1}h(p)\Big) \qquad (5.49)\end{aligned}$$

which, when taking the expectation, yields

$$\mathbb{E}\Big\{|X(K\tau+1)|^2-|X(1)|^2\Big\} < -2\varepsilon\eta\sum_{t=1}^{K\tau}\mathbb{E}\Big\{|X(t)|\Big\}+(K-1)(\tau Y+\sum_{p=0}^{\tau-1}h(p)) \qquad (5.50)$$

Rearranging gives

$$\begin{aligned}\frac{1}{K\tau}\sum_{t=1}^{K\tau}\mathbb{E}\Big\{|X(t)|\Big\} &< \frac{1}{2\varepsilon\eta K\tau}\mathbb{E}\Big\{|X(1)|^2-|X(K\tau+1)|^2\Big\}+\frac{\tau-1}{2\varepsilon\eta K\tau}\Big(\sum_{p=0}^{\tau-1}h(p)+\tau Y\Big)\Big]\\ &< \frac{1}{2\varepsilon\eta\ K\tau}\mathbb{E}\Big\{|X(1)|^2\Big\}+\frac{1}{2\varepsilon\eta\tau}\left(\sum_{p=0}^{\tau-1}h(p)+\tau Y\right) \qquad (5.51)\end{aligned}$$

By (5.12), the bound

$$2\varepsilon\eta\frac{1}{K\tau}\sum_{t=1}^{K\tau}\mathbb{E}\Big\{|X(t)|\Big\} < \tfrac{1}{K\tau}\mathbb{E}\Big\{|X(1)|^2\Big\}+\frac{1}{\tau}\sum_{p=0}^{\tau-1}h(p)+Y \qquad (5.52)$$

establishes that the cycle-based max pressure controller $u^{c*}(X(t))$ defined in (5.16) will stabilize a vertical queueing network with dynamics $X(t)$ as in (5.5)-(5.6).

Increase in queue bounds

The following bound on network queue state for a standard max pressure controller is derived in Appendix B (for Y as in (5.33)):

$$2\varepsilon\eta\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big\{|X(t)|\Big\} < \frac{1}{T}\mathbb{E}\Big\{|X(1)|^2\Big\}+Y \qquad (5.53)$$

Notice that by comparison between (5.53) and (5.52), the bound on the long-term sum of expected network queues in cycle-based max pressure is larger by a term that increases linearly in cycle length τ:

$$\frac{1}{2\varepsilon\eta\tau}\sum_{p=0}^{\tau-1}h(p) = (\tau-1)NB\Big(1+\frac{1}{\varepsilon\eta}\sum_{l,m}\big[f_l r(l,m)+c(l,m)\big]\Big) \tag{5.54}$$

5.4 Numerical implementation of Cb-MP

To demonstrate the effectiveness of a realistic implementation of max pressure, cycle-based max pressure controller was implemented on a network of 11 signalized intersections modeled in the Aimsun, a commonly-used micro-simulation platform. The model was generated as part of the I-15 Integrated Corridor Management project undertaken by the San Diego Association of Governments in San Diego, CA. Demand and other model parameters are calibrated to match the morning peak period (5:00 AM to 10:00 AM).

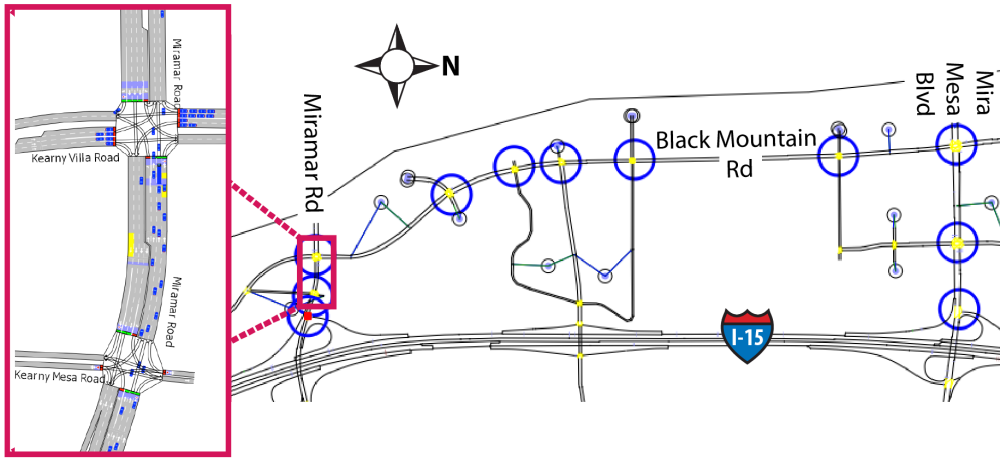

Figure 5.1: The chosen network was calibrated to represent realistic demands and physical parameters observed on a stretch of Black Mountain Road near the I-15 freeway in San Diego, California.

This section of road is currently controlled using an offset-optimized actuated-coordinated control scheme. Under this system, each signal operates with a fixed cycle time of 100 seconds and a fixed phase ordering, but uses instantaneous feedback of intersection vehicle approaches to adjust cycle green splits (effectively λ_i) within fixed minimum and maximum green time constraints per cycle. This control algorithm is coded and calibrated in the Interstate-15 Aimsun model to represent realistic conditions and was therefore deemed an appropriate benchmark for performance comparison to Cb-MP.

Six variations of Cb-MP were implemented. First, a version with a cycle length of 100 seconds and minimum green time constraints of 10 seconds for each available signal phase was

used to closely match the operational constraints of the existing fully-actuated controller. The relative offsets for the southbound coordination phases in this implementation were the same as those used in the actuated-coordinated system. We then ran variations which extended the cycle time for Cb-MP to 120, 140, 160, 180, and 200 seconds to demonstrate the effect of increased cycle time τ on observed queue lengths.

To compare performance, we calculated vehicle service rates, average delay, average number of stops and stopped time, and mean and maximum queue lengths that were modeled using each controller. These metrics were only calculated for vehicles and links corresponding to the southbound direction on Black Mountain Road as well as the short connections to the I-15 freeway on westbound Mira Mesa Blvd and eastbound Mirimar Rd. This pathway simulates a viable "freeway-alternative" in the congested direction during the morning peak period. During implementation, the Cb-MP algorithm most often chose to give actuation priority to this high-demand Southbound direction, as expected.

The comparison of network vehicle counts in Figure 5.2 suggests that Cb-MP is able to service approximately the same volumes as the optimized actuated-coordinated control when cycle times were comparable. The higher cycle length Cb-MP implementations are omitted from this plot for clarity; these controllers resulted in higher variations of vehicle service between 5-minute periods but ultimately only reduced total service rates slightly.

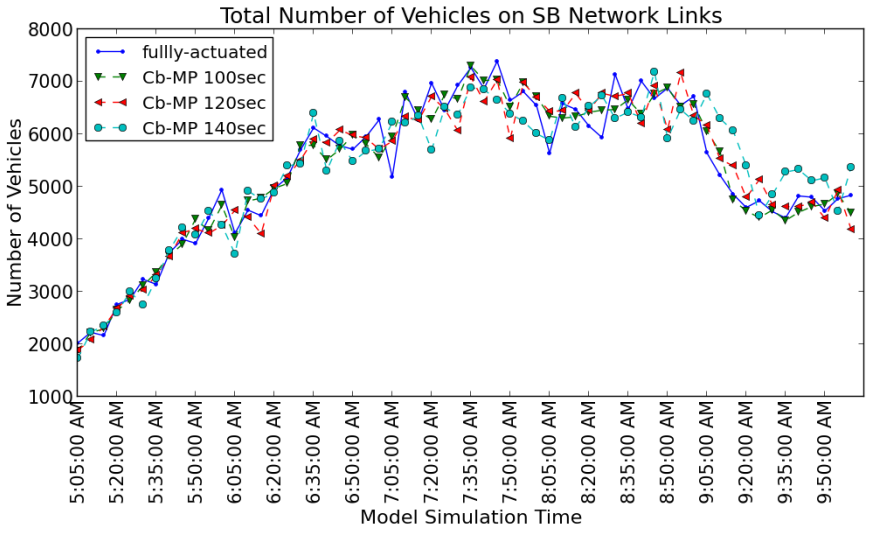

Figure 5.2: Cb-MP demonstrated service rates that are consistent with a fully-actuated control system for similar cycle lengths.

Yet distinct differences between the fully-actuated and Cb-MP controllers were observed in measurements of delays. Figure 5.3 compares the average vehicle delay given fully-actuated and max pressure control with the same cycle length. It is apparent that while the fully-actuated controller produces less delay when demand is far below network capacity, Cb-MP outperforms the existing controller given consistently high demand; it imposes less delay

with a noticeably smaller variance. This may not be surprising given known deficiencies in actuated controllers, however it is important to point out that this implementation of max pressure produces very promising network delays with almost no controller parameters that require tuning.

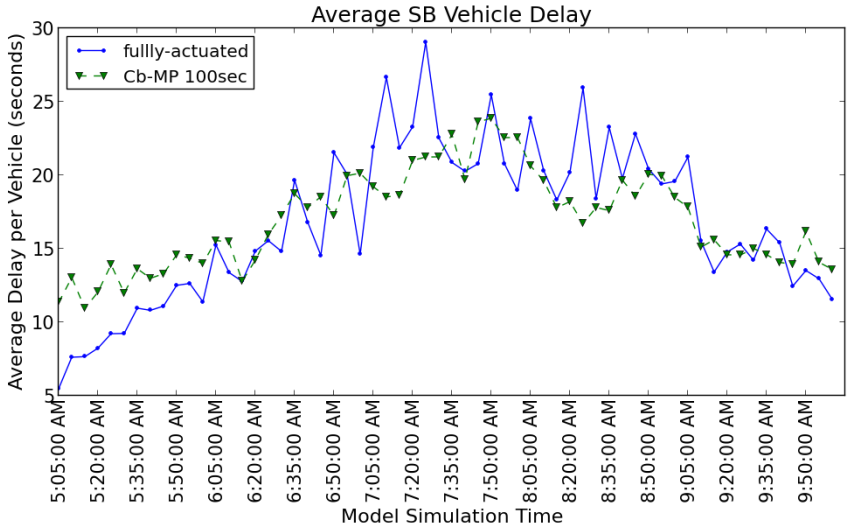

Figure 5.3: Cb-MP outperforms the actuated controller given high demand in terms of vehicle delay.

Despite maintaining the same relative offsets for actuation of the main (coordination) direction as the fully actuated controller, Cb-MP induced slightly more stops during a vehicle's south-bound journey across the network. Again, this may be expected given the stop-minimizing design objectives of the fully-actuated system: the small but consistent differences in average vehicle-stops shown in Figure 5.4 are likely caused by the on-demand service extensions provided for low density "back-of-queue" arrivals in the fully-actuated system. Notice that the average vehicle stopped time is actually lower in Cb-MP than with the fully actuated system during peak demand, which is consistent with the estimates of total delay demonstrated in Figure 5.3.

The higher cycle length Cb-MP implementations are again omitted from Figures 5.3-5.4 for clarity, yet it is important to note that higher cycle lengths predictably led the longer stops and more delay, as vehicles which encountered a red light would have to wait longer for the cycle to reach their desired green phase. This increase in wait time also corresponds to the predicted larger queues.

Figure 5.5 demonstrates the increase in mean queue lengths with increase in cycle length τ. While the linear increase in maximum queue length derived in (5.54) is not explicitly depicted in the observations, these results appear to remain consistent with such an upper bound.

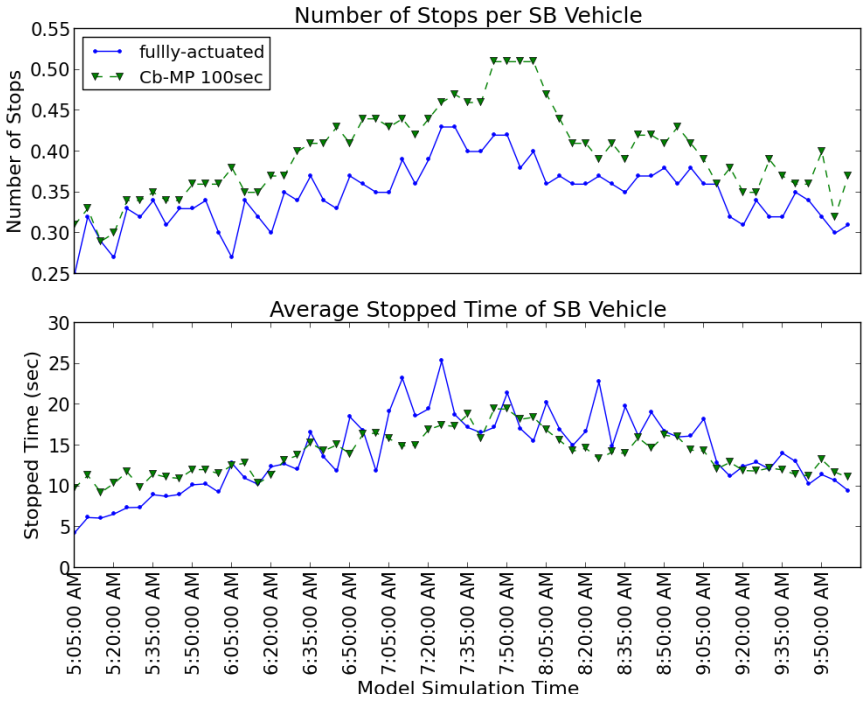

Figure 5.4: While Cb-MP caused more vehicle stop events, stoppage times were similar to those observed using the actuated controller.

The numerical implementation of cycle-based max pressure presented here suggests a promising alternative for signal control in periods of high demand where the performance of existing actuated controllers is known to deteriorate. It is an intuitive and scalable control algorithm that is appealing because it maintains theoretical network-wide performance guarantees without the need for centralized communication or control centers. The cyclical operation of Cb-MP can still maintain the flow-progression benefits obtained from existing offset optimization algorithms along a prioritized route, as demonstrated by the fact that the average number of vehicle-stops on the southbound route only increases slightly using Cb-MP over an implementation of the optimized actuated system. Because the cycle splits are more predictable in Cb-MP than in current actuated-coordinated algorithms, it may even be possible to further optimize progression on multiple (conflicting) routes with additional linear constraints on cycle splits in (5.17).

Furthermore, Cb-MP is a widely-applicable algorithm which requires significantly less tuning and site-specific adjustment than the typical fully-actuated control system. For example, the timing parameters for a fully-actuation system deployed on networks such as the San Diego site referenced above are often a result of many hours of both model-based and heuristic optimization procedures for a specific network geometry and expected demand. Yet in our implementation, the generalized Cb-MP algorithm with arbitrary reasonable minimum green parameters achieved approximately equal performance without requiring any knowl-

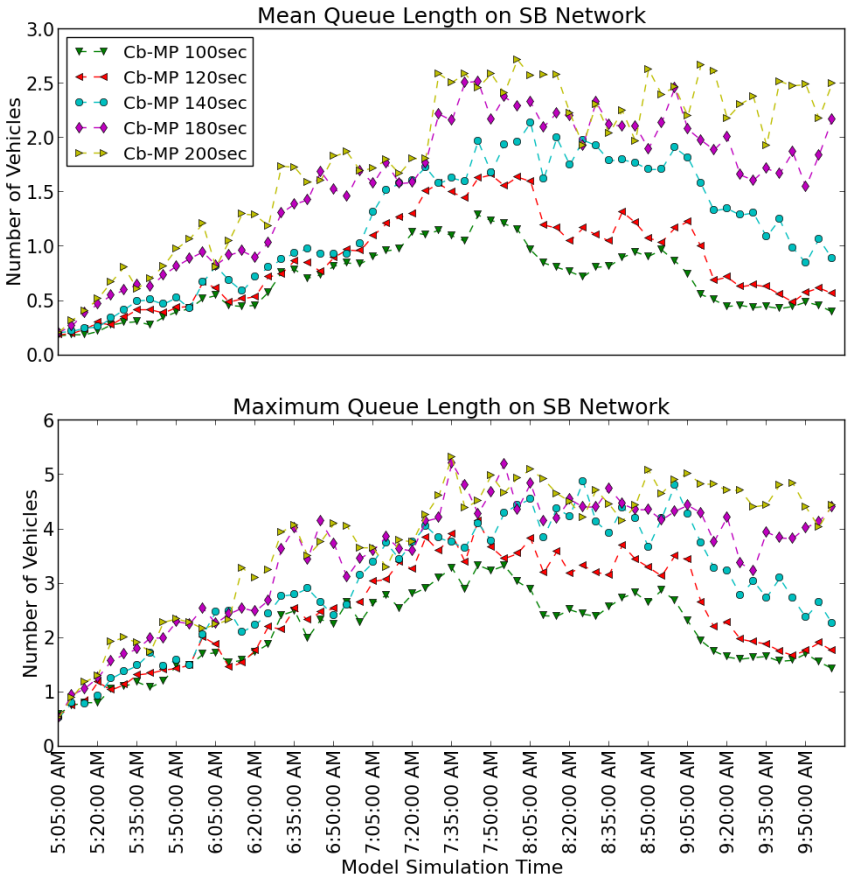

Figure 5.5: Observed queues increase with cycle length using Cb-MP control.

edge of location or demand.

By addressing the practical problem of lost capacity due to frequent switching, our extension of the proofs of [205] brings the concept of max pressure closer to a realistic implementation. Yet it is important to point out that existing sensing infrastructure does not typically provide the capabilities necessary to accurately measure approaching link-counts, nor does it provide any measurement of downstream link state. Future work should also address limitations inherent in the vertical queueing model framework such as the assumption of infinite link buffer size. In practice, oversaturation could cause unmodeled instabilities in a traffic network if expected queue bounds exceed physical road storage.

One could also consider multiple ways in which the performance of Cb-MP could benefit from heuristic modification, such as enforcing a maximum value of the green split λ_l as a function of queue measurement x_l to prevent wasted green in a given cycle. Our extended proof of stability should hold for any additional linear constraints on λ in (5.17) (which maintain the concept of a relaxed, τ-updated controller) given that there exists any controller satisfying these constraints for which the network demand is feasible according to (5.19).

Chapter 6

Facilitating implementation of traffic responsive plan selection operations

Typical traffic controllers can be operated with plan selection based either on time of day (using time-of-day or TOD mode) or on observed conditions (using a traffic responsive plan selection or TRPS mode). TRPS mode will enable a signal controller to use immediate feedback from local volume and/or occupancy sensors to choose a timing plan optimized for current conditions from a pre-programmed set of existing plans.

While most of the implementation-oriented research performed to date on TRPS has focused on either small networks of less than five intersections or on artificial (theoretical) networks, the following studies have shown that operating in a TRPS mode often has large potential for achieving delay reductions in highly-varying or abnormal traffic conditions:

- A TRPS implementation based on real-time use of the Traffic Network Study Tool (TRANSYT) software claims a 15 percent delay reduction over application of a fixed-time or vehicle-actuated control [217].
- An analysis of a simulated traffic responsive system on SR-28 in Lafayette, Indiana using 5 different plans (originally TOD plans for midday, morning, afternoon, event-inbound, event-outbound) suggested that delay could be reduced from 14-28% compared to TOD with these plans. But a lot of fine-tuning was required to prevent frequent unexpected switching which initially reduced system performance [153].
- A recent deployment on Reston Parkway in Northern Virginia demonstrated across-the-board improvements in delays, travel times, number of stops in both congested and un-congested conditions over previous TOD operations [1].

Despite its promise, TRPS is rarely used in normal signal operations; most jurisdictions simply default to a sub-optimal plan switching schedule. The difficulty and lack of intuition in calibrating the many weights and thresholds of a TRPS system is often cited as the main factor discouraging its implementation [3]. This has driven the search for ways to automate the process of designing and calibrating the required parameters.

One of the biggest obstacles in implementing TRPS is the inability to model the direct impacts of control parameters on a desired performance objective. For example, the primary method of quantifying intersection level-of-service (LOS) in the *Highway Capacity Manual 2010* is an evaluation of the control delay. Specifically, delay is calculated in the HCM by a prescribed equation (2.27)-(2.29). Yet this functional form is highly dependent on heuristic parameters, and there is much debate as to its accuracy and applicability. In fact, while intersection delay is an intuitive metric for the performance of a signal controller, there does not exist any universally-accepted model for accurately calculating delay at a signal-controlled intersection with arbitrary vehicle arrival patterns (as discussed in Section 2.2).

Many researchers have therefore suggested using new data-driven classification algorithms for designing a TRPS controller, for example hierarchical clustering/regression trees [195, 196], k-means clustering [184, 210], linear discriminant analysis (LDA) [168], genetic algorithms [165, 2], or neural networks [82]. While theoretically promising, these methods all require a great deal of flexibility in controller design, including the creation of new signal plans that are not guaranteed to adhere to the many practical political or safety constraints that must be considered by existing traffic signal operators or technicians.

Here we propose an new method for rapidly configuring TRPS system parameters for global delay reduction *using only the set of signal timing plans that is already encoded in network controllers*. This methodology is model-independent and therefore easy to implement on any network given reasonable knowledge of sensor placements and critical intersections. The "constraint" of using only existing signal plans makes our method more immediately useful than previous proposals, as it skirts the need for long and costly re-timing processes—but can still incorporate new plans as they become available. We believe that this will make our methodology very attractive to municipalities hoping to improve the efficiency of their existing automated signal control procedures without the expense of re-timing procedures or the need to acquire new hardware.

In the following sections, we describe the typical TRPS mechanism that is present in existing controller software, explain how our procedure could be implemented to calibrate this mechanism, and provide a proof-of-concept demonstration of the procedure which improves the theoretical performance of a real signal in terms of a simple estimation of intersection delay.

6.1 Background: existing traffic responsive plan selection functionalities

TRPS can be implemented on either a single controller or a set of neighboring controllers equipped with occupancy and count sensors and an appropriate system management software. Generally, a controller (either an isolated controller or *master controller* in a *coordination group*) aggregates a set of scaled and smoothed detector measurements into weighted linear combinations that are known as *Computational Channel (CC) parameters*. Functions

of these CC parameters then define three Plan Selection (PS) indices, which are used to reference an appropriate pre-encoded plan from a static three-dimensional look-up table stored in controller memore. Figure 6.1 illustrates the typical process hierarchy for calculation of PS indices.

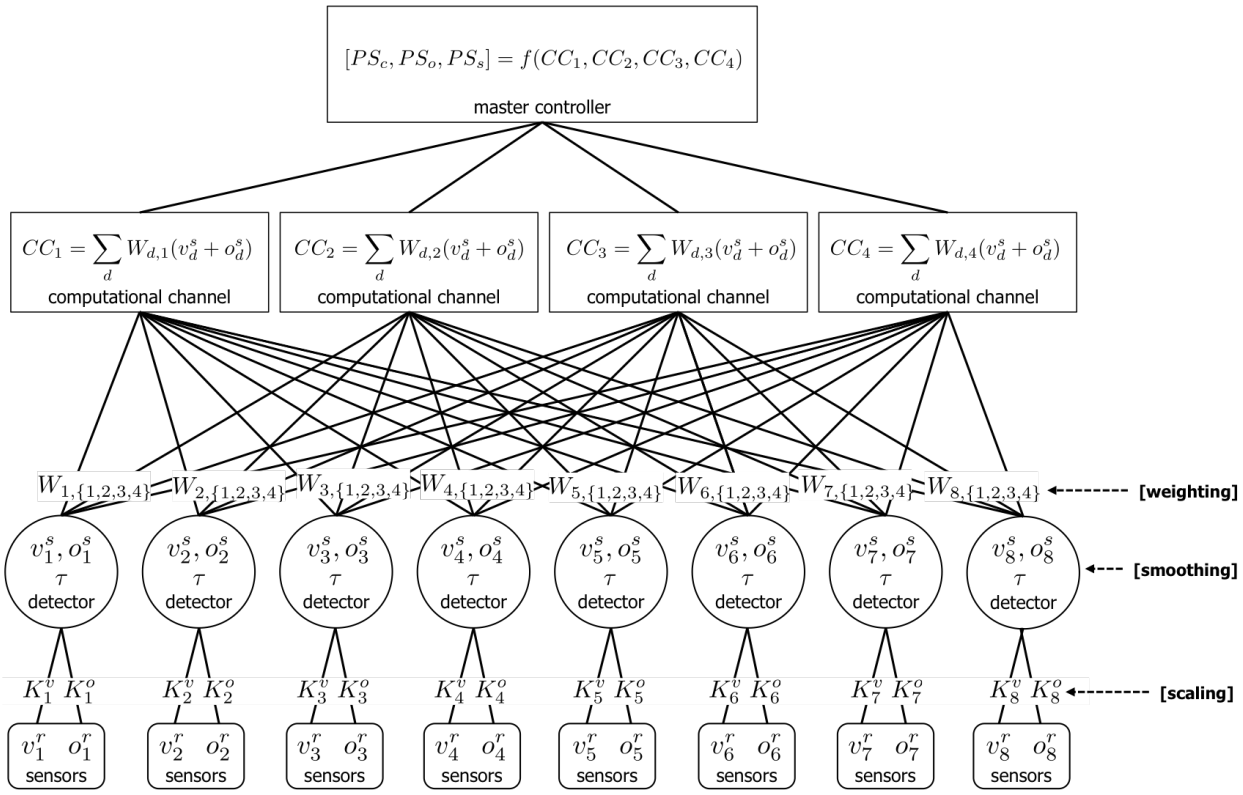

Figure 6.1: Individual sensor measurements are first scaled, smoothed, and combined into detector values. Computational channels (CC) are linear combinations of detector values, and plan selection (PS) parameters are calculated as a function of these computational channels. The PS parameters are used as indices in a look-up table to determine the proposed plan.

CC parameters

As illustrated in Figure 6.1, raw measurements can be treated with *scaling*, *smoothing*, and/or *weighting* factors in the process of calculating the CC parameters.

Scaling factors Scaling factors (or gains) convert raw count and occupancy measurements into a value which is approximately representative of the utilized capacity of the approach being detected. Assume that a scaling factor K for each individual sensor can range between 0 and 100.

Smoothing factors Some type of time-averaging operation is typically performed to smooth raw measurements, for example

$$x^{smoothed}[t+1] = x^{smoothed}[t] + \tau(x^{raw}[t+1] - x^{smoothed}[t]) \tag{6.1}$$

for some smoothing factor τ. This helps identify long-term trends from high-frequency variance in cycle-to-cycle movement demands.

Weighting factors Detectors are assigned weighting factors W by which the (scaled and smoothed) approach measurements are multiplied prior to calculation of the PS indices. Selection of detector weights W have a large influence on the resulting CC and eventual PS. Intuitively, detectors with high variance corresponding to the crucial distinctions in network state should be assigned a higher weight than those which have less variation in output.

Plan look-up table

The final step of the TRPS algorithm is the calculation of PS indices, which are some simple function of the CC parameters. The exact form of this function is dependent on the management software in use, but can be assumed to be a mapping that involves some sort of weighted average, rounding, $\max(\cdot)$ or $\min(\cdot)$ operator.

Importantly, *PS indices are not explicit suggested values for each of the plan characteristics*, rather they only represent a label to match detected conditions with suggested characteristics. As previously mentioned, differences in PS indices typically represent thresholds in optimal cycle length, green split, and offset timing features. These indices ultimately correspond to coordinates in a static look-up table such as that illustrated in Figure 6.2.

At the end of an evaluation period, each master controller uses the PS indices calculated from detectors within its coordination group to locate the appropriate plan from this predefined table and disseminates the intended plan to the other coordinated signals. Each individual controller then actuates its pre-encoded timings corresponding to the centrally-chosen plan.

Hysteresis thresholds

It is known that frequent switches in timing plans causes unintended congestion due to disruptions in planned signal coordination. Therefore TRPS systems typically have limitations on the frequency of plan switches. They also often have a built in "hysteresis" mechanism to increase system stability given rapid changes in congestion state.

Many management systems designate "entering" and "exiting" thresholds for PS states. These effectively define an "overlap" between adjacent measurement classifications where the PS indices will tend to remain in their previous state until the alternate state fully dominates (exceeds the exiting threshold of) the current measurement state.

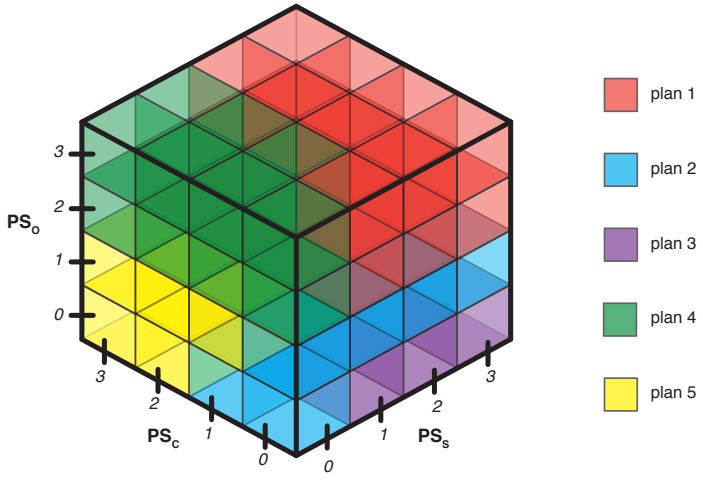

Figure 6.2: Plan selection (or PS) parameters correspond to indices of a lookup table containing plans available for the TRPS mechanism. Traditionally, the value of each PS index corresponds to the preferred value of one of three independent signal control parameters: cycle length (PS_c), green split (PS_s), and cycle offset (PS_o).

Alternative implementation of TRPS

Early implementations of TRPS depended on a formulation designed for the FHWA's Urban Traffic Control System (UTCS 1-GC) standard controller software [69]. This particular pattern matching algorithm is natively implemented on many versions of the firmware for existing 170 controllers. However today it is sometimes overridden by the TRPS algorithms of more advanced system management packages. We do not explicitly deal with this implementation in our algorithm, but it could be considered a special case of the general weighting and scaling procedure. More on this "vpko" implementation of TRPS is provided in Appendix C.

6.2 Analysis of the potential benefit of TRPS on signals in the I-210 corridor

To analyze an example of benefits which could be obtained in an ideal implementation of TRPS, we obtained four weeks worth of sensor data from a set intersections in Arcadia, California. Out of the 52 intersections that were equipped with sensors at the time (seen in Figure 6.3), we were only able to extract useable data from 30—and not a single intersection had every single one of its individual movements represented. Note that this does not imply that sensors were not present on all approaches for use in a TRPS application, only that all existing sensors were not designated to send data to the central server.

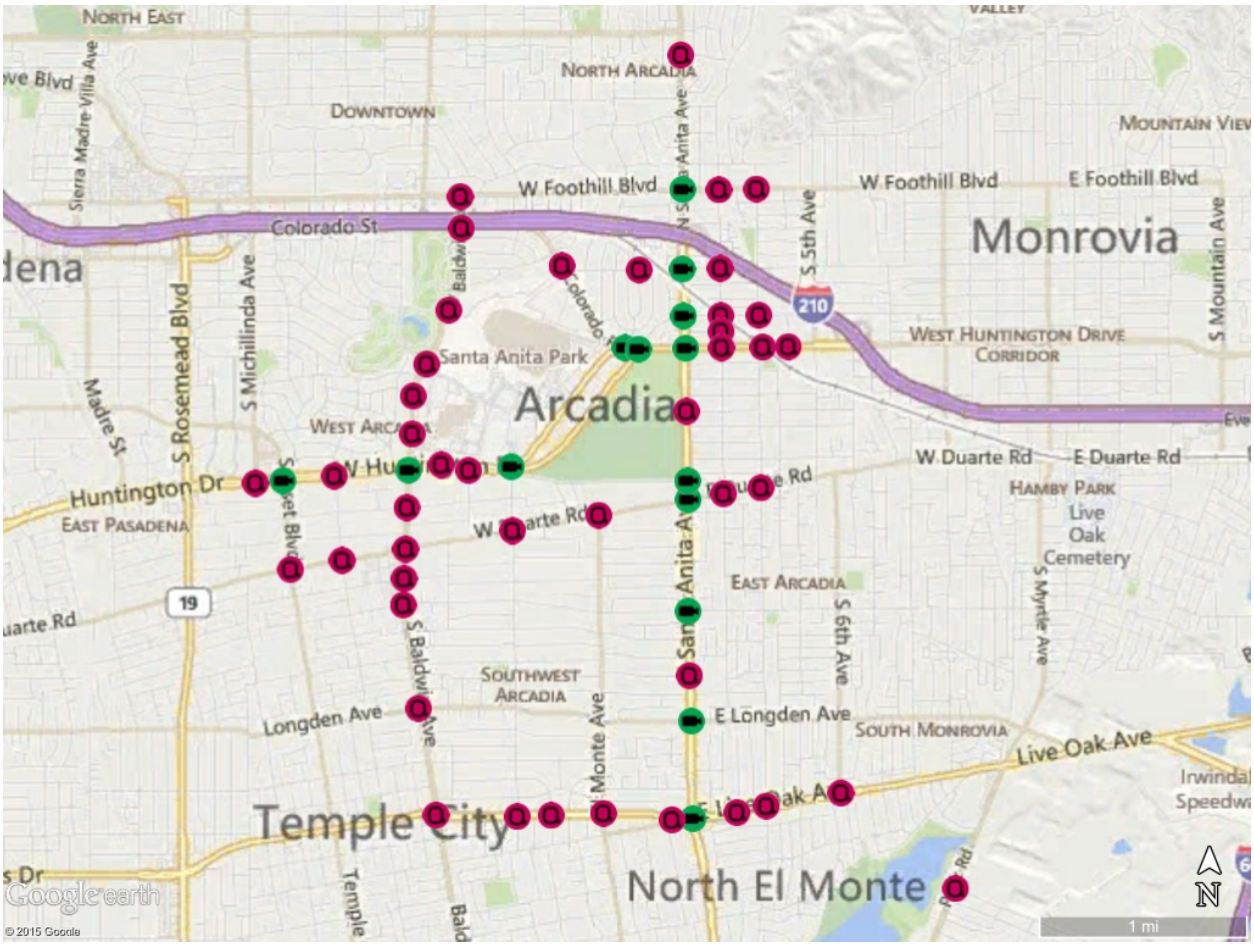

Figure 6.3: Traffic operators for the city of Arcadia, California provided approximately nine months of data collected from 52 intersections in their jurisdiction. Intersections equipped with loop detectors are indicated in red, and video detectors are indicated in green.

The functioning sensors include both video and loop detectors which measure counts and occupancies in various configurations. Data are aggregated into five-minute measurements. Most of the signals in the area operate with cycle lengths of 90-120 seconds, so five minutes represents multiple full signal cycles —thus the impact of imposed signal timings on observed demands should be minimal, assuming minimal cycle failures.

We choose delay as the primary performance metric we aim to minimize using TRPS. Because we do not have access to the intersection to experimentally measure delay, we use a the Webster delay formula (2.4) as a functional proxy to calculate an estimate of the expected control delays at an intersection.

Define $\mathcal{W}_m(q_m(t), p(t))$ to be Webster's delay formula (2.4) for a given intersection movement m, a function of measured volumes $q_m(t)$ and the intersection signal plan $p(t)$ (which is in the set of encoded plans $\mathcal{P}$). Using this model, the desired plan $p^{opt}(t)$ at every mea-

surement period t would solve the following optimization problem:

$$p^{opt}(t) = \arg\min_{p \in \mathcal{P}} \sum_{m} \mathcal{W}_m(q_m(t), p) \tag{6.2}$$

Consider the signal operating the intersection illustrated in Figure 6.4. This intersection is currently operating with three available plans, one designed for each of AM peak, PM peak, and off-peak hours.On a typical weekday, the AM peak plan (plan 2) is in operation from 0600-0900 hours (6:00-9:00 AM) and the PM peak plan (plan 3) is in operation from 1530-1900 hours (3:30-7:00 PM). Six of the eight relevant approach movements are represented by the seven accessible system sensors.

To demonstrate the sub-optimal performance of the current TOD plan switching schedule, we calculated the estimated intersection delay on the observed turn movements with each available plan, and compared the delay for scheduled plans compared to the delay-minimizing (optimal) plan $p^*(t)$. Only the movements for which measurements are available were considered; no effort is made to estimate delay on movements for which no data is provided. Results for a single weekday are demonstrated in Figure 6.5. Cumulative reduction in delay with optimal plan selection is shown in Figure 6.6. Over one week, ideal plan switching would result in a reduction of 13.82 vehicle-hours of delay as compared to scheduled operations. Scheduled operations would create an estimated 1,344.71 (cumulative) vehicle-hours of delay; hence this reduction is equivalent to only a little over 1% of estimated control delay at this intersection over the week.

If the expected onset and termination of peak hour demands remained consistent throughout a typical week, similar delay reductions could be achieved via a simple re-tuning of the TOD schedule. However we found that this was not the case: optimal switching behaviors were not in fact consistent from Monday to Friday on a typical week, as shown in Figure 6.7. Such day-to-day variations in demands corresponding to different optimal plans provides the strongest motivation for further investigation of TRPS, despite seemingly minimal delay reductions in this case. If the available plans had greater variation in green splits that are more attuned to the observed differences in AM and PM demand patterns, a greater variance in performance (and larger delay reductions) would be expected.

Yet is also important to note that no realistic TRPS algorithm could fully achieve the fully optimal plan switching patterns for one major reason: the controller must decide on future plans based on past feedback—it cannot choose ideal plans before the measured demands are served. Switches between plans will therefore be at least one measurement period behind optimal switching times.

Furthermore, the high-frequency switching behavior observed in the optimal plan patterns in Figures 6.5-6.7 is not desirable because intended signal coordinations (progression bandwidths) with neighboring intersections would be disrupted. In practice, plan switches will be minimized by the hysteresis functionalities of the TRPS system. This mechanism will however further delay the onset of desired plan switches.

Analysis of plan switching optimization over a network of intersections would involve repetition of this process for all included intersections. Accurate calculation of delay in a

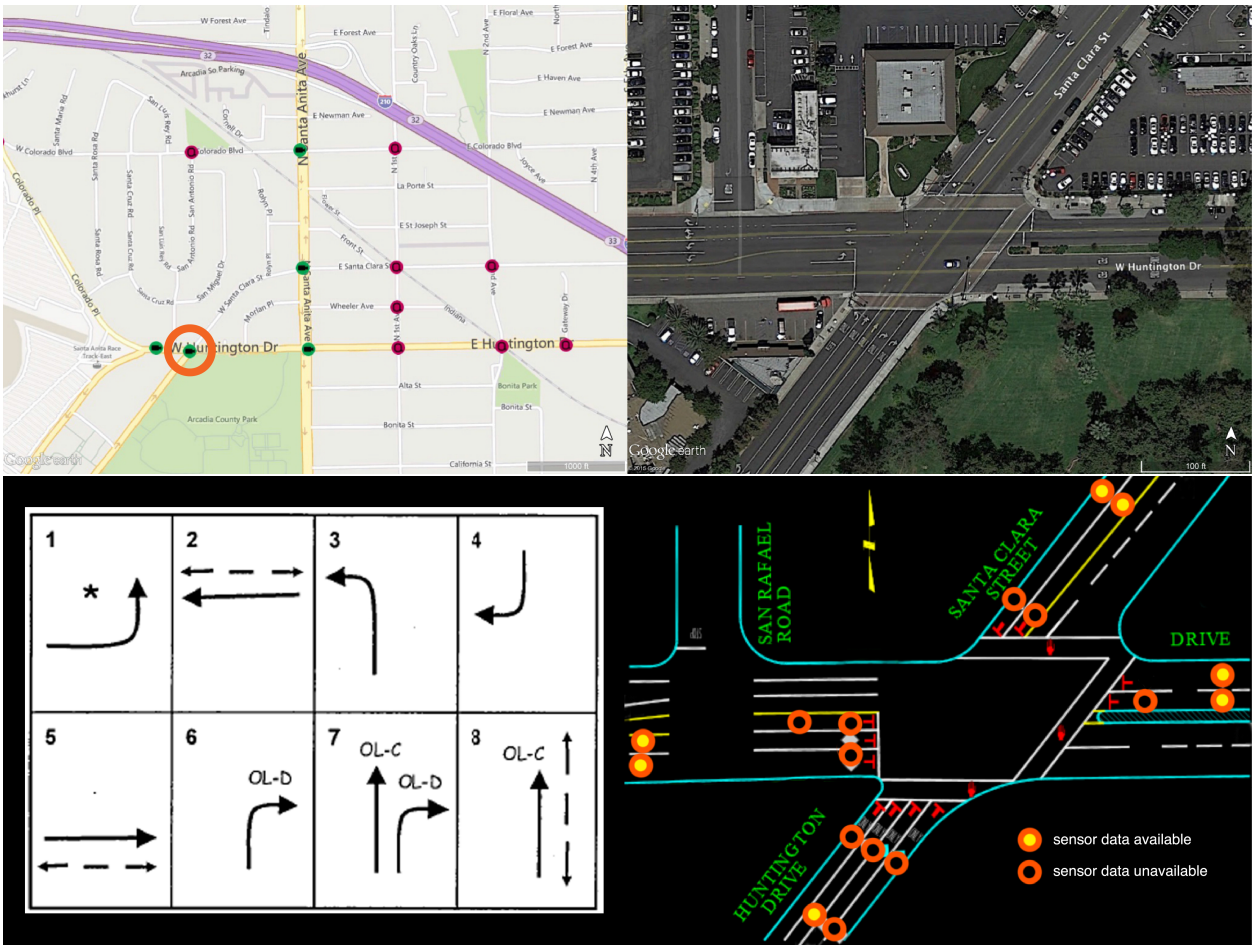

Figure 6.4: The intersection of Huntington Drive and Santa Clara Street in Arcadia, California has four approaches and three egresses. While sensors corresponding to each of the nine turning movements are installed, only data from seven approach sensors is collected centrally: measurements of north-bound right turns and north-bound left turns were not provided. Each sensor location returned 2 independent measurements: volume and time-occupancy. Thus there are a total of 14 measurements available for analysis.

network, however, should include a representation of the impacts of signal offsets, which is not the case with Webster's delay formula. While there is a term in the HCM delay formula which attempts to account for the effects of progression bandwidth, we currently lack knowledge of the intended or observed network bandwidths that is required to calculate the parameters of this delay term.

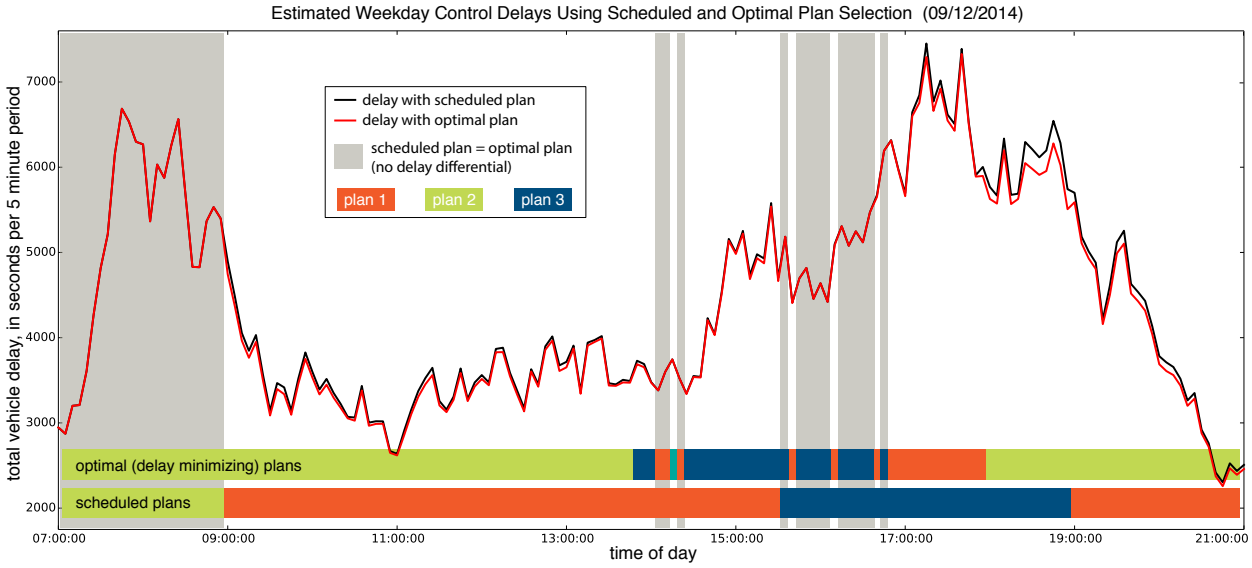

Figure 6.5: Ideal plan switching would reduce delay most notably during the hours surrounding the peak periods. In this example, it appears as if the intersection would benefit from an extension of the AM peak period plan (2) and a shift in the PM peak period plan (3) towards earlier hours. Also, the AM peak plan (2) could better serve late evening demand than either the off-peak (1) or PM peak (3) plans. However this pattern varies daily.

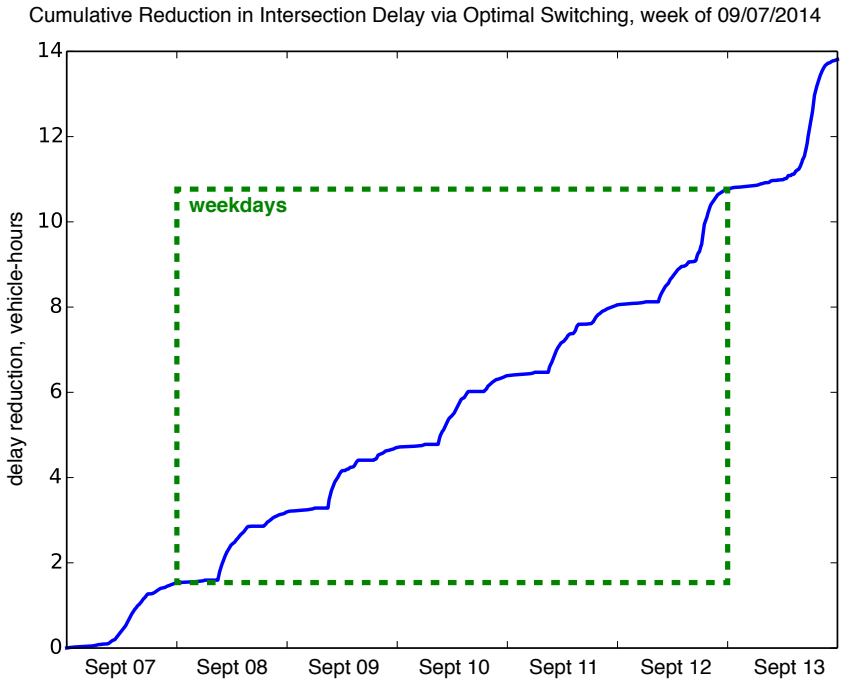

Figure 6.6: Ideal plan switching would save approximately 13.8 hours of delay (1%) as compared to continuous application of peak period plan scheduling over one week.

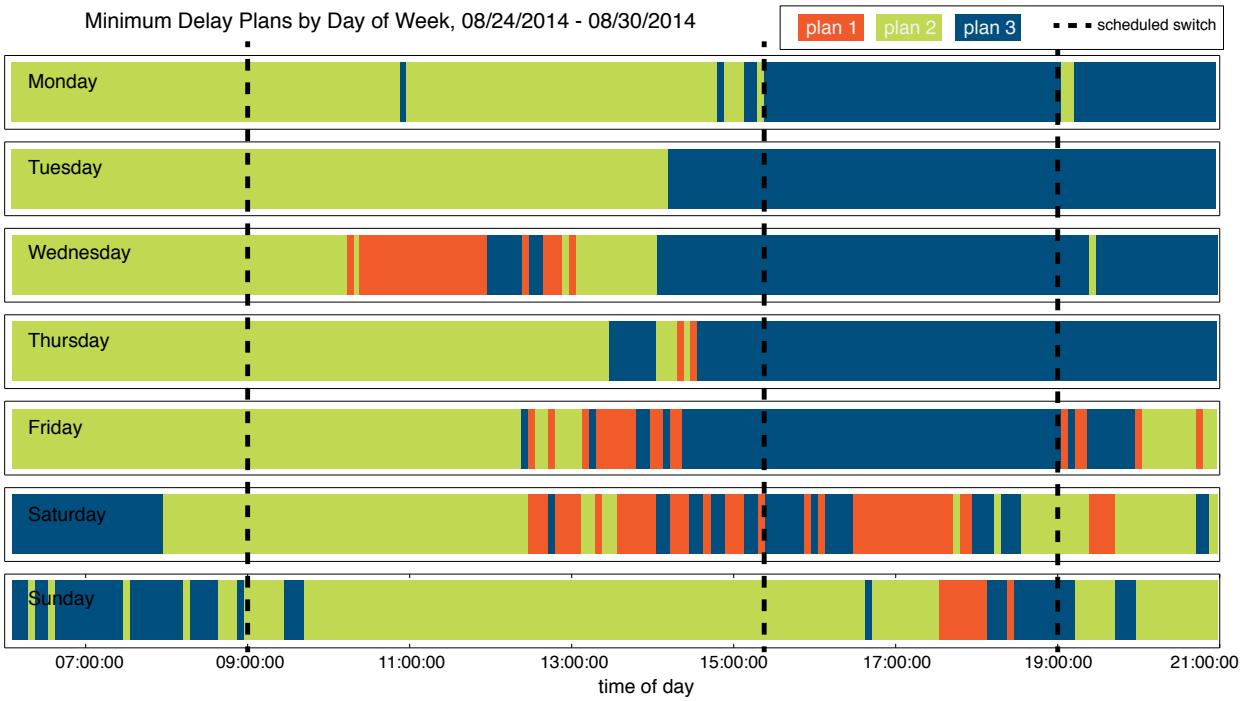

Figure 6.7: Time-of-day settings would inevitably be sub-optimal for a majority of the week days due to day-to-day variation in demands.

6.3 Formulating TRPS calibration as a supervised learning problem

We propose a mathematical methodology for calibrating the set of parameters to cause the mechanism to select the optimal plan (from the existing set of pre-designed plans) as often as possible.

This method requires a two inputs: the parameters of the existing signal timing plans encoded on a given controller, and an extensive set of measurements taken from the sensors that will ultimately be accessible to the TRPS system. It is assumed that the sensor measurements are aggregated at a rate of 5-30 minutes. If all intersection approaches are not represented by measurements, assumptions must be made regarding proportional relationships between observed and unobserved demands.

The proposed calibration procedure can theoretically be used to derive a system that selections plans to optimize any unknown (or known but non-calculable) function of available measurements. However, we choose to propose a system to specifically achieve the natural objective of minimum delay. The first step of our procedure is therefore to determine the delay that would be induced by each available control plan for each measured set of demands.

If an accurate analytical model of intersection delay was available (and all measurements and parameters required to calculate such a function were observable) an “optimal”

TRPS implementation would be trivial: the objective function for each set of sensor measurements could simply be calculated explicitly, and new sets of measurements could be functionally mapped PS indices in order to populate the optimal TRPS lookup table. However, as previously mentioned, no such analytical expression exists. This set must therefore be achieved using a numerical simulation of an intersection using, for example, microscopic simulation software (see Section 2.6 for a brief description of microsimulation tools). The desired outcome of the microsimulation procedure would be an assignment of "optimal" or delay-minimizing plan for each simultaneous set of sensor observations (i.e., a plan is assigned to every 5-minute measurement).

Modern microscopic simulation tools have built-in numerical methods for calculating vehicle delay at intersections given different signal plans. Hence the mapping of demands to delay-minimizing plan would be fairly straightforward. Note that the simulated network does not have to be excessively large or detailed, as it only needs to encompass a single intersection. However some knowledge of local coordination or platooning behaviors (depending on the specific tool used) would be required for accuracy in delay estimation. An intersection will only need to be simulated once per available signal plan with the same (deterministic) set of input demands. The computational effort required for this procedure would then of course depend on the length of the observation set.

The next step is to determine how the observable sensor measurements can best be "compressed" into computational channels (CCs) that can be used to make distinctions between optimal plans. One major difficulty faced by traffic technicians is the rigidity of the existing TRPS feedback mechanism, which operates via the system of scaling, smoothing, and weighting parameters illustrated in Figure 6.1. In practice, traffic signal technicians are left to heuristically decide which sensor measurements are "most important" to consider in the calculation of optimal cycle length, green splits, and cycle offset—and furthermore, how to design the set of CC parameters that makes the desired threshold between appropriate controller parameters as distinct as possible.

Abstractly, this procedure is equivalent to the process of compressing data into the most relevant or critical components—a process which is commonly required in modern data science applications, and is therefore very well studied. Here we consider the application of existing two data reduction techniques, Principle Component Analysis and Linear Discriminant Analysis, to compress available sensor measurements into linear combinations that are most likely to provide the information required to distinguish between assigned optimal plans. The coefficients of the resulting linear combinations can be considered equivalent to the weighting parameters $W_{d,c}$ which map the measurements of detector d to computational channel c; they can be scaled as necessary to lie within the range which can by implemented by TRPS software system.

Finally, we use machine learning techniques to develop thresholds on the space of the computational channels that distinguish between the appropriate plan assignments. At this point we have a set of "training data" (computational channel values) with known "labels" (optimal plans) and we desire to develop a set of rules to assign new data to the appropriate labels, as would be encoded in a look-up table. This is the task of a *supervised classification*

algorithm.

The relationship between these steps of the proposed calibration procedure are illustrated in Figure 6.8.

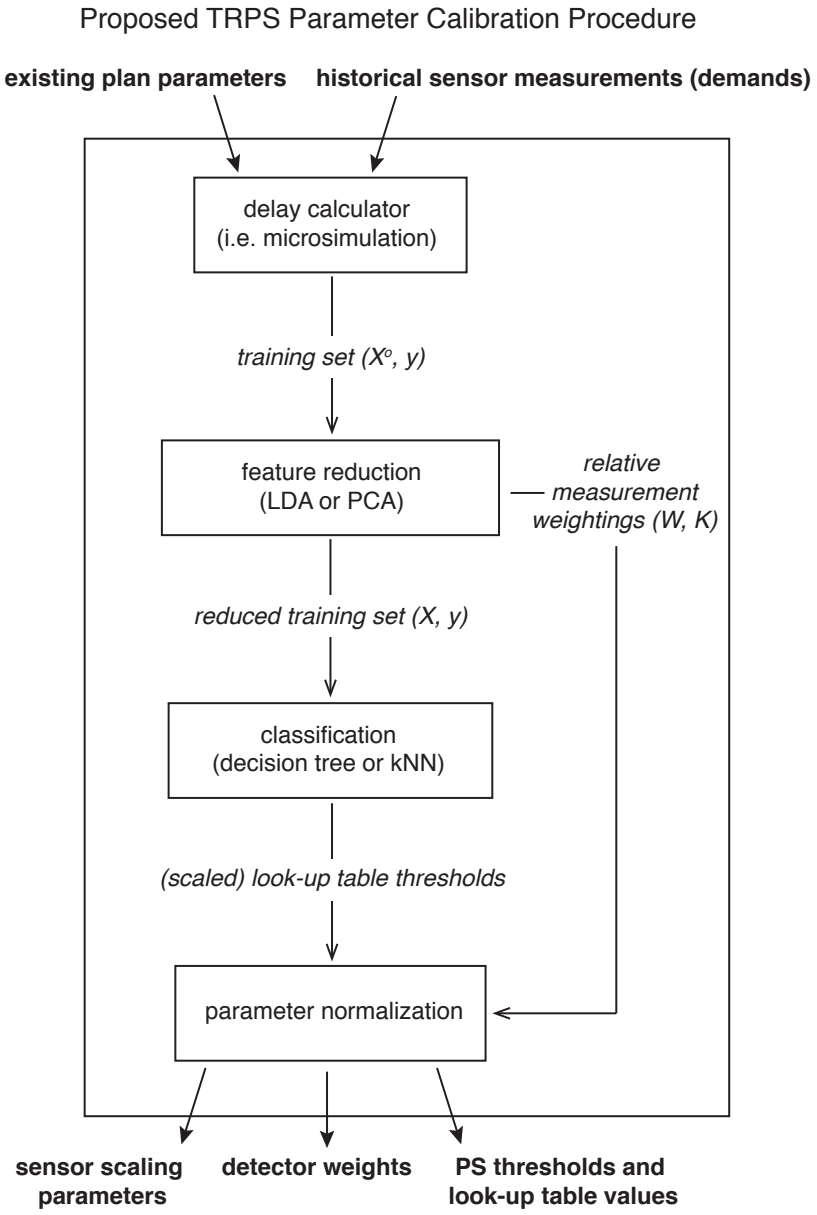

Figure 6.8: A traffic operator will existing plan parameters and historical data to receive an appropriate set of scaling, weighting, and PS threshold parameters that reduce estimated delay in a TRPS implementation.

Derivation of computational channels via feature extraction algorithms

The computational channels illustrated in Figure 6.1 primarily serve to compress the many independent sensor measurements into a smaller number of features that are considered most relevant to plan selection. There are no formal guidelines offered for constructing these features in practice. They may logically be used to highlight expected changes in demands for certain movements based on the design of the relevant plans. For example, sensors on the major arterial through movements may be given larger relative weights than those on minor through movements to aid distinction between peak period plans that are traditionally used to accommodate unidirectional spikes in demands on a major arterial. Previous academic implies that computational channels can also serve a "first pass" in distinguishing the available measurements which relate to the control variables (cycle length, splits, and offsets) to which plan selection parameters are assigned. Yet it is not intuitive how this can be accomplished, even with advanced knowledge of the geometric properties and flow patterns on a specific network.

We propose to replace the heuristic procedure with an automatic *feature extraction algorithm* to design appropriate computational channels. Explicitly, we use a feature reduction algorithm $f_r(\cdot)$ to map training measurements $\{\hat{q}^{train}, \hat{o}^{train}\}$ to an "optimal" space for classification, and then use the same mapping on all future measurements in the calibrated TRPS controller:

$$\mathbf{W} = f_r(\{\hat{q}^{train}, \hat{o}^{train}\}) \tag{6.3}$$

$$\mathcal{X}(t) = \mathbf{W} \cdot [\hat{\mathbf{q}}(t), \hat{\mathbf{o}}(t)] \tag{6.4}$$

Dimensionality reduction through feature extraction is common practice for pre-processing data in modern data science applications; it is a standard way to improve the efficiency of a learning procedure (i.e. regression or classification) by "filtering out" data fields that are of little impact to the final result before costly computations are attempted. Our proposed methodology brings benefit to the TRPS calibration procedure because we effectively remove the influence of "traditional" heuristic knowledge of relative sensor importance and instead generate an unbiased quantification of which features truly show high variation.

We evaluate the performance of the following two feature extraction methods for use in this system.

Principle Component Analysis

The simplest method of dimensionality reduction is known as *principle component analysis* (PCA) [167, 98]. PCA is an intuitive eigenspace-based analysis method that is widely used in many fields [190, 149, 104].

Consider a data set of N samples $\{\hat{q}^{train}, \hat{o}^{train}\}$ which each contain $m << N$ unique sensor measurements. Arrange this data as a m-by-N matrix $\mathbf{X_o}$ in which each column corresponds to a specific sample time and each row corresponds to measurements from a

single sensor (measuring either a volume q or occupancy o). Denote the i^{th} column of this matrix x_i. The *scatter matrix* of $\mathbf{X_o}$ (an un-normalized covariance matrix, assuming $\mathrm{E}(x_i)$ is equal to the column-mean vector μ) can be calculated as follows:

$$S = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \tag{6.5}$$

PCA seeks the direction of maximum variance, or the projection ϕ which achieves

$$\phi_{PCA} = \arg\max \left| \phi^T S \phi \right| \tag{6.6}$$

The *principle components* are simply the set of $n < m$ normalized eigenvectors of S that correspond to the n highest eigenvalues. Intuitively, these eigenvectors indicate the orthogonal dimensions which capture the most variance in the columns of $\mathbf{X}_o$ (the training data).

Explicitly, ϕ is a m-by-n matrix with the principle component vectors (the first m eigenvectors of S) as columns. The dimension of a data set X_o is reduced via the linear transformation

$$\mathcal{X} = \phi^T X_o \tag{6.7}$$

In this application, we choose n to be equal to the number of desired computation channels in the TRPS mechanism. The set of sensor weights W_i corresponding to CC_i would then be equal to the (scaled) elements of the i^{th} eigenvector.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), also known as *Fisher's Linear Discriminant*, is often the default choice for feature reduction in *supervised* learning applications such as this one [21, 133]. While an *unsupervised* feature extraction algorithm such as PCA generates a set of features which account for the most overall variance in a dataset, LDA targets those features which have the largest influence specifically on the corresponding classifications. In other words, the relative sensor weightings derived by this algorithm indicate the relative impact of specific sensors measurements directly on the space of the outcomes instead of on the space of the measurements themselves. A projection derived via LDA would therefore ideally result in maximal separation of the labeled classes.

Organize the training data into matrix $\mathbf{X}_o$ as previously described. Define a separate mean vector μ_k of the measurements corresponding to each available class $k \in \{1, \ldots, K\}$. The sum of the un-normalized co-variance matrices calculated for each class-subset of the training data is called the *within-class scatter matrix*:

$$S_w = \sum_{k=1}^{K} \sum_{i=1}^{N} (x_i - \mu_i)(x_i - \mu_i)^T \tag{6.8}$$

The *between-class scatter matrix* is then defined as

$$S_b = \sum_{k=1}^{K} N_k(\mu_i - \mu)(\mu_i - \mu)^T \tag{6.9}$$

where N_k is the number of samples (columns) classified as class k and μ is the mean of *all* training measurements.

LDA selects a projection that maximizes the ratio of the determinants of the between-class scatter matrix and the within-class scatter matrix, known as *Fisher's criterion*:

$$\phi_{LDA} = \arg\max \frac{\left|\phi^T S_b \phi\right|}{\left|\phi^T S_w \phi\right|} \tag{6.10}$$

It can be shown that the columns of the projection matrix which maximizes (6.10) are equal to the subset of the eigenvectors of the matrix $[S_w^{-1} S_b]$ which correspond to the largest eigenvectors.

Note that this matrix is of dimension m-by-K, and therefore its rank is necessarily $\leq K$. Hence LDA can at most derive $(K-1)$ feature vectors. In our application, this is a strong limitation: as will be seen in our simulated implementation, the features derived via LDA will not be able to generate a three-dimensional plan selection table for a signal which has fewer than four plans available for selection.

Mapping CCs to PS indices

Very little information is available about how to translate the values of the computational channels into plan selection parameters. It appears to vary according to the manufacturer of a specific controller or signal management software.

Define this mapping between CC parameters and PS indices as $PS = \psi(CC)$. Consider only the following constraints on the PS indices:

- The table is indexed by three PS indices, which may or may not be linked to the three degrees of freedom in a signal plan (cycle length, splits, offsets). This assumption provides flexibility for improved distinction between plans in cases where, for example, a cycle-length PS parameter would become irrelevant because all available plans share a common cycle time.
- The function $\psi(\cdot)$ is user-defined, but can only use relatively simple operations such as summations, averages, integer-rounding, $\min(\cdot)$, or $\max(\cdot)$.
- PS indices must be positive integers between 0 and a fixed maximum index M.

The first assumption highlights one of the initial challenges we encountered while experimenting with TRPS calibration procedures. Literature implies that the PS indices are typically assumed to correspond directly to the three degrees-of-freedom in a signal plan

design process. While this would be a reasonable approach if one were to have full freedom in plan design when calibrating a TRPS (in other words, the plans were being designed with this TRPS mechanism in mind), it is a severe limitation in our case because we are attempting to demonstrate TRPS only with existing plans.

In practice, differences in green splits are almost universally the primary distinction between plans. In the case of coordinated controllers, offsets are also a key source of plan variation. But in our analysis of the plans implemented on Arcadia intersections, we commonly ran across cases in which all available plans for a given intersection shared a common cycle length. In this case, the plan selection parameter designated for the cycle length characteristics would become irrelevant, and our plan selection table would be effectively reduced to two dimensions. Furthermore, because we do not have full observation of all relevant geometric parameters of the network, we lack knowledge of the intentions of impacts of the offset parameter on signal performance, so it was also difficult to make plan distinctions based on this PS axis. We were therefore only left with a single index (the PS corresponding to splits) upon which to base our plan choice.

Here we defy the precedent that each of the PS indices must correspond to a plan characteristic. We found no reference to specifically evidence that this was a necessary assumption in the functionality of existing TRPS mechanisms. Furthermore, the process for deriving CCs we described above is actually designed in a manner that is blind to this type of physical interpretation—but instead provides a theoretical robustness that cannot necessarily be achieved with the artificial division between individual plan characteristics.

We choose a very simple mapping for computational channels to PS indices: the i^{th} PS equal to the i^{th} CC after normalization, rounding, and scaling to a valid positive integer between 0 and M (to enforce maximum possible precision).

$$PS_i = \psi(CC) = \min\left\{\max\left\{\mathbf{int}\left[\frac{M}{b_i^{max} - b_i^{min}}(CC_i + b_i^{min})\right],\ 0\right\},\ M\right\} \tag{6.11}$$

where b_i^{max} and b_i^{min} are the maximum and minimum values of the i^{th} CC observed in the training set, respectively, and the $\mathbf{int}[\cdot]$ implies rounding to the nearest integer.

Notice that this method does limit our procedure to using (at most) three CCs, given that there are only three PS dimensions available. We discuss the implications of this on our test implementation in the following section.

Plan selection table design: comparison of supervised learning alternatives

The last step of the calibration procedure is the population of the lookup table. As previously mentioned, this is accomplished via a supervised classification procedure to map all possible combinations of PS indices to the best available plan.

We define *classification* as the process of assigning a discrete label to an object or state based on a set of known characteristics. In this case, we are attempting to assign an appropriate signal plan $p \in \mathcal{P}$ for the signal controller based on feedback from sensor measurements

$\{\hat{q}, \hat{o}\}$ —or more specifically, the set of features derived from the measurements using the previously discussed data reduction algorithms, $\mathcal{X}(t)$:

$$p = \mathcal{C}\big(\mathcal{X}(t)\big) \tag{6.12}$$

In a TRPS implementation, we actually use this classification algorithm slightly differently: because we cannot explicitly compute the classifier on the controller hardware, we use the classifier to populate a discrete look-up table $\mathbf{T}$ which stores the values of the classifier *at each possible combination of PS indices* $\mathbf{PS} = \psi(CC)$.

$$p^c = \mathbf{T}[PS] \tag{6.13}$$

$$\mathbf{T} = \mathcal{C}(\mathcal{X}^{train}) \tag{6.14}$$

for a set of training features $\mathcal{X}^{train}$ derived from a set of historical sensor measurements. We are effectively creating a discretized representation of the classifier output for reference in the TRPS controller.

Three multi-class classification algorithms were considered for this application.

Decision Tree

A *decision tree* iteratively analyzes groups of training data to determine a variable which "optimally" splits the group into two subsets. It begins by selecting the entire training data set and splitting it into two sub-nodes called "branches". It then acts on each branch individually until it reaches "leaves" which attain full classification or some other pre-defined stopping criterion. Decision trees are one of the most popular tools for data mining applications, and there are many in depth references on their implementation [35, 84].

At every node of a decision tree, the division between subsequent branch node groups are based on some threshold on feature values that produces the most information gain (or reduction in impurity) between the parent node and the resulting sub-groups. There can be many measures of optimality in the splitting criterion, but the goal is generally to progressively reduce the variance of the data in the separated groups as they progress through subsequent branches. In this work we choose a split with minimum *Gini impurity*.

Consider selecting a single measurement from a data set, and randomly labeling this measurement by selecting a value from the overall distribution of labels present in the whole set. Gini impurity (I_G) is a measure of how often this element would be labeled *incorrectly*. For some group of data with K classes, define f_k to be the fraction of the data that belongs to class k. The Gini impurity of this set is defined as

$$I_G(f) = \sum_{k=1}^{K} f_k(1 - f_k) = \sum_{k=1}^{K} (f_k - {f_k}^2) = 1 - \sum_{k=1}^{K} {f_k}^2 \tag{6.15}$$

Notice that (6.15) reaches its minimum (0) in a group where all data belongs to the same class.

As previously mentioned, node groups are split based on a threshold in one of the continuous feature values. A node n chooses a single *split point* c_n for some feature f, such that all elements in the resulting sub nodes either have $f < c_n$ or $f > c_n$. The choice of attribute (or small group of attributes) to split on can be determined differently in different implementations of a decision tree algorithm. Often when the choice of optimal splits is not unique, one of the alternatives is chosen at random. This step of the procedure may introduce randomness into the algorithm output: it may be the case that two runs of the same decision tree implementation will produce different trees for the same input data. A variety of parameters (such as the maximum number of bins to use when discretizing continuous input features, or the minimum information gain required to induce a split) can be tuned to make the path of a decision tree more predictable, but no there are no guarantees of optimality when fixing (or not not fixing) these parameters.

If the construction of a decision tree classifier is not impeded by a pre-defined maximum tree depth, minimum leaf size, or other stopping criterion, it will theoretically provide classifier that correctly identifies every single data point in a training set (assuming that there is a true one-to-one mapping of features to classes in the data). This is not necessarily a desirable characteristic because it could cause *overfitting*, or the state in which the classifier is so specific to the training data that it cannot be generalized to perform optimally on new stochastic feature sets.

Random Forest

A *random forest* is an implementation of the decision tree concept that reduces risks of overfitting by generating sub-samples of the training data and creating decision trees on each sample [34]. New predictions are generated by labeling the input feature set using each tree individually and ultimately selecting the "most popular" mode class label. By the Strong Law of Large Numbers, a random forest is guaranteed to converge to a deterministic classifier as the number of individual trees used grows. Not surprisingly, this ensemble method has been shown to produce classification results that are much more robust and accurate than individual trees.

k-Nearest-Neighbors

A simpler method for classification is the *k-nearest-neighbor* (k-NN) algorithm. As the name implies, a prediction for any new element is simply the "majority vote" of the k "nearest" training examples in the feature space. The appropriate number of neighbors to choose will vary according to the specific characteristics of the training data set; it can be determined by testing the algorithm on a validation data set (with known classes) with many values of k to identify which k minimizes root-mean-square error in the resulting predictions.

This algorithm is intuitive and easy to implement, but has significant drawbacks. The resulting classification can easily be biased towards one class that is more frequent than others in the training data, because logically these samples are more likely to be the closest

"neighbors" to any point in the feature space. To reduce the impact of this effect, the "votes" of the k neighbors can be given weights according to their Euclidean distance from the queried location in the feature space. Yet data with a large amount of "unimportant" features (that are not relevant to the resulting class) can also degrade the results of the k-NN algorithm, and the aforementioned weighting scheme may amplify this problem. For additional reference on this algorithm, see [84].

Robustness to missing data or sensor malfunction

The inherent TRPS system can unfortunately be very sensitive to unexpected sensor failures. Once such a malfunction is positively identified, however, the proposed method of calibration makes adjusting for such failures relatively simple. Assuming that the complete mapping of sensor measurements to optimal plan (i.e., from microsimulation) is still available, only the data reduction and table generation steps would need to be reconstructed to complete a full recalibration of the TRPS controller. In practice, these two steps are highly efficient and can be fully automated—no additional manual analysis would be required. To account for the missing data, one would simply remove the missing sensors from the training set which is input into the feature extraction step without changing the corresponding plan assignment, and in practice assign the faulty sensors a weight of 0 for all resulting CCs.

In other words, the essential functionality of the calibration methodology proposed here would not be impacted by the reduction of a single sensor. If the removed sensors were highly weighted in any CC of the previous calibration, performance of the newly calibrated system could theoretically be significantly reduced—however this would be the case with any reasonable TRPS system.

6.4 Performance of proposed parameter selection system

We tested a proof-of-concept implementation of our proposed calibration method by simulating its use on the controller at Huntington and Santa Clara that was described above. We had a total of 37 weeks of sensor measurements available; thus we used the first 25 weeks of data to train our feature extraction and classification algorithms and validated the performance of the resulting TRPS mechanism using the subsequent 12 weeks of data. Importantly, instead of using microsimulation for generation of a training set (as proposed above), we used the simple analytical expression of Webster's delay (as in (6.2)) to generate our training set.

Based on a preliminary investigation of classification error, we selected the following parameters for the three classification algorithms described above:

- The decision tree and random forest was derived with a maximum depth (number of consecutive branches in the tree structure) of 10. While limiting the tree depth could

result in unnecessarily mis-classified spaces in the feature space, we found that this stopping criteria actually improved results on the validation set. Because we simply aimed for a proof-of-concept implementation, we did not investigate further restrictions on the stopping criteria.

- We chose $k = 3$ neighbors in the k-nearest-neighbor implementation based on high-level observations of the level of detail in the resulting class boundaries for ks ranging from two to five. Again, our proof-of-concept implementation did not require "optimization" of this classifier, although it could easily be performed with a more rigorous analysis of the results on the validation data set using different alternative for k.

Recall that LDA is usually considered the preferred methodology for feature extraction in supervised learning applications, however it is limited by the fact that it cannot derive more components than available classes. Hence we were not able to generate more than two sets of CC weights using LDA in this case (as there were only three plans encoded into this specific controller). However, Figure 6.9 demonstrates how classification on the validation data set using the two features extracted from the training set via LDA outperformed a low-dimensional feature set derived using PCA in all of the tested algorithms. This matches common expectations in such a supervised learning application. The fact that classification accuracy continues to increase with the number of PCA parameters in all algorithms suggests that overfitting is not a factor here, rather the lower dimensional feature reductions are not actually sufficient to capture all relevant variations in data. Yet as previously mentioned, the three-dimensional lookup table available in TRPS systems constrain our feature dimension to at most three—and at this low dimensionality, we have clear evidence that feature extraction via LDA will lead to more accurate classification than via PCA.

Figure 6.9 also suggests that the most accurate classification (with any set of three or fewer parameters) is attained using a Random Forest algorithm. Yet accuracy in classification alone cannot ultimately predict performance: the impact of sub-optimal plan selection due to the resulting misclassifications from each algorithm is not guaranteed to be the same with all algorithms. Thus we simulated the performance of a TRPS controller using each of the three previously described classification algorithms using features derived from both two-feature LDA and three-feature PCA.

Examples of the 2-D and 3-D lookup tables are shown in Figures 6.10-6.13. The controller classification accuracy (and resulting delay reduction) is highly influenced by the resolution of the plan selection table (M in equation (6.11)). With 10 or fewer possible values for each PS index, the rounding used to map each continuous CC value to a discrete PS index caused relatively large areas of the CC-space to be mis-classified. The choice of 32 index values was made via comparison of the resulting selected plans to the optimal plans (see Figure 6.14); improvement in classification accuracy was questionable in all algorithms at table resolutions higher than this.

Our TRPS controller implementation was simplified for the purposes of demonstration and analysis: upon receipt of a measurement, the controller calculates CC and resulting PS

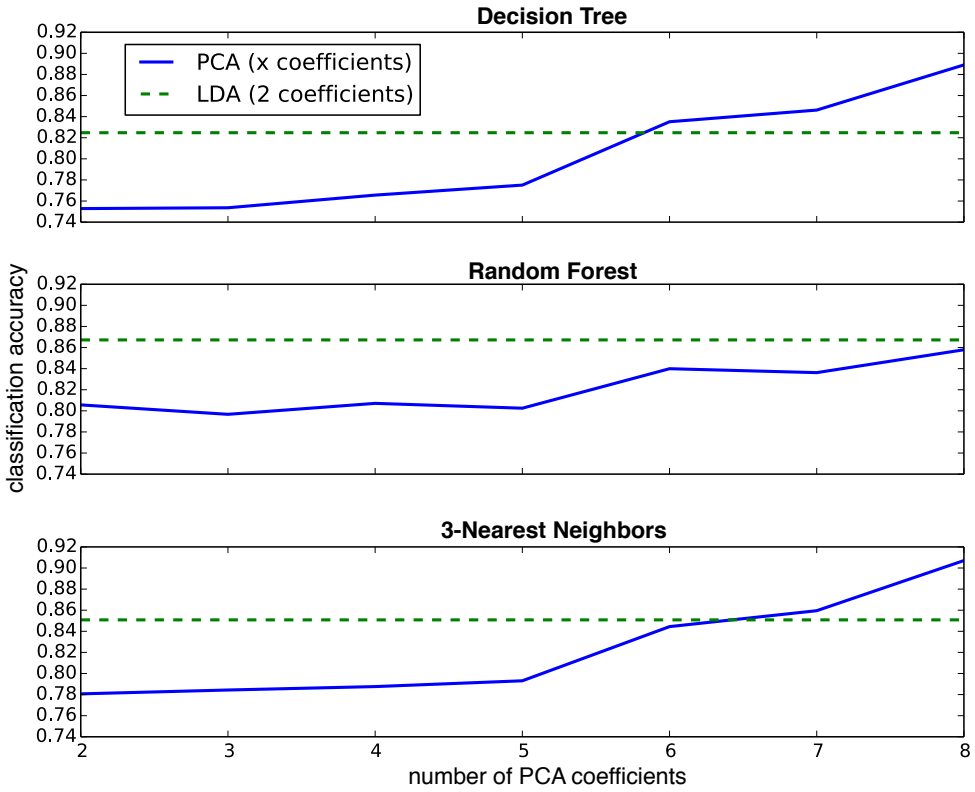

Figure 6.9: In general, the LDA feature reduction method resulted in higher accuracy of the subsequent classification algorithms as compared to lower-dimensional PCA; six or more features were required from the unsupervised feature extraction procedure to produce results equivalent to a two-feature LDA reduction. Additionally, a comparison between the three tested classification algorithms reveals that random forest was the most successful.

parameters and immediately implements the plan indicated by the lookup table at these indices at the subsequent time step, illustrated in Figure 6.15. Explicitly, the plan $p^c(t)$ selected for application at time step t is a function of volume and occupancy measurements from previous time step t^-, $\{\hat{q}(t^-), \hat{o}(t^-)\}$. These measurements are mapped through feature extraction method f_r and a table generated using classifier $\mathcal{C}$ trained on a set of historical

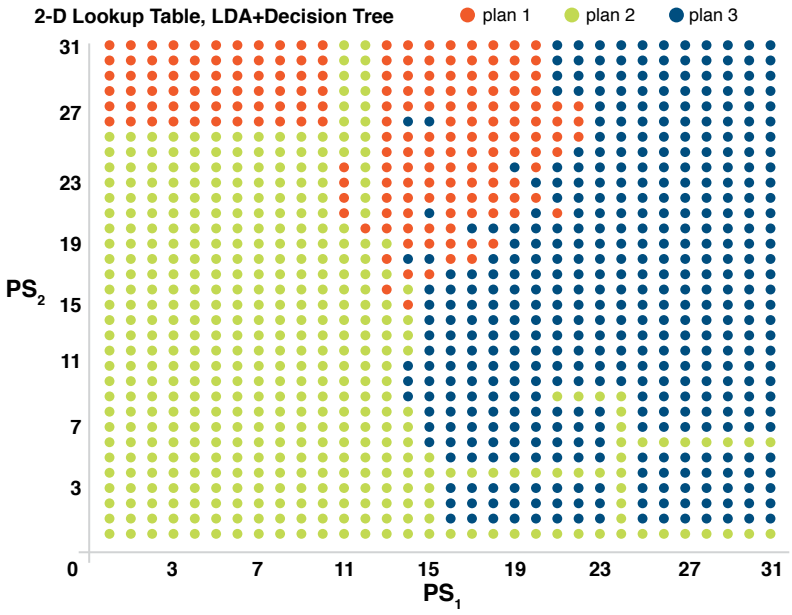

Figure 6.10: The creeping class boundaries in the lookup table generated by a single decision tree shows evidence of over-fitting.

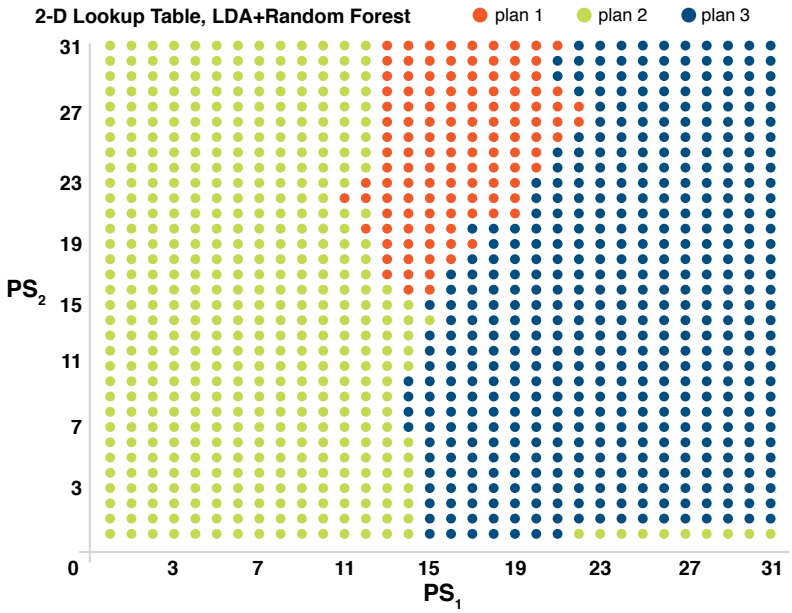

Figure 6.11: The lookup table generated by the random forest algorithm shows more regular class boundaries than the decision tree table, as expected using this ensemble method.

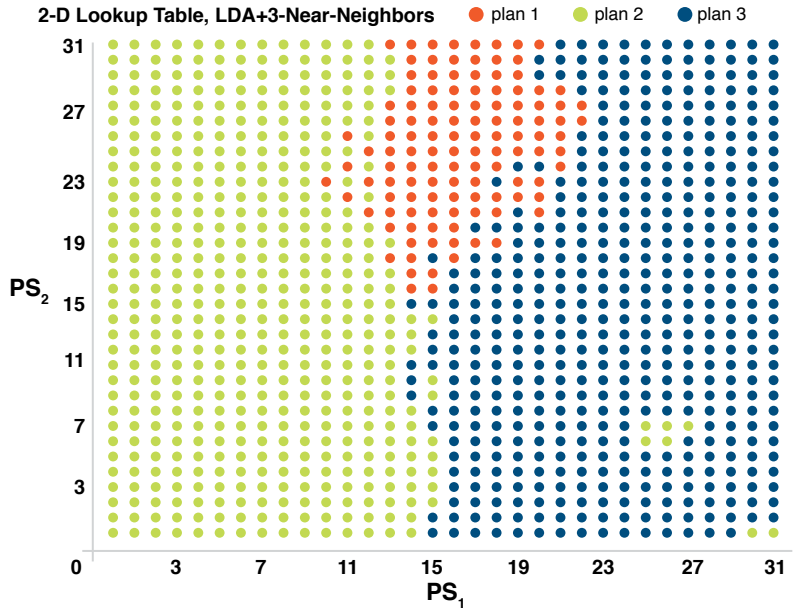

Figure 6.12: The discretization inherent in creating a lookup table filters noise in the feature space that would otherwise appear in the feature space using a k-nearest neighbors classifier.

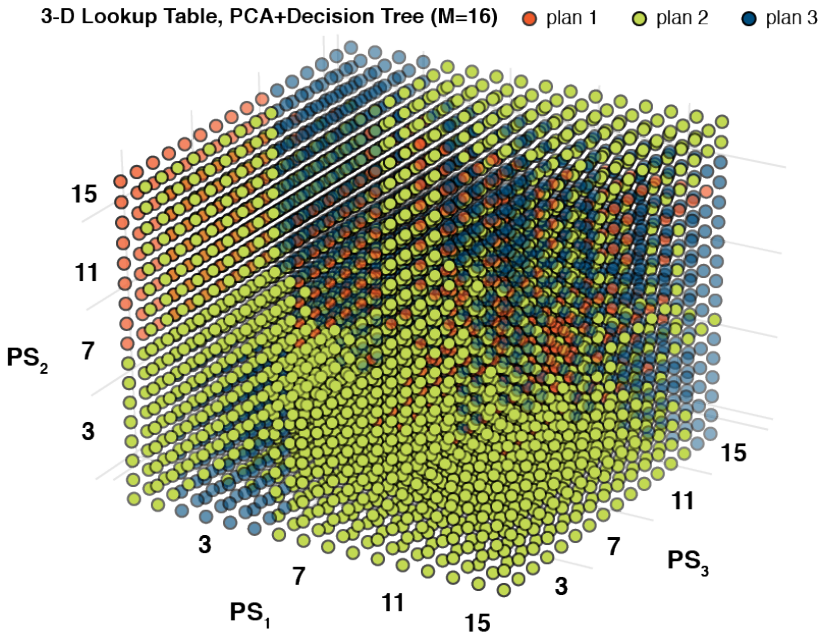

Figure 6.13: The 3-dimensional lookup table shown here is only an under-sampled representation (rendering the full version is difficult).

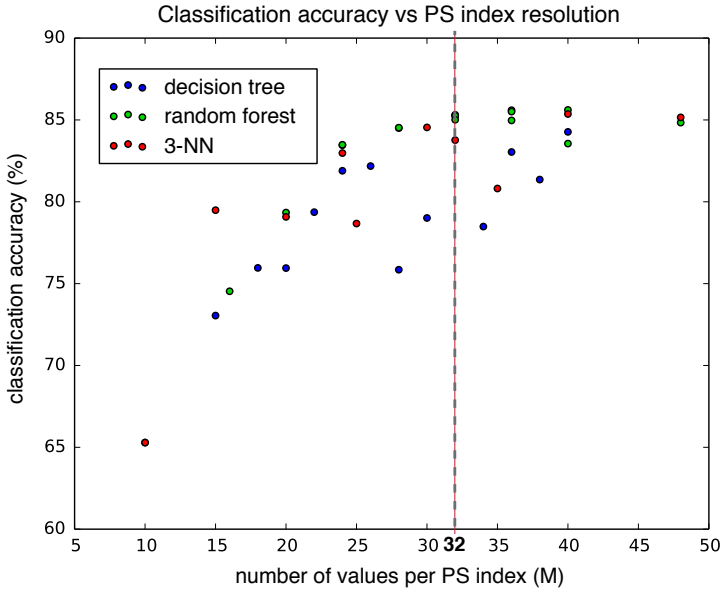

Figure 6.14: The look-up table resolution (M in (6.11)) intuitively has a large impact on the optimality of the resulting plan selection table. We chose $M = 32$ based on the observation that oscillations in accuracy with increasing resolution could be caused by the inherent error due to overfitting of the classifiers, which is "smoothed out" by under-sampling at lower resolutions. Variations in the accuracies of the decision tree and random forest classifiers are expected due to the randomness involved in these algorithms.

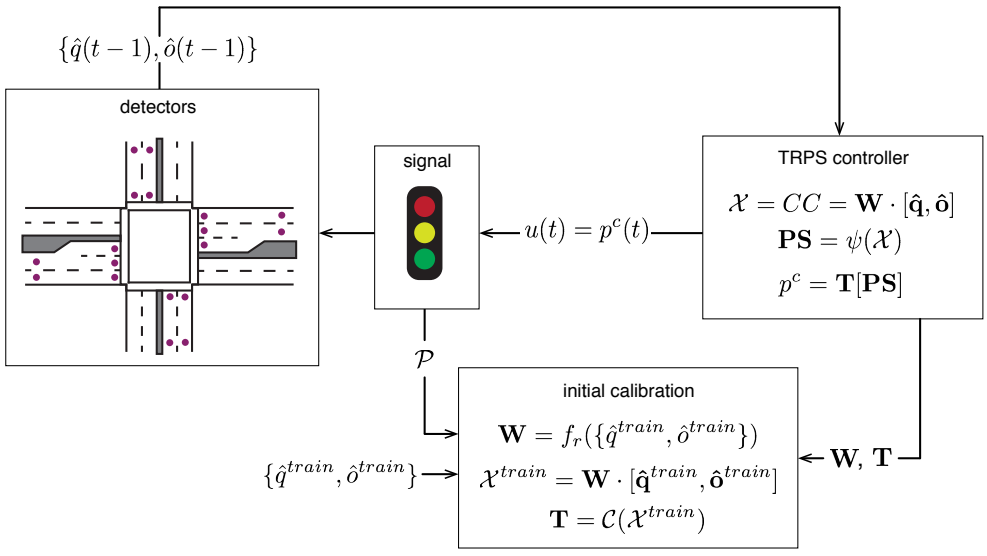

Figure 6.15: We implement a simple feedback controller for our system simulation. The selected plan is a function of volume and occupancy measurements from all available sensors.

Classification accuracy (percent)

classifier	2 LDA features	3 PCA features
decision tree*	**85.2913**	66.4018
random forest*	**85.2358**	74.2851
3-nearest neighbors	83.7657	69.8779

Table 6.1: In general, classification using two LDA features was significantly more successful than that using three PCA features. The decision tree and random forest algorithms were more accurate than k(3)-nearest-neighbors. Values for the decision tree and random forest algorithms (marked with *) are an average of four simulation results.

sensor data, $\{\hat{q}^{train}, \hat{o}^{train}\}$.

$$u(\{\hat{q}(t), \hat{o}(t)\}) = p^c(t) = \mathbf{T}[PS(t^-)] \tag{6.16}$$

$$\text{where} \quad \mathbf{W} = f_r(\{\hat{q}^{train}, \hat{o}^{train}\}) \tag{6.17}$$

$$\mathcal{X}^{train} = \mathbf{W} \cdot [\hat{\mathbf{q}}^{train}, \hat{\mathbf{o}}^{train}] \tag{6.18}$$

$$\mathbf{T} = \mathcal{C}(\mathcal{X}^{train}) \tag{6.19}$$

$$\mathcal{X}(t) = \mathbf{CC}(t) = \mathbf{W} \cdot [\hat{\mathbf{q}}(t), \hat{\mathbf{o}}(t)] \tag{6.20}$$

$$\mathbf{PS}(t) = \psi(\mathcal{X}(t)) \tag{6.21}$$

This is a "simplified" TRPS controller because it does not incorporate any smoothing on the measurements and it does not demonstrate hysteresis boundaries on the plan switching mechanism. Both of these features, if implemented, would theoretically reduce undesirable high-frequency switching behavior. Figure 6.16 shows how our simplified controller often recommends switching plans twice or more in a 15-minute interval.

Table 6.1 lists the final classification accuracies achieved by the simulated TRPS controller referencing a 32×32 plan selection table (for 2 LDA features) or a $32 \times 32 \times 32$ plan selection table (for 3 PCA features). Despite the reduced dimensions, each of the tables generated by the LDA features clearly outperforms the tables indexed by the PCA-based CC parameters. LDA should be the default option for feature reduction—especially if four or more plans are available such that a 3-dimensional table can be produced.

The delay induced by the LDA-feature TRPS controllers was very close to the optimized delay, as seen in Table 6.2. While 0.9 % appears to be a relatively small reduction in delay, absolute delay reductions are far from insignificant. Given "optimal" plan selection, the cumulative total control delay over the entire 12 weeks would be 170,729 vehicle-hours. Our simulated TRPS controller would induce a cumulative total delay of 170,829 vehicle-hours. The scheduled TOD plans, meanwhile, generate 172,385 vehicle-hours of delay over this time period. Assuming a similar reduction given TRPS application on the surrounding 51 signals, the annual delay reduction could be over 300,000 vehicle-hours.

While we recognize that the frequency of switching that is required to achieve these delay reductions is unreasonable, this work serves as a proof-of-concept for a realistic TRPS

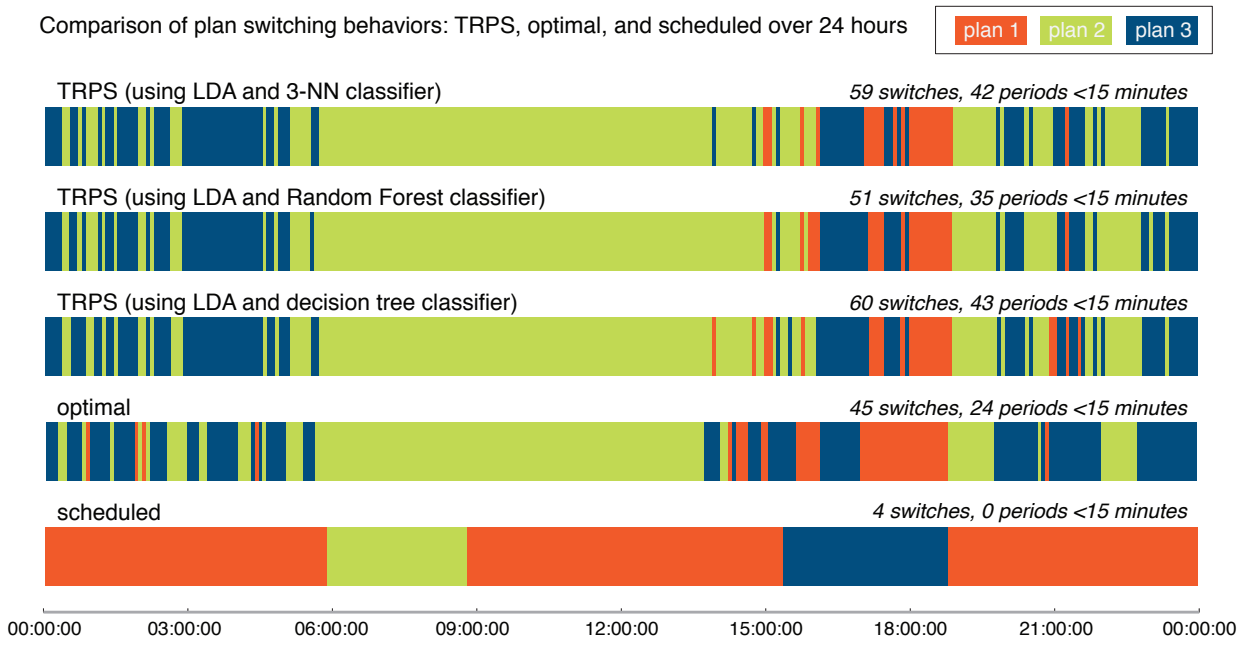

Figure 6.16: In this example day, the "best" performing classifier (decision tree using LCA-derived features) incorporated the most plan switches—even more than the optimal switching pattern. This suggests an overfitting of the data in the classification algorithm. The random forest and k-nearest-neighbor classifiers returned exactly the same results in this case. In practice, a functioning TRPS controller would have to eliminate high-frequency plan switches, as the process of switching between plans will realistically take around 15-20 minutes. More frequent plan switches also have a negative (short-term) impact on the intended progression bandwidths.

Delay reduction (percent, compared to TOD plans)

classifier	2 LDA features	3 PCA features
decision tree*	**0.900502**	0.584689
random forest*	**0.902846**	0.782999
3-nearest neighbors	0.884022	0.697648
optimal	*0.960819*	

Table 6.2: Optimal delay reduction was observed using a decision tree or random forest classifier (with a maximum tree depth of 10) operating on two features that were derived via linear discriminant analysis. Values for the decision tree and random forest algorithms (marked with *) are an average of four simulation results.

calibration mechanism. Only small modifications are required to make the proposed system directly ready for implementation. For example, a simple constraint on switching frequency (and "memory" of recent measurements) would provide a desired smoothing of controller switches. Furthermore, a procedure could be developed to identify situations in which adding an additional plan to the encoded set would significantly increase controller performance.

Chapter 7

Conclusion

The research presented in this dissertation represents a variety of theoretical and practical contributions to the study of signalized traffic networks, but more globally demonstrates the challenges of merging theory into the practice of urban traffic management. In the concluding chapter, we briefly summarize the main contributions presented in this body of work. We then return to the initial motivations that were described in the first chapter, and address the outcomes of our work on the prospect of a comprehensive arterial management system. We finally suggest some directions for future research and development that have arisen from the work compiled in this dissertation.

7.1 Summary of contributions

In Chapter 1, we introduced three primary objectives of arterial traffic management: modeling/prediction, estimation, and control. To conclude the work, we present the contributions of the preceding work in relation to each of these objectives.

Modeling

While the traffic community has (relatively) settled on methods of representing uninterrupted freeway flows using kinematic wave models, no such consensus has been reached in the case of signalized urban roadways. The validation and analysis of a reproducible implementation of two dynamic traffic models in Chapter 3 addresses this fact. To the knowledge of this author, this work represents the first comprehensive evaluation of the performance of CTM on short signalized roadways. It also presents an entirely new formulation of a cell-based vertical queueing model variant specifically for signalized networks called the Vertical Cell Model (VCM). Because no physical model can be expected to reproduce the often-unpredictable instabilities in congestion dynamics with complete accuracy, we demonstrated that a simplified vertical queuing model such as VCM achieves an acceptable compromise

between accuracy and analytical simplicity in its representation of the evolution of arterial traffic state.

Estimation

We introduced an analytically-based state estimation procedure which can incorporate measurements taken from multiple disparate sources into a single model of link dynamics in Chapter 4. Specifically, we extended a methodology that has been previously demonstrated on freeways to an application on signalized roadways through incorporation of constraints to impose the artificial impediment generated by a downstream signal controller. We also presented a method of incorporating end-to-end travel time measurements into the model-based state estimation. This is a more flexible way to incorporate information gained from a GPS trajectory into a queue estimate than techniques which depend on highly-accurate positioning information, as individual trajectories with such high resolution can provide a poor representation of the macroscopic behaviors that are being estimated.

Control

Control was approached from two very different directions in this dissertation.

Chapter 5 presents a theoretically optimal signal controller and extends the controller's underlying modeling framework closer towards an accurate representation of the constraints imposed on existing signal controllers. The contribution of this effort brings max pressure signal control one step closer to reality.

On the other hand, Chapter 6 addresses the very practical problem of calibrating a rigid controller software framework to achieve theoretically-optimal results. Even without any real effort to tune or "optimize" the tools implemented in the proposed calibration procedure, the resulting controller was shown to achieve a delay reduction that was very close to the optimal performance possible with the set of existing signal plans. Because the concept was designed to be applied on existing hardware using the set of signal plans that is already available, the algorithm is something that could be implemented practically immediately.

7.2 Primary challenges identified in the development of a unified urban traffic management system

Perhaps the most significant limiting factors of this work has been a lack of available data for realistic analysis of traffic on signalized road networks. In an era where surveillance (and, in some sense, control) of human behavior is practically ubiquitous, it seems strange that we do not yet have an effective system to monitor and/or control how people use their local transportation networks. Yet consider that currently we do not even have a robust real-time method of monitoring the proportion of vehicles that turn left at any given approach to a

traffic intersection. Such an observation of turn-ratios alone would make a huge differences in the expected accuracies of any type of dynamic model of urban traffic flow.

But the lack of data was not limited to the dynamic characteristics of a road network. In fact, it was even incredibly difficult to compile the (relatively) *static* data that more fundamentally influence vehicle flows, such as lane counts, turn bay capacities, and artificial movement restrictions.

This data gap is often overlooked by academic researchers driving the development of traffic theory, as they can (and sometimes must) use fully-synthetic simulated models to demonstrate their theories and algorithms. Yet in Chapter 3, we concluded that the representation—or misrepresentation—of this type of data data can have more impact on model predictions than the choice of model itself.

While it was not always addressed specifically, this conclusion could also be considered highly applicable to the rest of the work described in this dissertation. For instance, the estimation algorithm described in Chapter 4 not only relies on the inherent assumption that traffic behaves according to the LWR model, but it also assumes that traffic divides into a predictable number of queues corresponding to the number of lanes available in the roadway. For truly realistic results, it would need to be aware, for example, of the ability of right-turning vehicles to consistently bypass queued vehicles to proceed on a red light by creating an unofficial turn lane. While this type of unexpected behavior may seem trivial to a theoretician, it must be considered if we ever hope to depend on a model to accurately predict the onset of intersection spill-back. Considering that one of the objectives of projects such as Integrated Corridor Management is to facilitate operation of a traffic network when demands are very near its theoretical capacity, this desired use-case is actually very possible.

One overarching outcome of this work is therefore a reconsideration of the initial assumptions outlined in the introduction to this work: the first step towards a unified urban traffic management system is not in fact a "comprehensive" model of traffic flow dynamics, but rather a "comprehensive" representation of the road network itself.

There is evidence of past efforts to systematize the geometric data required to accurately represent network dynamics with an incredibly high level of detail. For example, the formula presented in the 2010 Highway Capacity Manual for determining saturation flow at an intersection approach has 11 separate geometric and behavioral adjustment factors. Calculating the value of each adjustment factor would require a distinct and detailed physical measurement of road geometry or typical driver behavior. The presentation of this equation in the HCM is quickly followed by a reference to default values, and a disclaimer that measured saturation flows will be more accurate than the adjustment procedure anyway—which implies that someone with sufficient access to observe all of the physical characteristics required for accurate estimation of saturation flow should just measure it instead!

Before future efforts to use data-driven methods of calibrating physical network parameters, more study needs to be done to determine and standardize the minimum set of information required to accurately represent the characteristics of a signalized road network. Such a process would greatly facilitate future development of sufficiently-accurate yet efficient queueing models for signalized networks.

It is also important to actively seek areas where the traditional heuristic models and traditional beliefs about urban traffic management can be supplemented or even replaced by more rigorous data-driven algorithms, such as the TRPS calibration algorithm presented in Chapter 6. This work clearly demonstrates an area where a departure from traditional methods can make operations much more efficient and effective. But it is also important to recognize that this does not mean that there is no need for the heuristic knowledge of experienced traffic engineers and technicians. In fact, our simplified implementation of the TRPS controller could greatly benefit from advice of traffic practitioners to improve the realism of our modeled objective function and assumptions on reasonable switching behaviors.

7.3 Future directions

The original motivation behind developing VCM was in fact to simplify the pathway towards applying traditional control techniques derived from linear systems theory to traffic signals. This is still a very valid objective for future work. The linear link model is attractive for analysis and estimation purposes as well. In fact, existing efforts at feedback-based queue estimation applied a Kalman filter to a similar vertical queueing representation using measurements that could be realistically available [44, 206]. It would be valuable to explore how such an approach to queue estimation could be extended to incorporate measurements from neighboring links using the VCM framework.

In the same vein, the next logical step for the Moskowitz PDE-based queue estimation technique is an extension to a network of signalized intersections which are realistically coupled such that a queue length estimate could be informed not only by the measurements that are within the domain of a single link, but also by those taken on upstream links. Research would be required to determine how to best project upstream measurements onto the area of analysis in a way that is consistent with the assumptions of the underlying LWR model.

There is a great deal of work that would need to be done to make the theoretically-beneficial max pressure controller more practically applicable. Unfortunately, there are many unrealistic assumptions of the inherent store-and-forward modeling framework that must be addressed before the theoretical guarantees of optimality can be directly translated to performance on a road network. One of the biggest assumptions is the infinite storage capacity of network links. An extension of the modeling framework to handle finite link storage is challenging, as it introduces an additional non-linearity (and downstream dependence) on link discharge rates. Yet such an extension would be a huge step forward in the relevant line of research.

Another unrealistic assumption which limits the application of max pressure is that the controllers have knowledge of the expected intersection turning ratios. In fact, almost all of the algorithms discussed in this dissertation have required estimation of turning proportions. Turn ratio estimation using existing sources of data has been explored from many angles by

the traffic community, yet no strong consensus on the optimal approach has been reached. Emerging data sources such as those discussed at the end of Chapter 2 offer new promise in this line of work; researchers should continue to explore creative ways to use probe vehicles and advanced detectors to improve estimation of network parameters as well as congestion state.

The TRPS calibration procedure described in Chapter 6 remains only a proof-of-concept. Areas for improvement are numerous: classifier parameters could be optimized, mechanisms could be implemented to reduce high-frequency switching, and, importantly, the benefit analysis should be extended to implementation on a network of coordinated signals with a representation of progression (or number of stops) incorporated into the objective function.

The lessons learned from this process could also motivate the search for additional sub-problems in arterial traffic operations where, given increased availability of data from signals and traffic sensors, model-less estimation or control techniques could be applied. Because of the challenges associated with modeling signalized intersections, algorithms that are not explicitly dependent on a specific model of traffic dynamics have wide appeal. The suggestion of new use cases for high-resolution sensor data would furthermore motivate traffic operators to begin to centralize and archive the data from their signalized networks, which could benefit all of the primary objectives of the work described in this dissertation.

Bibliography

[1] M. M. Abbas, S. Abdelaziz, and C. C. McGhee. *Evaluation of Traffic Responsive Control on the Reston Parkway Arterial Network*. Tech. rep. VTRC 09-CR6. Virginia Department of Transportation, Feb. 2009.

[2] M. M. Abbas and A. Sharma. "Multiobjective Plan Selection Optimization for Traffic Responsive Control". In: *Journal of Transportation Engineering* 132.5 (2006), pp. 376–384.

[3] M. M. Abbas et al. *Methodology for Determination of Optimal Traffic Responsive Plan Selection Control Parameters*. Tech. rep. FHWA/TX-04/0-4421-1. Texas Department of Transportation, Research and Technology Implementation Office, June 2004.

[4] K Aboudolas, M. Papageorgiou, and E Kosmatopoulos. "Store-and-forward based methods for the signal control problem in large-scale congested urban networks". In: *Transportation Research Part C: Emerging Technologies* 17 (2009), pp. 163–174.

[5] K Aboudolas, M. Papageorgiou, and A. Kouvelas. "A rolling-horizon quadratic programming approach to the signal control problem in large-scale congested urban road networks". In: *Transportation Research Part C: Emerging Technologies* 18 (2010), pp. 680–694.

[6] G. Abu-Lebdeh and R. F. Benekohal. "Development of traffic control and queue management procedures for oversaturated arterials". In: *Transportation Research Record: Journal of the Transportation Research Board* 1603 (1997), pp. 119–127.

[7] *ACS-Lite: The Next Generation of Traffic Signal Control*. Federal Highway Administration. Apr. 2007.

[8] V. Alexiadis et al. *Guidance on the Level of Effort Required to Conduct Traffic Analysis Using Microsimulation*. Tech. rep. FHWA-HRT-13-026. Federal Highway Administration, Mar. 2014.

[9] R. E. Allsop. "Estimating the traffic capacity of a signalized road junction". In: *Transportation Research* 6.3 (1972), pp. 245–255.

[10] L. Anderson, G. Gomes, and A. M. Bayen. "Evaluation of horizontal and vertical queueing models in relation to observed trajectory data in a signalized urban traffic network". In: *Proceedings of the 94th Annual Meeting of the Transportation Research Board*. Jan. 2015.

[11] L. Anderson et al. "Optimization-based queue estimation on an arterial traffic link with measurement uncertainties". In: *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*. Jan. 2014.

[12] L. Anderson et al. "Stability and Implementation of a Cycle-based Max Pressure Controller for Signalized Traffic Networks".

[13] M. Andrews et al. "Scheduling in a queuing system with asynchronously varying service rates". In: *Probability in the Engineering and Informational Sciences* 18.2 (2004), pp. 191–217.

[14] J. Argote et al. "Estimation of Arterial Measures of Effectiveness with Connected Vehicle Data". In: *Proceedings of the 91st Annual Meeting of the Transportation Research Board*. Washington, DC, USA, Jan. 2012.

[15] J.-P. Aubin, A. M. Bayen, and P. Saint-Pierre. "Dirichlet Problems for some Hamilton–Jacobi Equations with Inequality Constraints". In: *SIAM Journal on Control and Optimization* 47.5 (Jan. 2008), pp. 2348–2380.

[16] X. J. Ban, P. Hao, and Z. Sun. "Real time queue length estimation for signalized intersections using travel times from mobile sensors". In: *Transportation Research Part C: Emerging Technologies* 19.6 (Oct. 2011), pp. 1133–1156.

[17] H. Bar-Gera. "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel". In: *Transportation Research Part C: Emerging Technologies* 15.6 (2007), pp. 380–391.

[18] E. N. Barron and R Jensen. "Semicontinuous Viscosity Solutions For Hamilton–Jacobi Equations With Convex Hamiltonians". In: *Communications in Partial Differential Equations* 15.12 (Jan. 1990), pp. 293–309.

[19] C. Beard and A. Ziliaskopoulos. "System optimal signal optimization formulation". In: *Transportation Research Record: Journal of the Transportation Research Board* 1978 (2006), pp. 102–112.

[20] M Beckmann, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*. New Haven: Yale University Press, 1956.

[21] P. N. Belhumeur, J. P. Hespanha, and J Kriengman. "Eigenfaces vs fisherfaces recognition using class specific linear projections". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7 (1997), pp. 711–720.

[22] M van den Berg, A Hegyi, and B. De Schutter. "Integrated traffic control for mixed urban and freeway networks: A model predictive control approach". In: *European Journal of Transport and Infrastructure Research* 7 (2007), pp. 223–250.

[23] M van den Berg et al. "A macroscopic traffic flow model for integrated control of freeway and urban traffic networks". In: *Proceedings of the 42nd IEEE Conference on Decision and Control*. IEEE, 2003, pp. 2774–2779.

[24] C. Berge and A Ghouila-Houri. *Programming, Games and Transportation Networks*. Methuen. London, 1965.

[25] M. Berkow, M. Wolfe, and C. M. Monsere. "Using signal system data and buses as probe vehicles to define the congested regime on arterials". In: *Proceedings of the 87th Annual Meeting of the Transporation Research Board*. Washington, DC, 2008.

[26] M. Berkow et al. "Prototype for Data Fusion Using Stationary and Mobile Data". In: *Transportation Research Record: Journal of the Transportation Research Board* 2099 (Sept. 2009), pp. 102–112.

[27] A. Bhaskar, E. Chung, and A.-G. Dumont. "Fusing Loop Detector and Probe Vehicle Data to Estimate Travel Time Statistics on Signalized Urban Networks". In: *Computer-Aided Civil and Infrastructure Engineering* 26.6 (2011), pp. 433–450.

[28] S. Blandin et al. "On sequential data assimilation for scalar macroscopic traffic flow models". In: *Physica D* 241.17 (Sept. 2012), pp. 1421–1440.

[29] F. Boillot, S. Midenet, and J.-C. Pierrelée. "The real-time urban traffic control system CRONOS: Algorithm and experiments". In: *Transportation Research Part C: Emerging Technologies* 14 (Feb. 2006), pp. 18–38.

[30] F. Boillot et al. "Optimal signal control of urban traffic networks". In: *6th International Conference on Road Traffic Monitoring and Control*. IET, 1992, pp. 75–79.

[31] J. A. Bonneson, A. Sharma, and D. M. Bullock. *Measuring the Performance of Automobile Traffic on Urban Streets*. Tech. rep. 3-79. National Cooperative Highway Research Program, Transportation Research Board of The National Academies, Jan. 2008.

[32] D. Boyce. "Forecasting Travel on Congested Urban Transportation Networks: Review and Prospects for Network Equilibrium Models". In: *Networks and Spatial Economics* 7.2 (Jan. 2007), pp. 99–128.

[33] M. Brackstone and M. McDonald. "Car-following: a historical review". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 2.4 (1999), pp. 181–196.

[34] L Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[35] L. Breiman et al. *Classification and Regression Trees*. Chapman and Hall/CRC, Jan. 1984.

[36] R. Brockett. "Stabilization of motor networks". In: *Proceedings of the 34th IEEE Conference on Decision and Control, 1995*. IEEE, 1995, pp. 1484–1488.

[37] E. S. Canepa and C. G. Claudel. "Exact solutions to traffic density estimation problems involving the Lighthill-Whitham-Richards traffic flow model using Mixed Integer Programming". In: *Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems*. 2012.

[38] C. G. Claudel and A. M. Bayen. "Convex Formulations of Data Assimilation Problems for a Class of Hamilton–Jacobi Equations". In: *SIAM Journal on Control and Optimization* 49.2 (Jan. 2011), pp. 383–402.

[39] C. G. Claudel and A. M. Bayen. "Lax-Hopf Based Incorporation of Internal Boundary Conditions Into Hamilton-Jacobi Equation. Part I: Theory". In: *IEEE Transactions on Automatic Control* 55.5 (May 2010), pp. 1142–1157.

[40] G. Comert and M. Cetin. "Analytical Evaluation of the Error in Queue Length Estimation at Traffic Signals From Probe Vehicle Data". In: *IEEE Transactions on Intelligent Transportation Systems* 12.2 (June 2011), pp. 563–573.

[41] G. Comert and M. Cetin. "Queue length estimation from probe vehicle location and the impacts of sample size". In: *European Journal of Operational Research* 197.1 (Aug. 2009), pp. 196–202.

[42] *Congestion Reduction Toolbox*. URL: `http://www.fhwa.dot.gov/congestion/toolbox/service.htm` (visited on 05/11/2015).

[43] *CORSIM Microscopic Traffic Simulation Model*. URL: `http://www-mctrans.ce.ufl.edu/featured/TSIS/Version5/corsim.htm` (visited on 05/11/2015).

[44] M. Cremer and T. C. Henninger. "Estimation of queue lengths in urban road networks". In: *3rd Pacific Rim TransTech Conference*. Seattle, 1993, pp. 29–35.

[45] W. B. Cronje. "Analysis of existing formulas for delay, overflow, and stops". In: *Transportation Research Record: Journal of the Transportation Research Board* 905 (1983), pp. 89–93.

[46] C. F. Daganzo. "A variational formulation of kinematic waves: basic theory and complex boundary conditions". In: *Transportation Research Part B: Methodological* 39.2 (Feb. 2005), pp. 187–196.

[47] C. F. Daganzo. "On the variational theory of traffic flow: well-posedness, duality and applications". In: *Networks and Heterogeneous Media* 1.4 (2006), pp. 601–619.

[48] C. F. Daganzo. "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory". In: *Transportation Research Part B: Methodological* 28B.4 (1994), pp. 296–287.

[49] C. F. Daganzo. "The cell transmission model, part II: network traffic". In: *Transportation Research Part B: Methodological* 29B (1995), pp. 79–93.

[50] J. G. Dai and W. Q. Lin. "Maximum pressure policies in Stochastic processing networks". In: *Operations Research* 53.2 (2005), pp. 197–218.

[51] X. Dai, M. A. Ferman, and R. P. Roesser. "A simulation evaluation of a real-time traffic information system using probe vehicles". In: *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2003, pp. 475–480.

[52] G. C. D'Ans and D. C. Gazis. "Optimal control of oversaturated store-and-forward transportation networks". In: *Transportation Science* 10.1 (1976), pp. 1–19.

[53] E. J. Davison and U Ozguner. "Decentralized control of traffic networks". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.4 (1983), pp. 476–487.

[54] C. M. Day and D. M. Bullock. *Final Report: Performance Based Management of Arterial Traffic Signal Systems*. Tech. rep. 3-79a. National Cooperative Highway Research Program, Aug. 2010.

[55] B. De Schutter and B De Moor. "Optimal Traffic Light Control for a Single Intersection". In: *European Journal of Control* 4.3 (1998), pp. 260–276.

[56] P Dell'Olmo and P. B. Mirchandani. "REALBAND: An approach for real-time coordination of traffic flows on networks". In: *Transportation Research Record: Journal of the Transportation Research Board* 1494 (1995), pp. 106–116.

[57] C. Diakaki, M. Papageorgiou, and K Aboudolas. "A multivariable regulator approach to traffic-responsive network-wide signal control". In: *Control Engineering Practice* 10.2 (2002), pp. 183–195.

[58] C. Diakaki, M. Papageorgiou, and T McLean. "Integrated traffic-responsive urban corridor control strategy in Glasgow, Scotland - Application and evaluation". In: *Transportation Research Record: Journal of the Transportation Research Board* 1727 (2000), pp. 101–111.

[59] F. Dion, H. A. Rakha, and Y.-S. Kang. "Comparison of delay estimates at undersaturated and over-saturated pre-timed signalized intersections". In: *Transportation Research Part B: Methodological* 38.2 (Feb. 2004), pp. 99–122.

[60] M. C. Dunne and R. B. Potts. "Algorithm for Traffic Control". In: *Operations Research* 12.6 (Nov. 1964), pp. 870–881.

[61] U. Dutta et al. *Evaluation of the SCATS control system*. Tech. rep. RC-1545. Michigan Department of Transportation, Dec. 2008.

[62] M Egerstedt and Y Wardi. "Multi-process control using queuing theory". In: *Proceedings of the 41st IEEE Conference on Decision and Control, 2002*. IEEE, 2002, pp. 1991–1996.

[63] J. L. Farges, I Khoudour, and J. B. Lesort. "PRODYN: on site evaluation". In: *Road Traffic Control, 1990., Third International Conference on*. IET, 1990, pp. 62–66.

[64] Federal Highway Administration. *It's About Time... Traffic Signal Management*. 2001. URL: `https://archive.org/details/gov.fhwa.ttp.vh-434` (visited on 05/11/2015).

[65] M. Fellendorf and P. Vortisch. "Microscopic Traffic Flow Simulator VISSIM". In: *Fundamentals of Traffic Simulation*. New York, NY: Springer New York, June 2010, pp. 63–93.

[66] L. R. Ford and D. R. Fulkerson. *Flows in Networks*. Tech. rep. R-375-PR. Santa Monica, CA, USA: United States Air Force Project Rand, Aug. 1962.

[67] H. Frankowska. "Lower Semicontinuous Solutions of Hamilton-Jacobi-Bellman Equations". In: *SIAM Journal on Control and Optimization* 31.1 (Jan. 1993), pp. 257–272.

[68] L. Fu, B. Hellinga, and Y. Zhu. "An adaptive model for real-time estimation of overflow queues on congested arterials". In: *2001 IEEE Intelligent Transportation Systems Conference Proceedings*. Oakland, CA, USA: IEEE, 2001, pp. 219–226.

[69] *Functional Description: Urban Traffic Control System*. Tech. rep. FHWA-TS-79-228. Federal Highway Administration, Aug. 1979.

[70] N. H. Gartner. "OPAC: A demand-responsive strategy for traffic signal control". In: *Transportation Research Record: Journal of the Transportation Research Board* 906 (1983), pp. 75–81.

[71] N. H. Gartner, F. J. Pooran, and C. M. Andrews. "Implementation of the OPAC adaptive control strategy in a traffic signal network". In: *2001 IEEE Intelligent Transportation Systems Conference Proceedings*. Oakland, CA USA: IEEE, 2001, pp. 195–200.

[72] D. C. Gazis. "Modeling and optimal control of congested transportation systems". In: *Networks* 4 (1974), pp. 113–124.

[73] D. C. Gazis. "Optimum Control of a System of Oversaturated Intersections". In: *Operations Research* 12.6 (1964), pp. 815–831.

[74] D. C. Gazis and R. B. Potts. *The Oversaturated Intersection*. Tech. rep. RC-929. Yorktown Heights, NY: IBM Research Center, Apr. 1963.

[75] N. Geroliminis and A. Skabardonis. "Identification and Analysis of Queue Spillovers in City Street Networks". In: *IEEE Transactions on Intelligent Transportation Systems* 12.4 (Nov. 2011), pp. 1107–1115.

[76] P Giaccone, E Leonardi, and D Shah. "On the maximal throughput of networks with finite buffers and its application to buffered crossbars". In: *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE, 2005, pp. 971–980.

[77] S. K. Godunov. "A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics". In: *Matematicheskiĭ sbornik* 47(89).3 (1959), pp. 271–306.

[78] G. Gomes and R. Horowitz. "Optimal freeway ramp metering using the asymmetric cell transmission model". In: *Transportation Research Part C: Emerging Technologies* 14.4 (Aug. 2006), pp. 244–262.

[79] G. Gomes et al. "Behavior of the cell transmission model and effectiveness of ramp metering". In: *Transportation Research Part C: Emerging Technologies* 16.4 (Aug. 2008), pp. 485–513.

[80] H. Greenberg. "An analysis of traffic flow". In: *Operations Research* 7.1 (1959), pp. 79–85.

[81] B. D. Greenshields, W Channing, and H Miller. "A Study of Traffic Capacity". In: *Proceedings of the 14th Annual Meeting of the Highway Research Board*. 1935, pp. 488–477.

[82] F Gundogan and M Fellendorf. "Pattern recognition method for simplified coordinated traffic signal control". In: *2011 IEEE Forum on Integrated and Sustainable Transportation System*. Vienna, Austria: IEEE, July 2011, pp. 283–288.

[83] A. Haoui, R. Kavaler, and P. Varaiya. "Wireless magnetic sensors for traffic surveillance". In: *Transportation Research Part C: Emerging Technologies* 16.3 (June 2008), pp. 294–306.

[84] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Data Mining, Inference, and Prediction, Second Edition. New York, NY: Springer Science & Business Media, Aug. 2009.

[85] *HCM2010: Highway Capacity Manual*. Vol. Special Report 209. Transportation Research Board of the National Academies.

[86] K. L. Head, P. B. Mirchandani, and D. Sheppard. "Hierarchical Framework for Real-time Traffic Control". In: *Transportation Research Record: Journal of the Transportation Research Board* (1992), pp. 82–88.

[87] B. Hellinga and R Gudapati. "Estimating link travel times for advanced traveler information systems". In: *Proceedings of the Canadian Society of Civil Engineers, 3rd Transportation Specialty Conference*. 2000.

[88] B. Hellinga et al. "Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments". In: *Transportation Research Part C: Emerging Technologies* 16.6 (Dec. 2008), pp. 768–782.

[89] D. Henry. *Signal Timing on a Shoestring*. Tech. rep. FHWA-HOP-07-006. Federal Highway Administration, Mar. 2005.

[90] J. J. Henry, J. L. Farges, and J Tuffal. "The PRODYN Real-Time Traffic Algorithm". In: *IFAC Control in Transportation Systems*. Baden-Baden, Germany: Elsevier, 1984, pp. 305–310.

[91] J. C. Herrera and A. M. Bayen. "Incorporation of Lagrangian measurements in freeway traffic state estimation". In: *Transportation Research Part B: Methodological* 44.4 (May 2010), pp. 460–481.

[92] J. C. Herrera et al. "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment". In: *Transportation Research Part C: Emerging Technologies* 18.4 (Aug. 2010), pp. 568–583.

[93] J.-C. Herrera and A. M. Bayen. "Traffic flow reconstruction using mobile sensors and loop detector data". In: *Proceedings of the 87th Annual Meeting of the Transportation Research Board*. Jan. 2008, pp. 1–18.

[94] *Highway Safety Information System, California database*. URL: `http://www.hsisinfo.org/` (visited on 05/11/2015).

[95] A. Hofleitner and A. M. Bayen. "Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model". In: *14th International IEEE Conference on Intelligent Transportation Systems*. Washington, DC USA: IEEE, 2011, pp. 815–821.

[96] A. Hofleitner, C. G. Claudel, and A. M. Bayen. "Reconstruction of boundary conditions from internal conditions using viability theory". In: *Proceedings of the 2012 American Control Conference*. Fairmont Queen Elizabeth, Montreal, Canada, May 2012, pp. 640–645.

[97] S. Hoogendoorn and P. H. L. Bovy. "State-of-the-art of vehicular traffic flow modelling". In: *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering* 215.4 (June 2001), pp. 283–303.

[98] H. Hotelling. *Analysis of a Complex of Statistical Variables Into Principal Components*. Warwick & York, Inc. 1933.

[99] T. C. Hu. *Integer programming and network flows*. Reading, Massachusetts: Addison-Wesley, 1969.

[100] P. B. Hunt et al. *SCOOT - a Traffic Responsive Method of Coordinating Signals*. Transport and Road Research Laboratory. UK, 1981.

[101] T. Hunter, P. Abbeel, and A. M. Bayen. "The Path Inference Filter: Model-Based Low-Latency Map Matching of Probe Vehicle Data". In: *IEEE Transactions on Intelligent Transportation Systems* 15.2 (Mar. 2014), pp. 507–529.

[102] T. P. Hutchinson. "Delay at a Fixed Time Traffic Signal—II: Numerical Comparisons of some Theoretical Expressions". In: *Transportation Science* 6.3 (1972), pp. 286–305.

[103] H Ishii and B. A. Francis. "Stabilizing a linear system by switching control with dwell time". In: *Proceedings of the 2001 American Control Conference*. Arlington, VA, USA: IEEE, 2001, pp. 1876–1881.

[104] I Jolliffe. *Principal component analysis*. Wiley StatsRef: Statistics Reference Online. 2002.

[105] H. R. Kashani and G. N. Saridis. "Intelligent Control for Urban Traffic Systems". In: *Automatica* 19.2 (1983), pp. 191–197.

[106] R. Kavaler et al. "Arterial Performance Measurement System with Wireless Magnetic Sensors". In: *First International Conference on Transportation Information and Safety*. Reston, VA: American Society of Civil Engineers, Apr. 2012, pp. 377–385.

[107] L. A. Klein, M. K. Mills, and D. R. P. Gibson. *Traffic Detector Handbook: Third Edition-Volume 1*. Tech. rep. FHWA-HRT-06-108. Federal Highway Administration, Oct. 2006.

[108] F. H. Knight. "Some Fallacies in the Interpretation of Social Cost". In: *The Quarterly Journal of Economics* 38.4 (Aug. 1924), pp. 582–606.

[109] P. Koonce. *Traffic Signal Timing Manual*. Tech. rep. FHWA-HOP-08-024. Federal Highway Administration, June 2008.

[110] A. Kouvelas et al. "Maximum Pressure Controller for Stabilizing Queues in Signalized Arterial Networks". In: *Transportation Research Record: Journal of the Transportation Research Board* 2421 (Oct. 2014), pp. 133–141.

[111] H. W. Kuhn and A. W. Tucker. "Nonlinear Programming". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. University of California Press. Berkeley, CA USA, 1951, pp. 481–492.

[112] K. Kwong et al. "Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors". In: *Transportation Research Part C: Emerging Technologies* 17.6 (Dec. 2009), pp. 586–606.

[113] K. Kwong et al. "Real-Time Measurement of Link Vehicle Count and Travel Time in a Road Network". In: *IEEE Transactions on Intelligent Transportation Systems* 11.4 (Nov. 2010), pp. 814–825.

[114] M. Kyte et al. "Testing Incremental Queue Accumulation Method Using Lankershim Boulevard NGSIM Data Set: A Replacement for HCM Signalized Intersection Uniform Delay and Queue Method in Los Angeles, California". In: *Transportation Research Record: Journal of the Transportation Research Board* 2071 (Jan. 2009), pp. 63–70.

[115] J.-P. Lebacque. "The Godunov scheme and what it means for first order traffic flow models". In: *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*. Ed. by J. B. Lesort. College Park, Maryland, USA: Elsevier, 1996, pp. 647–677.

[116] M. J. Lighthill and G. B. Whitham. "On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 229.1178 (May 1955), pp. 317–345.

[117] J. H. Lim et al. "Hierarchical optimal control of oversaturated urban traffic networks". In: *International Journal of Control* 33.4 (Apr. 1981), pp. 727–737.

[118] S. Lin and Y. Xi. "An Efficient Model for Urban Traffic Network Control". In: *Proceedings of the 17th World Congress of the International Federation of Automatic Control*. Seoul, Korea, July 2008, pp. 14066–14071.

[119] S. Lin et al. "Efficient network-wide model-based predictive control for urban traffic networks". In: *Transportation Research Part C: Emerging Technologies* 24 (Oct. 2012), pp. 122–140.

[120] W. H. Lin and C Wang. "An Enhanced 0-1 Mixed-Integer LP Formulation for Traffic Signal Control". In: *IEEE Transactions on Intelligent Transportation Systems* 5.4 (Dec. 2004), pp. 238–245.

[121] J. D. C. Little. "The Synchronization of Traffic Signals by Mixed-Integer Linear Programming". In: *Operations Research* 14.4 (1966), pp. 568–594.

[122] J. D. C. Little, M. D. Kelson, and N. H. Gartner. "MAXBAND: A Versatile Program for Setting Signals on Arteries and Triangular Networks". In: *Transportation Research Record: Journal of the Transportation Research Board* 795 (Jan. 1981).

[123] H. X. Liu, X. Wu, and P. G. Michalopoulos. "Improving Queue Size Estimation for Minnesota's Stratified Zone Metering Strategy". In: *Transportation Research Record: Journal of the Transportation Research Board* 2012 (2007), pp. 38–46.

[124] H. X. Liu et al. "Real-time queue length estimation for congested signalized intersections". In: *Transportation Research Part C: Emerging Technologies* 17.4 (Aug. 2009), pp. 412–427.

[125] H. X. Liu et al. "SMART-SIGNAL: Systematic Monitoring of Arterial Road Traffic Signals". In: *2008 11th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2008, pp. 1061–1066.

[126] H. K. Lo. "A cell-based traffic control formulation: strategies and benefits of dynamic timing plans". In: *Transportation Science* 35.2 (2001), pp. 148–164.

[127] H. K. Lo. "A novel traffic signal control formulation". In: *Transportation Research Part A: Policy and Practice* 33 (1999), pp. 433–448.

[128] H. K. Lo, E. Chang, and Y. C. Chan. "Dynamic network traffic control". In: *Transportation Research Part A: Policy and Practice* 35 (2001), pp. 721–744.

[129] D Longley. "A Control Strategy for a Congested Computer-Controlled Traffic Network". In: *Transportation Research* 2.4 (1968), pp. 391–408.

[130] P. R. Lowrie. *SCATS: Sydney Coordinated Adaptive Traffic System: a traffic responsive method of controlling urban traffic. Roads and Traffic Authority*. Tech. rep. 1990.

[131] J. Y. Luk, A. G. Sims, and P. R. Lowrie. "SCATS-application and field comparison with a TRANSYT- optimised fixed time system". In: *International Conference on Road Traffic Signalling*. London, UK, 1982, pp. 71–74.

[132] M. Maher. "A comparison of the use of the cell transmission and platoon dispersion models in TRANSYT 13". In: *Transportation Planning and Technology* 34.1 (Feb. 2011), pp. 71–85.

[133] A. M. Martinez and A. C. Kak. "PCA versus LDA". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.2 (2001), pp. 228–233.

[134] *MATSim: Multi-Agent Transport Simulation*. URL: `http://www.matsim.org/` (visited on 05/11/2015).

[135] V Mauro and C. d. Taranto. "UTOPIA". In: *Proceedings of the 6th IFAC symposium on Control, computers, communications in transportation*. 1990, pp. 245–252.

[136] P.-E. Mazaré et al. "Analytical and grid-free solutions to the Lighthill-Whitham-Richards traffic flow model". In: *Transportation Research Part B: Methodological* 45.10 (2011), pp. 1727–1748.

[137] P.-E. Mazaré et al. "Trade-offs between inductive loops and GPS probe vehicles for travel time estimation: A Mobile Century case study". In: *Proceedings of the 91st Annual Meeting of the Transportation Research Board*. Washington, DC, USA, 2012.

[138] D. R. McNeil. "A Solution to the Fixed-Cycle Traffic Light Problem for Compound Poisson Arrivals". In: *Journal of Applied Probability* 5.3 (1968), pp. 624–635.

[139] C. J. Messer et al. "A Variable Sequence Multiphase Progression Optimization Program". In: *Highway Research Record* 445 (1973), pp. 24–33.

[140] P. G. Michalopoulos and G. Stephanopoulos. "Oversaturated signal systems with queue length constraints—I: Single intersection". In: *Transportation Research* 11 (1977), pp. 413–421.

[141] P. G. Michalopoulos and G. Stephanopoulos. "Oversaturated signal systems with queue length constraints—II: Systems of intersections". In: *Transportation Research* 11 (1977), pp. 423–428.

[142] P. G. Michalopoulos, G. Stephanopoulos, and V. B. Pisharody. "Modeling of Traffic Flow at Signalized Links". In: *Transportation Science* 14.1 (1980), pp. 9–41.

[143] A. J. Miller. *Australian road capacity guide : provisional introduction and signalized intersections*. Australian Road Research Board. Vermont South, Victoria, 1968.

[144] A. J. Miller. "Settings for Fixed-Cycle Traffic Signals". In: *Operational Research Quarterly* 14.4 (Dec. 1963), pp. 373–386.

[145] P. B. Mirchandani and K. L. Head. "A real-time traffic signal control system: architecture, algorithms, and analysis". In: *Transportation Research Part C: Emerging Technologies* 9.6 (Dec. 2001), pp. 415–432.

[146] J. E. Moore II et al. "Anaheim Advanced Traffic Control System Field Operations Test: A Technical Evaluation of SCOOT". In: *Transportation Planning and Technology* 28.6 (Dec. 2005), pp. 465–482.

[147] J. T. Morgan and J. D. C. Little. "Synchronizing Traffic Signals for Maximal Bandwidth". In: *Operations Research* 12.6 (1964), pp. 896–912.

[148] L Munoz et al. "Traffic density estimation with the cell transmission model". In: *Proceedings of the 2003 American Control Conference*. Denver, Colorado, USA: IEEE, 2003, pp. 3750–3755.

[149] H. Murase and S. K. Nayar. "Visual learning and recognition of 3-d objects from appearance". In: *International Journal of Computer Vision* 14.1 (1995), pp. 5–24.

[150] *National Traffic Signal Report Card*. Tech. rep. National Transportation Operations Coalition, May 2012.

[151] *National Traffic Signal Report Card: Executive Summary*. Tech. rep. National Transportation Operations Coalition, May 2012.

[152] M. J. Neely, E Modiano, and C. E. Rohrs. "Dynamic power allocation and routing for time-varying wireless networks". In: *IEEE Journal on Selected Areas in Communications* 23.1 (Jan. 2005), pp. 89–103.

[153] E. J. Nelson et al. *Development of Closed Loop System Evaluation Procedures*. Tech. rep. FHWA/IN/JTRP-2000/5. Indiana Department of Transportation, Dec. 2000.

[154] *NEMA Standards Publication No. TS-1: Traffic Control Systems*. 1989.

[155] G. F. Newell. "A simplified theory of kinematic waves in highway traffic, part I: General theory". In: *Transportation Research Part B: Methodological* 27.4 (Aug. 1993), pp. 281–287.

[156] G. F. Newell. "A simplified theory of kinematic waves in highway traffic, Part II: Queueing at freeway bottlenecks". In: *Transportation Research Part B: Methodological* 27.4 (1993), pp. 289–303.

[157] *Next Generation Simulation (NGSIM)*. URL: `http://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm` (visited on 05/11/2015).

[158] *NGSIM FACTSheet: Lankershim Boulevard Dataset*. Tech. rep. FHWA-HRT-07-029. Federal Highway Administration, July 2007.

[159] K Ohno. "Computational algorithm for a fixed cycle traffic signal and new approximate expressions for average delay". In: *Transportation Science* 12.1 (1978), pp. 29–47.

[160] M. Pajic and G. J. Pappas. "Stabilizability over Deterministic Relay Networks". In: *Proceedings of the 52nd IEEE Conference on Decision and Control*. Florence, Italy, 2013, pp. 4018–4023.

[161] M. Papageorgiou. "An integrated control approach for traffic corridors". In: *Transportation Research Part C: Emerging Technologies* 3.1 (1995), pp. 19–30.

[162] M. Papageorgiou and P. Varaiya. "Link Vehicle-Count–the Missing Measurement for Traffic Control". In: *Proceedings of the 12th IFAC Symposium on Transportation Systems*. Redondo Beach, CA, USA, Sept. 2009, pp. 224–229.

[163] M. Papageorgiou et al. "Review of road traffic control strategies". In: *Proceedings of the IEEE* 91.12 (Dec. 2003), pp. 2043–2067.

[164] *Paramics Microsimulation*. URL: `http://www.sias.com/2013/sp/sparamicshome.htm` (visited on 05/11/2015).

[165] B. B. Park, P Santra, and I Yun. "Optimization of time-of-day breakpoints for better traffic signal control". In: *Transportation Research Record: Journal of the Transportation Research Board* 1867.1 (2004), pp. 217–223.

[166] E. S. Park et al. "Hierarchical optimal control of urban traffic networks". In: *International Journal of Control* 40.4 (Oct. 1984), pp. 813–829.

[167] K. Pearson. "On lines and planes of closest fit to systems of points in space ". In: *Philosophical Magazine* 2.11 (1901), pp. 559–572.

[168] G. Pesti, M. M. Abbas, and N. A. Chaudhary. "Traffic State Classification in Condition-Responsive Traffic Control". In: *2007 International Conference on Intelligent Engineering Systems*. Budapest, Hungary: IEEE, July 2007, pp. 33–37.

[169] L. J. Pignataro et al. *Traffic Control in Oversaturated Street Networks*. Tech. rep. Transportation Research Board of the National Academies, 1978.

[170] A. C. Pigou. *The Economics of Welfare*. London, England: MacMillan, 1920.

[171] T. Pumir, L. Anderson, and A. M. Bayen. "Stability of Modified Max Pressure Controller with Application to Signalized Traffic Networks". In: *Proceedings of the 2015 American Control Conference (to appear)*.

[172] M. Ramezani and N. Geroliminis. "On the estimation of arterial route travel time distribution with Markov chains". In: *Transportation Research Part B: Methodological* 46.10 (Dec. 2012), pp. 1576–1590.

[173] M. Ramezani and N. Geroliminis. "Queue Profile Estimation in Congested Urban Networks with Probe Data". In: *Computer-Aided Civil and Infrastructure Engineering* (Sept. 2014), pp. 1–19.

[174] P. I. Richards. "Shock Waves on the Highway". In: *Operations Research* 4.1 (Feb. 1956), pp. 42–51.

[175] D. I. Robertson. "Research on the TRANSYT and SCOOT Methods of Signal Coordination". In: *ITE Journal* 56.1 (1986), pp. 36–40.

[176] D. I. Robertson. *TRANSYT: A Traffic Network Study Tool*. Tech. rep. RRL Report LR 253. Crowthorne, Berkshire: Ministry of Transport Road Research Laboratory, 1969.

[177] J. Rørbech. "Determining the length of the approach lanes required at signal-controlled intersections on through highways". In: *Transportation Research* 2.3 (1968), pp. 283–291.

[178] E Rowe. "The Los-Angeles Automated Traffic Surveillance and Control (ATSAC) System". In: *IEEE Transactions on Vehicular Technology* 40.1 (1991), pp. 16–20.

[179] R. O. Sanchez. "Wireless Magnetic Sensor Applications in Transportation Infrastructure". PhD thesis. Berkeley, CA, USA: University of California, Berkeley, Dec. 2012.

[180] R. O. Sanchez, R. Horowitz, and P. Varaiya. "Analysis of Queue Estimation Methods Using Wireless Magnetic Sensors". In: *Transportation Research Record: Journal of the Transportation Research Board* 2229 (Sept. 2011), pp. 34–45.

[181] R. O. Sanchez et al. "Vehicle re-identification using wireless magnetic sensors: Algorithm revision, modifications and performance analysis". In: *2011 IEEE International Conference on Vehicular Electronics and Safety*. 2011, pp. 226–231.

[182] D. L. Schrank and T. J. Lomax. *TTI's 2012 Urban Mobility Report*. Tech. rep. Texas A&M Transportation Institute, Dec. 2012.

[183] S Sen and K. L. Head. "Controlled optimization of phases at an intersection". In: *Transportation Science* 31.1 (1997), pp. 5–17.

[184] A. Sharma. "Determination of traffic responsive plan selection factors and thresholds using artificial neural networks". MA thesis. College Station, TX: Texas A&M University, 2004.

[185] A. Sharma, D. M. Bullock, and J. A. Bonneson. "Input-Output and Hybrid Techniques for Real-Time Prediction of Delay and Maximum Queue Length at Signalized Intersections". In: *Transportation Research Record: Journal of the Transportation Research Board* 2035 (Jan. 2008), pp. 69–80.

[186] S. G. Shelby et al. "An Overview and Performance Evaluation of ACS Lite – A Low Cost Adaptive Signal Control System". In: *Proceedings of the 87th Annual Meeting of the Transportation Research Board*. Washington, DC USA, Jan. 2008.

[187] *Signal Timing Process Final Report*. Tech. rep. DTFH61-01-C-00183. Federal Highway Administration, Dec. 2003.

[188] A. G. Sims and K. W. Dobinson. "The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits". In: *IEEE Transactions on Vehicular Technology* 29.2 (1980), pp. 130–137.

[189] M. G. Singh and H Tamura. "Modelling and hierarchical optimization for oversaturated urban road traffic networks". In: *International Journal of Control* 20.6 (Dec. 1974), pp. 913–934.

[190] L Sirovich and M Kirby. "Low-dimensional procedure for the characterization of human faces". In: *Journal of the Optical Society of America. A, Optics and image science* 4.3 (1987), pp. 519–524.

[191] A. Skabardonis and N. Geroliminis. "Real-time estimation of travel times on signalized arterials". In: *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*. 2005.

[192] A. Skabardonis and N. Geroliminis. "Real-Time Monitoring and Control on Signalized Arterials". In: *Journal of Intelligent Transportation Systems* 12.2 (Apr. 2008), pp. 64–74.

[193] K. R. Smilowitz and C. F. Daganzo. *Predictability of Time-dependent Traffic Backups and Other Reproducible Traits in Experimental Highway Data*. Tech. rep. UCB-ITS-PWP-99-5. California PATH, Mar. 1999.

[194] K. R. Smilowitz, C. F. Daganzo, and M. J. Cassidy. "Some observations of highway traffic in long queues". In: *Transportation Research Record: Journal of the Transportation Research Board* 1678.1 (1999), pp. 225–233.

[195] B. L. Smith, W. T. Scherer, and T. A. Hauser. "Data-mining tools for the support of signal-timing plan development". In: *Transportation Research Record: Journal of the Transportation Research Board* 1768.1 (2001), pp. 141–147.

[196] B. L. Smith et al. "Data–Driven Methodology for Signal Timing Plan Development: A Computational Approach". In: *Computer-Aided Civil and Infrastructure Engineering* 17.6 (Nov. 2002), pp. 387–395.

[197] K. Sohn and D. Kim. "Dynamic Origin-Destination Flow Estimation Using Cellular Communication System". In: *IEEE Transactions on Vehicular Technology* 57.5 (Sept. 2008), pp. 2703–2713.

[198] G. Stephanopoulos and P. G. Michalopoulos. "Modelling and analysis of traffic queue dynamics at signalized intersections". In: *Transportation Research Part A: Policy and Practice* 13.5 (1979), pp. 295–307.

[199] A. Stevanovic. *Adaptive Traffic Control Systems: Domestic and Foreign State of Practice*. Tech. rep. NCHRP Synthesis 403. Transportation Research Board of the National Academies, 2010.

[200] A Stolyar. "MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic". In: *Annals of Applied Probability* 14.1 (Feb. 2004), pp. 1–53.

[201] C. Tampere, S. Hoogendoorn, and B. van Arem. "A Behavioural Approach to Instability, Stop and Go Waves, Wide Jams and Capacity Drop". In: *Transportation and Traffic Theory* 16 (2005), pp. 205–228.

[202] L Tassiulas. "Adaptive Back-Pressure Congestion Control-Based on Local Information". In: *IEEE Transactions on Automatic Control* 40.2 (Feb. 1995), pp. 236–250.

[203] L Tassiulas and A. Ephremides. "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks". In: *IEEE Transactions on Automatic Control* 37.12 (Dec. 1992), pp. 1936–1948.

[204] *TSS-Transport Simulation Systems: AIMSUN*. URL: `http://www.aimsun.com/wp/` (visited on 05/11/2015).

[205] P. Varaiya. "Max pressure control of a network of signalized intersections". In: *Transportation Research Part C: Emerging Technologies* 36 (Nov. 2013), pp. 1–19.

[206] G. Vigos, M. Papageorgiou, and Y. Wang. "Real-time estimation of vehicle-count within signalized links". In: *Transportation Research Part C: Emerging Technologies* 16.1 (Feb. 2008), pp. 18–35.

[207] F. Viti and H. Van Zuylen. "The Dynamics and the Uncertainty of Queues at Fixed and Actuated Controls: A Probabilistic Approach". In: *Journal of Intelligent Transportation Systems* 13.1 (Jan. 2009), pp. 39–51.

[208] C. E. Wallace et al. *TRANSYT-7F Users Manual*. Federal Highway Administration, United States Department of Transportation. 1998.

[209] P. Wang et al. "Understanding Road Usage Patterns in Urban Areas". In: *Scientific Reports* 2.1001 (Dec. 2012), pp. 1–6.

[210] X. Wang, W Cottrell, and S. Mu. "Using k-means clustering to identify time-of-day break points for traffic signal timing plans". In: *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*. Vienna, Austria, Sept. 2005, pp. 586–591.

[211] J. G. Wardrop. "Some Theoretical Aspects of Road Traffic Research". In: *Proceedings of the Institute of Civil Engineers*. Jan. 1952, pp. 325–362.

[212] F. V. Webster. *Traffic signal settings*. Tech. rep. Road Research Technical Paper No. 39. London, UK: Department of Scientific and Industrial Research Road Research Laboratory, 1958.

[213] C Wenk, R Salas, and D Pfoser. "Addressing the Need for Map-Matching Speed: Localizing Globalb Curve-Matching Algorithms". In: *18th International Conference on Scientific and Statistical Database Management*. IEEE, 2006, pp. 379–388.

[214] W.-M. Wey. "Model formulation and solution algorithm of traffic signal control in an urban network". In: *Computers, Environment and Urban Systems* 24 (2000), pp. 355–377.

[215] W.-M. Wey and R Jayakrishnan. "Network traffic signal optimization formulation with embedded platoon dispersion simulation". In: *Transportation Research Record: Journal of the Transportation Research Board* 1683 (1999), pp. 150–159.

[216] *What Have We Learned About Intelligent Transportation Systems?* Tech. rep. Federal Highway Administration, Nov. 2000.

[217] A Wilson. "TRENT: A Traffic-Responsive Control Method for Small Networks". In: *Proceedings of the 5th World Congress on Intelligent Transportation Sysyems*. Seoul, Korea, Oct. 1998.

[218] T Wongpiromsarn et al. "Distributed traffic signal control for maximum network throughput". In: *Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems*. Anchorage, Alaska USA, 2012, pp. 588–595.

[219] D. B. Work et al. "A Traffic Model for Velocity Data Assimilation". In: *Applied Mathematics Research eXpress* 2010.1 (Apr. 2010).

[220] J. Wu, X. Jin, and A. J. Horowitz. "Methodologies for Estimating Vehicle Queue Length at Metered On-Ramps". In: *Transportation Research Record: Journal of the Transportation Research Board* 2047 (Sept. 2008), pp. 75–82.

[221] J. Wu et al. "Experiment to Improve Estimation of Vehicle Queue Length at Metered On-Ramps". In: *Transportation Research Record: Journal of the Transportation Research Board* 2099 (Sept. 2009), pp. 30–38.

[222] Y. Yim and R. Cayford. *Investigation of Vehicles as Probes Using Global Positioning System and Cellular Phone Tracking: Field Operational Test*. Tech. rep. California PATH Working Paper UCB-ITS-PWP-2001-9. California PATH, Institute of Transportation Studies at the University of California, Berkeley, Feb. 2001.

[223] H. M. Zhang, Y. M. Nie, and Z. S. Qian. "Modelling network flow with and without link interactions: the cases of point queue, spatial queue and cell transmission model". In: *Transportmetrica B: Transport Dynamics* 1.1 (2013), pp. 33–51.

Appendix A

Descriptions of existing adaptive traffic control systems

- The TRANSYT model is the basis for an automated online optimization method called the **Signal, Cycle, and Offset Optimization Technique (SCOOT)** [100, 175]. SCOOT adjusts traffic at small intervals (not a "one-shot" optimal) based on the optimal settings derived from TRANSYT given inputs derived from the most recent knowledge of traffic conditions. While it requires a central processing unit, the resulting strategy is functionally a decentralized actuation of individual signal splits, offsets, and cycle times. SCOOT was implemented in the 1980's in Britain with demonstrated success compared to typical heuristic fixed-time plans. It is more recently commonly deployed in Australia and Asia, and occasionally in North America (notably, see the results of a field test in Anaheim, CA [146]).

- The **Sydney Coordinated Adaptive Traffic System (SCATS)** is a closed-loop traffic control system that updates intersection cycle length and splits based on information gathered in real-time at stop-line detectors [188, 131]. It also showed substantial improvement in arterial flow over a static optimized fixed-time control scheme. As of 1990, there were plans to implement SCATS on over 1,800 signals in Sydney [130]. It requires a central processing center to supervise a set of independent regional computers which each directly advise up to 200 local controllers (sets of signals). The regional computers essentially choose which controllers to "marry" to create groups of locally connected sub-systems over which to optimize flows. The selection of these "marriages" is dependent on measured traffic conditions. The regional system then selects the cycle length, offset, and plan (out of a pre-defined set) for each controller in each sub-system that optimizes either for minimum delay, minimum stops, or maximum throughput in that subsystem. SCATS was the first adaptive control system deployed in the U.S. with a field test as part of the Fast-Trac program in Oakland County, Michigan [61]. This test used video detectors instead of loop detectors.

- The FHWA developed a centralized traffic control system known as the **Urban Traffic**

Control System (UTCS) in the 1970's. This uses historical data to develop timing plans which may vary by time of day or day of week. The first generation system (1-GC) selects fixed or variable timing plans based on time of day or by matching recent measurement to previously observed congestion patterns. If operating in traffic responsive mode, it can update the plan at 15 minute intervals based on real-time demand measurements. A feature called critical intersection control permits some adjustment of green splits from pre-determined plans. Details of the traffic responsive capabilities of UTCS are provided in Appendix C.

- **UTCS 2-GC** is a online network-wide control strategy that uses real-time measurements and a prediction model (fed historical data) to optimize signal timings at 5 minute intervals. Plans can be switched at most once every ten minutes, with a molded transition time. **UTCS 3-G** shortens the revision of the optimal to once every 3 minutes, with possible control updates every 5 minutes. It also incorporates current measurements into modeled predictions, and includes the option to vary cycle lengths among signals.

- The **Optimization Policies for Adaptive Control (OPAC)** is a real-time, decentralized, demand-responsive signal timing optimization algorithm which acts to minimize a function of total intersection delay and stops on a two-level basis [70]. It uses a combination of measured and modeled demands to select optimal phase durations within constraints on minimum and maximum green times at a time resolution of 4-5 seconds. There can also be (demand-responsive) considerations of signal coordination via virtual constraints on cycle length and offsets. However, it is important to note that there is not a rigidly fixed cycle time in the OPAC algorithm; in fact, it is considered the first practical system to break from the traditional concept of cycle-based signal timing plans. It requires on-line data from upstream link detectors at all controllable links, but can otherwise be implemented on existing hardware (2070 controllers). Multiple approaches to the optimization problem have been tested, including a dynamic programming approach, a sequential optimization approach, a rolling horizon approach, and, most recently, a "Virtual-Fixed-Cycle" approach developed to fulfill the requirements of the RT-TRACS program. [71]. The dynamic programming approach requires advanced knowledge of arrival data and neighboring signal data, which is typically not available and has to be estimated. The optimal sequential constrained search divides the optimization process into stages of 50-100 seconds, and during each stage there is at least one signal change and at most three signal changes. Total delay is evaluated for each feasible switching sequence, and the minimal pattern is selected.

- **RHODES** is a real-time, hierarchical, distributed traffic signal control framework that uses high-resolution predictive models to react to stochastic variations in demand patterns at intersections. It is based on an architecture that separates low-level intersection control, mid-level network flow control (signal coordination), and high-level network loading dynamics [86]. Low-level logical signal control decisions such as signal

phase and duration decisions are made on a second-by-second basis based on observed vehicular flows and predictions as well as operational constraints. These are distributed controllers. Coordination constraints are made at a 200-300 second interval based on predictions of platoon flows. This requires communications with other local controllers and sensors. General travel demand over longer periods of time (about one hour) are used centrally at the highest level to anticipate general demand patterns and future platoon sizes at network boundaries. RHODES uses a stochastic model called PREDICT to estimate intersection arrival patterns based on upstream measurements [145]. The individual intersection control logic (called Controlled Optimization of Phases, or COP) is found by solving a dynamic programming problem that minimizes a function of stops or delay over a rolling horizon [183]. While it does not require a pre-defined phase sequence (cycle), this can be enforced when preferred by operators. A platoon-tracking model called APRES-NET is then used at the sub-network level, and signal offsets are optimized based on an adaptive bandwidth controller called REALBAND [56]. The offsets generated from REALBAND appear as constraints to the COP controller. A prototype of the RHODES architecture is proposed for field testing in Tuscon, AZ, Seattle, WA, and Tempe, AZ [145].

- **Automated Traffic Surveillance and Control (ATSAC)** is a system created by the city of Los Angeles. It is a customization of UTCS that incorporates measurements from loop detectors as well as CCTV, and features signal optimization software and real-time centralized control of traffic signals. Time-of-day signal timing plans are generated using historical data and TRANSYT models, then they are fine-tuned manually by operators based on conditions observed at link detectors and CCTV cameras. There is also capability to use automated traffic-responsive control to select the appropriate signal timing plan from a pre-defined set using an algorithm that matches observed congestion patterns to those that were used to generate each plan [178]. Link detectors are placed either 250 feet upstream or 100 feet downstream of each signal. At intersections of urban and local streets, semi-actuated control is used and data is not centralized. ATSAC was first used in connection with congestion generated around the Coliseum due to the 1984 Olympic Games on a network of 118 intersections and 396 detectors, and is currently deployed on approximately 1170 intersections in Los Angeles [216]. It has been found to be highly effective at clearing event traffic. It also reduces stops by 35%, intersection delay by 20%, travel time by 13%, and fuel consumption by 12.5%. While the cost of installation and operation is more than $70,000 per intersection, it is estimated to have paid for itself in less than one year upon initial deployment.

- **PRODYN** uses forward dynamic programming to minimize intersection delay and decomposition coordination techniques to optimize signal coordination [90]. Like OPAC and RHODES, it does not follow a phase-cycle pattern but rather chooses flexible phase durations to attain closer-to-optimal performance. The local intersections de-

cides at a 5-second interval whether to shift from one phase to the next based on a dynamic programming problem operating on a heuristic delay objective function. It requires a sensor at each upstream junction output (to predict downstream arrivals) and another sensor at about 50 meters from the stop line. The combination of these measurements are used to estimate the probability density function of the queue state of the line based on a vertical queuing model. This estimate is used as an initial state in a forecast model running over a 75-second time horizon. The forecast model also has information sent by neighboring intersections on their intended signal timings during this time. The optimization procedure is used to determine the future green timings which minimize the sum of delays over the 75-second horizon. PRODYN was field-tested on two intersections in France [63].

- **Urban Traffic Optimization by Integrated Automation (UTOPIA)** is another hierarchal traffic control system that performs bilevel optimization [135]. Local intersection signal optimization is done using a rolling horizon optimal controller with 120-second time horizon with 3-second actuation updates using an algorithm called SPOT. SPOT applies a microscopic model to estimate the state and time-varying parameters of an intersection from loop detector measurements. The area level uses a less detailed model to validate local detection and compare measurements to historical (or nominal) congestion levels to detect significant congestion or possible incidents. It can also perform regional actions such as activate VMS. A central "supervisor" level collects area information and integrates outside information such as bus arrivals and travel times into a less detailed macroscopic model. UTOPIA has been deployed in many places in Europe, notably in Denmark.

- **Control of Networks by Optimization of Switchovers (CRONOS)** is an algorithm developed in the 1990's with the objective of building a non-exponential and fast optimization method to provide signal states for the next second in less than one second — it was motived by the desire to react as quickly as possible to variations in traffic conditions [30, 29]. It also desired to use image-processing-based traffic measurements and vehicle spatial occupancy inside the intersection as state feedback. It uses a state prediction based on a rolling average of past arrivals to calculate the value of a chosen traffic parameter (usually total delay) over a finite future time horizon. The optimization procedure is a modified version of the Box algorithm, which has polynomial complexity as the number of intersections increases. Typically the optimization procedure is run with a time horizon of around one minute for one controlled intersection. When the procedure is run on a set of intersections in a single zone, intersection arrivals are calculated from the decision of the upstream intersection at the previous time step, preventing the need for more heuristic predictions and allowing the optimal signal pattern for the entire zone to be calculated at once. Like OPAC/UTOPIA/PRODYN, CRONOS does not stick to the concept of a fixed cycle. In fact, it goes even farther to eliminate the need for pre-defined stages at all; intersections are instead only defined

by a set of safety constraints on simultaneous actuations [29].

- **ACS Lite** is the result of a recent project by the US FHWA that was initiated in response to widespread concerns over the installation and maintenance costs of previously developed ATCS [186]. To reduce costs, it leverages existing capabilities on typical National Electrical Manufacturers of America (NEMA) model controllers and existing detectors that communicate via the standard National Transportation Communications for ITS Protocol (NTCIP). In fact, ACS Lite depends on existing hardware and logic for underlying signal timing plan characteristics and immediate decision-making. It instead functions "on top" of existing second-to-second operations by making incremental adjustments to the baseline split and offset parameters on a 5-10 minute time step. The suggested timing adjustments are based on measurements of phase utilization from local stopline detectors, and can be constrained to remain within predetermined limits on minimum or maximum green times. Early field tests have shown delay reductions of up to 35% [7].

Appendix B

Stability of max pressure controller, original formulation

For reference, we provide the proof of the following theorem that was originally published in [205]. All terminology and variables are defined as in Chapter 5.

Theorem: *The max pressure controller*

$$u^*(X(t)) = \arg\max\{\gamma(S)(X(t))|S \in U\}$$

is stabilizing whenever the average demand vector $d = \{d_l\}$ is within the set of feasible demands D^0.

Proof. To prove that the max pressure controller is stabilizing, we must prove that when the controller is applied to the system dynamics, the following quantity

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big\{|X(t)|_1\Big\}$$

remains bounded. In fact, it is sufficient to show that there exists constants K and $\varepsilon > 0$ such that

$$\mathbb{E}\Big\{|X(t+1)|_2^2 - |X(t)|_2^2\big|X(t)\Big\} < -\varepsilon|X(t)|_1 + K \tag{B.1}$$

where $|X|_2^2 = \sum_{l,m}|x(l,m)|^2$. This is because (B.1) immediately implies that

$$\mathbb{E}\Big\{|X(T+1)|_2^2\Big\} - \mathbb{E}\Big\{|X(1)|_2^2\Big\} < -\varepsilon\sum_{t=1}^{T}\mathbb{E}\Big\{|X(t)|_1\Big\} + KT \tag{B.2}$$

which can be rewritten as a bound on the desired quantity:

$$\frac{\varepsilon}{T}\sum_{t=1}^{T}\mathbb{E}\Big\{|X(t)|_1\Big\} < K + \frac{1}{T}\mathbb{E}\Big\{|X(1)|_2^2\Big\} \tag{B.3}$$

The following proof shows that for a network with feasible demand and a standard max pressure control, there exists a $\varepsilon > 0$ and a $K > 0$ satisfying (B.1).

Consider the expectation of the following function of queue state with perturbation

$$\delta(t) = X(t+1) - X(t) \tag{B.4}$$

conditioned on the past queue state:

$$\begin{aligned} |X(t+1)|^2 - |X(t)|^2 &= |X(t) + \delta(t)|^2 - |X(t)|^2 \\ &= 2X(t)^T\delta(t) + |\delta(t)|^2 \\ &= 2\alpha(t) + \beta(t) \end{aligned} \tag{B.5}$$

with

$$\alpha(t) = X(t)^T\delta(t) \tag{B.6}$$

and

$$\beta(t) = |\delta(t)|^2 \tag{B.7}$$

We address bounds on β and α separately.

Bound on $\beta(t) = |\delta(t)|^2$

Define known (or measurable) parameters:

- $\overline{C}(l,m)$ is the maximum realized saturation flow on (l,m),
- $\overline{d}(l,m)$ is the maximum possible value of the demand on (l,m),
- $D(l)(t)$ is the realized demand on link l at time t, and
- $D(l,m)(t) = D(l)(t)R(l,m)(t)$.

Then if $l \in \mathcal{L}_{\text{ent}}$ and $m \in Out(l)$,

$$\begin{aligned} \big|\delta(l,m)(t)\big| &= \Big| - [C(l,m)(t+1)S(l,m)(t) \wedge x(l,m)(t)] + D(l,m)(t+1)\Big| \\ &\leq \max\big\{\overline{C}(l,m), \overline{d}(l,m)\big\} \end{aligned} \tag{B.8}$$

This is because both $C(l,m)(t+1)S(l,m)(t) \wedge x(l,m)(t)$ and $D(l,m)(t+1)$ are non-negative, so the absolute value of their difference must be less than either of the two quantities individually.

Similarly, if $l \in \mathcal{L}\backslash\mathcal{L}_{\text{ent}}$ and $m \in Out(l)$:

$$\begin{aligned} \big|\delta(l,m)(t)\big| &= \Big| - \big[C(l,m)(t+1)S(l,m)(t) \wedge x(l,m)(t)\big] \\ &\qquad + \sum_k \big[C(k,l)(t+1)S(k,l)(t) \wedge x(k,l)(t)\big]R(l,m)(t+1)\Big| \\ &\leq \max\left\{\overline{C}(l,m), \sum_k \overline{C}(k,l)\right\} \end{aligned} \tag{B.9}$$

If we define B as the maximum of all of the quantities $\left\{\overline{C}(l,m), \sum_k \overline{C}(k,l), \overline{d}(l,m)\right\}$ and N as the number of queues in the network, we can derive a bound for β which depends on only B and N:

$$\beta(t) = \left|\delta(t)\right|^2 \leq NB^2 \tag{B.10}$$

Bound on $\alpha(t) = X(t)^T\delta(t)$

The term $\alpha(t)$ in (B.6) is explicitly defined in terms of queue state $X(t)$ as follows:

$$\begin{aligned}
\alpha(t) &= X(t)^T\left[X(t+1) - X(t)\right] \\
&= \sum_{l\in\mathcal{L}\setminus\mathcal{L}_{\text{ent}},m}\sum_k \left[C(k,l)(t+1)S(k,l)(t) \wedge x(k,l)(t)\right]R(l,m)(t+1)x(l,m)(t) \\
&\quad - \left(\sum_{l\in\mathcal{L},m}\left[C(l,m)(t+1)S(l,m)(t)\wedge x(l,m)(t)\right] + \sum_{l\in\mathcal{L}_{\text{ent}},m} d(l,m)(t+1)\right)x(l,m)(t) \\
&= \sum_{l\in\mathcal{L},m}\Big[C(l,m)(t+1)S(l,m)(t)\wedge x(l,m)(t)\Big]\Big(\sum_p R(m,p)(t+1)x(m,p)(t) - x(l,m)(t)\Big) \\
&\qquad + \sum_{l\in\mathcal{L}_{\text{ent}},m} d(l,m)(t+1)x(l,m)(t)
\end{aligned} \tag{B.11}$$

Observe that because $R(m,p)(t+1)$ is independent of $C(l,m)(t+1)$ and $X(t)$,

$$\begin{aligned}
&\mathbb{E}\Big\{\left[C(l,m)(t+1)S(l,m)(t)\wedge x(l,m)(t)\right]R(m,p)(t+1)x(m,p)(t)\big|X(t)\Big\} \\
&\qquad = \mathbb{E}\Big\{\left[C(l,m)(t+1)S(l,m)(t)\wedge x(l,m)(t)\right]\big|X(t)\Big\}r(m,p)x(m,p)(t)
\end{aligned} \tag{B.12}$$

Also, note that the the expectation of demand $d(l,m)$ is equal to the measured demand on link l scaled by the expected split ratio $r(l,m)$:

$$\mathbb{E}\Big\{d(l,m)\Big\} = d_l r(l,m) \tag{B.13}$$

Hence we derive the expectation of (B.11) as follows:

$$\begin{aligned}
&\mathbb{E}\Big\{\alpha(t)|X(t)\Big\} = \\
&\quad\sum_{l\in\mathcal{L},m}\mathbb{E}\Big\{\Big[C(l,m)(t+1)S(l,m)(t)\wedge x(l,m)(t)\Big]\Big|X(t)\Big\}\Big(\sum_p r(m,p)x(m,p)(t) - x(l,m)(t)\Big) \\
&\qquad\qquad + \sum_{l\in\mathcal{L}_{\text{ent}},m} d_l r(l,m)x(l,m)(t)
\end{aligned} \tag{B.14}$$

$$\begin{aligned}
&= -\sum_{l\in\mathcal{L},m}\mathbb{E}\Big\{\left[C(l,m)(t+1)S(l,m)(t)\wedge x(l,m)(t)\right]\big|X(t)\Big\}w(l,m)(t) \\
&\qquad\qquad + \sum_{l\in\mathcal{L}_{\text{ent}},m} d_l r(l,m)x(l,m)(t)
\end{aligned} \tag{B.15}$$

where $w(l,m)(t) = w(l,m)(X(t))$ is the max pressure *weight* of a link, as defined in equation (5.9). We then incorporate the following relation:

$$\begin{aligned}
\sum_{l\in\mathcal{L},m} f_l r(l,m) w(l,m)(t) &= \sum_{l\in\mathcal{L},m} f_l r(l,m) \left[x(l,m) - \sum_p r(m,p) x(m,p)(t) \right] \\
&= \sum_{l\in\mathcal{L},m} f_l r(l,m) x(l,m)(t) - \sum_m \left[\sum_{l\in\mathcal{L}} f_l r(l,m) \sum_p r(m,p) x(m,p)(t) \right] \\
&= \sum_{l\in\mathcal{L},m} f_l r(l,m) x(l,m)(t) - \sum_{m\in\mathcal{L}\setminus\mathcal{L}_{\text{ent}},p} f_m r(m,p) x(m,p)(t) \\
&= \sum_{l\in\mathcal{L}_{\text{ent}},m} d_l r(l,m) x(l,m)(t)
\end{aligned}$$

So (B.15) is further simplified to:

$$\mathbb{E}\{\alpha(t)|X(t)\} = \sum_{l\in\mathcal{L},m} \Big[f_l r(l,m) - \mathbb{E}\Big\{ \big[C(l,m)(t+1)S(l,m)(t) \wedge x(l,m)(t)\big] \big| X(t) \Big\} \Big] w(l,m)(t) \tag{B.16}$$

By adding the 0-valued term $[c(l,m)S(l,m)(t)w(l,m)(t) - c(l,m)S(l,m)(t)w(l,m)(t)]$ to (B.16), we split this expression into the following sub-terms for convenience:

$$\mathbb{E}\{\alpha(t)|X(t)\} = \alpha_1(t) + \alpha_2(t) \tag{B.17}$$

where

$$\alpha_1(t) = \sum_{l\in\mathcal{L},m} \left[f_l r(l,m) - c(l,m)S(l,m)(t) \right] w(l,m)(t) \tag{B.18}$$

and

$$\alpha_2(t) = \sum_{l\in\mathcal{L},m} \Big[c(l,m) - \mathbb{E}\Big\{ \big[C(l,m)(t+1) \wedge x(l,m)(t)\big] \big| X(t) \Big\} \Big] S(l,m)(t) w(l,m)(t) \tag{B.19}$$

Note that because $S(l,m)(t) \in \{0,1\}$, this term is brought outside of the internal minimization function in the $\alpha_2(t)$ expression without impact on the result.

Lemma B.1. *For all l, m, t,*

$$\alpha_2(t) \leq \sum_{l\in\mathcal{L},m} c(l,m)\overline{C}(l,m) \tag{B.20}$$

where $\overline{C}(l,m)$ is the maximum value of the random service rate $C(l,m)(t)$.

Proof of Lemma B.1:
By Jensen's inequality,

$$\begin{aligned}\mathbb{E}\Big\{C(l,m)(t+1)\wedge x(l,m)(t)\big|X(t)\Big\} &\le \mathbb{E}\Big\{C(l,m)(t+1)\big|X(t)\Big\}\wedge x(l,m)(t)\\ &= c(l,m)\wedge x(l,m)(t)\\ &\le c(l,m)\end{aligned}$$

Furthermore, it is known that

$$\Big[c(l,m)-\mathbb{E}\Big\{\big[C(l,m)(t+1)\wedge x(l,m)(t)\big]\big|X(t)\Big\}\Big]\ge 0 \tag{B.21}$$

and $\Big[c(l,m)-\mathbb{E}\Big\{\big[C(l,m)(t+1)\wedge x(l,m)(t)\big]\big|X(t)\Big\}\Big]=0$ only when $x(l,m)(t)>\overline{C}(l,m)$. Using these relations and the observations that $w(l,m)(t)\le x(l,m)(t)$ and $S(l,m)(t)\in\{0,1\}$, the following must hold

$$\begin{aligned}\alpha_2(t) &= \sum_{l\in\mathcal{L},m}\Big[c(l,m)-\mathbb{E}\Big\{\big[C(l,m)(t+1)\wedge x(l,m)(t)\big]\big|X(t)\Big\}\Big]S(l,m)(t)w(l,m)(t)\\ &\le \sum_{l\in\mathcal{L},m}\Big[c(l,m)-\mathbb{E}\Big\{\big[C(l,m)(t+1)\wedge x(l,m)(t)\big]\big|X(t)\Big\}\Big]S(l,m)(t)x(l,m)(t)\\ &\le \sum_{l\in\mathcal{L},m}c(l,m)\overline{C}(l,m)\end{aligned}$$

Lemma B.2. *If $S^*(t)=u^*(X(t))=\arg\max\{\gamma(S)(X(t))|S\in U\}$ and demand d is in the set of feasible demands D^o, then there exists an $\varepsilon>0$, $\eta>0$ such that*

$$\alpha_1(t)\le-\varepsilon\eta\big|X(t)\big| \tag{B.22}$$

Proof of Lemma B.2:
Applying the definition of max pressure control in (5.11) as S^* and using long term proportion matrix M as defined in (5.7),

$$\begin{aligned}\sum_{l,m}S^*(l,m)(t)c(l,m)w(l,m)(t) &= \max_{S\in U}\sum_{l,m}S(l,m)c(l,m)w(l,m)(t)\\ &= \max_{M\in co(U)}\sum_{l,m}M(l,m)c(l,m)w(l,m)(t)\end{aligned} \tag{B.23}$$

As in (5.8), since $d\in D^o$ there exists an ε and M^+ such that $C(l,m)M^+(l,m)>f_l r(l,m)+\varepsilon\ \forall(l,m)$. Logically, any M' such that $0\le M'\le M^+$ (component-wise) must also be in $co(U)$. Therefore choose a $M'<M^+$ such that

$$M'(l,m)c(l,m)=\begin{cases}f_l r(l,m)+\varepsilon & \text{if } w(l,m)>0\\ 0 & \text{if } w(l,m)\le 0\end{cases}$$

Then

$$\begin{aligned}\alpha_1(t) &= \sum_{l\in\mathcal{L},m} \left[f_l r(l,m) - c(l,m)S^*(l,m)(t)\right] w(l,m)(t) \\ &\leq \sum_{l\in\mathcal{L},m} \left[f_l r(l,m) - M'(l,m)c(l,m)\right] w(l,m)(t) \\ &= -\varepsilon \sum_{l\in\mathcal{L},m} \max\{w(l,m)(t),0\} + \sum_{l\in\mathcal{L},m} f_l r(l,m) \min\{w(l,m)(t),0\} \\ &\leq -\varepsilon \sum_{l\in\mathcal{L},m} |w(l,m)(X(t))| \end{aligned} \tag{B.24}$$

Notice that $w(l,m)(X(t)) = x(l,m)(t) - \sum_p r(m,p)x(m,p)(t)$ is a linear, invertible function of the array $X(t)$, and therefore there exists a $\eta > 0$ such that $\sum_{l,m} |w(l,m)(X(t))| \geq \eta |X(t)|$. Substituting this expression into (B.24) defines a bound on α_1 (t):

$$\alpha_1(t) \leq -\varepsilon\eta\, |X(t)| \tag{B.25}$$

Combining the results of Lemmas B.1 and B.2 generates the desired bound on $\mathbb{E}\Big\{\alpha(t)|X(t)\Big\}$:

$$\mathbb{E}\Big\{\alpha(t)|X(t)\Big\} \leq -\varepsilon\eta\, |X(t)| + \sum_{l\in\mathcal{L},m} c(l,m)\overline{C}(l,m) \tag{B.26}$$

Explicit bound on queues

Combining (B.26) and (B.10), we obtain

$$\begin{aligned}\mathbb{E}\Big\{|X(t+1)|^2 - |X(t)|^2 | X(t)\Big\} &= \mathbb{E}\Big\{2\alpha(t) + \beta(t)\Big\} \\ &< -2\varepsilon\eta\, |X(t)| + 2\sum_{l\in\mathcal{L},m} \left[c(l,m)\overline{C}(l,m)\right] + NB^2 \end{aligned} \tag{B.27}$$

where N is the number of links in the network and $B = \max\left\{\overline{C}(l,m), \sum_k \overline{C}(k,l), \overline{d}(l,m)\right\}$. The expression (B.27) demonstrates stability.

□

Appendix C

"VPKO" traffic responsive functionalities

The following paragraphs describe the standard for traffic responsive plan selection developed for Type-170 controllers as part of FHWA's Urban Traffic Control System (UTCS) standard. More on objectives of the UTCS project is described in Section 2.5.

UTCS pattern matching algorithm

Assume that there are between 4 and 16 detectors per intersection. In a typical 4-approach intersection, these detectors are loop detectors that are placed both at the stopbar of each approach and approximately 200 feet upstream of these stopbars. Each detector l returns a measurement of raw volume $v_l^r[k]$ and raw occupancy $o_l^r[k]$ over a fixed period of time denoted by time step k.

Define the following *smoothed volume/occupancy* values from a loop detector l:

$$v_l^s[k] = \tau v_l^s[k-1] + (1-\tau)v^r[k] \tag{C.1}$$

$$o_l^s[k] = \tau o_l^s[k-1] + (1-\tau)o^r[k] \tag{C.2}$$

where τ is a filter parameter related to the user-defined time constant T_c (typically set to be equal to the time to reduce filter error by 63%) and the measurement time step Δt:

$$\tau = e^{-\Delta t/T_c} \tag{C.3}$$

Flow and occupancy measurements from each of intersection i's system loop detectors $l \in \{1, \ldots, L_i\}$ are used to calculated a "volume plus K occupancy" or VPKO value :

$$VPKO_{i,l}^{obs} = v_{i,l}^s l + K \times o_{i,l}^s \tag{C.4}$$

for a user-defined (system-wide) occupancy weighting factor $K \in (0, 100)$. The measured state of the intersection $\hat{x}_i(t)$ is represented by an array of VPKO values for all system detectors at that intersection:

$$\hat{x}_i[k] = \left[VPKO_{i,1}^{obs}[k],\ VPKO_{i,2}^{obs}[k],\ \ldots,\ VPKO_{i,L_i}^{obs}[k]\right]^T \tag{C.5}$$

UTCS-1G controllers each store a fixed set of signal timing plans P. Each plan $p \in P$ is associated with pre-defined volume and occupancy signatures for each system detector (respectively, $\overline{v}_l^p$ and $\overline{o}_l^p$ for all l). To choose an appropriate plan to implement, controllers calculate a linear combination of the differences between each the measured VPKO and that corresponding to each available plan:

$$F_p[k] = \sum_l W_{i,l} \left| (v_{i,l}^s[k] + K \times o_{i,l}^s[k]) - (\overline{v}_l^p + K \times \overline{o}_l^p) \right| \tag{C.6}$$

given user-defined detector weights $W_l \in (0, 10)$ [69]. A controller selects the candidate plan with minimum value of its comparison function:

$$p_i^*[k] = u_i(\hat{x}_i[k]) = \arg\min_{\phi \in P_i} F_{i,\phi}[k] \tag{C.7}$$

But a transition is only initiated when the following two conditions are met:

1. the current plan must have been in place for at least time $t_s^{\min}$ time steps:

$$t_s^-[k] \geq t_s^{\min} \tag{C.8}$$

   where $t_s^-[k]$ is the number of time steps since the previous plan switch, and

2. the value of the comparison function $F_{p[k]}$ for the current plan $p[k]$ differs from that of the candidate plan p_i^* by more than a pre-defined threshold:

$$E = AF_{p_i[k]} - F_{p_i^*} > 0 \tag{C.9}$$

   where A is a weighting factor defined separately for each section.

Control switching decisions are typically limited to a frequency on the order of once per 600-900 seconds (10-15 minutes).

Section synchronization

Intersection controllers can act somewhat independently, but are synchronized within a coordination zone by a pre-designated *master* controller. Every time that the previously described pattern-matching algorithm selects a new plan in a *slave* controller, the software compares the cycle lengths of the candidate plan with that currently operating at the master controller. If the difference between these cycle lengths is less than an operator-defined limit, the timing plan number for the slave intersection controller is simply set to the same number as that of the master controller [69]. In other words, the feedback-dictated plans at the slave controllers may often be overridden if they differ significantly from the desires of the master.

Critical intersection control

UTCS software also has the capability of implementing a feature called *Critical Intersection Control* (CIC) which enables a controller to explicitly calculate the total green demand for all phases in a *critical* intersection using the previous time step's approach volume and occupancy measurements. Green demands are calculated in one of two ways:

1. The standard UTCS formula from the Traffic Control Systems Handbook defines green demand as:
$$\overline{g}_l = K_1 o_l^s + K_2 v_l^s + K_3 (v_l^s \cdot o_l^s) \tag{C.10}$$

2. In the Los Angeles DOT's ATSAC system (a customization of UTCS), a user defines coefficients A, B, C, and D to obtain an approach greed demand $\overline{g}_l$:
$$\overline{g}_l = A(v_l^s)^B + C(o_l^s)^D \tag{C.11}$$

The critical controller can then redistribute the stage green splits $\{g_s\}$ to reduce detected excess greens for each stage (e_s) given minimum green time parameters $g_s^{\min}$:

$$e_s = \max\left\{ \max_l\{\overline{g}_l\} - g_s^{\min}, 0 \right\} \tag{C.12}$$

$$g_s' = \frac{e_s}{\sum_\sigma e_\sigma} \times G^{\text{free}} + g_s^{\min} \tag{C.13}$$

$$\text{where } G^{\text{free}} = C - \sum_s (g_s^{\min} + y_s + r_s) \tag{C.14}$$

Note that while green splits may be adjusted independently at these critical intersections, cycle times and yield points (relative offsets) are held constant to maintain network-level coordination.