New Approaches for Online Control of Urban Traffic Signal Systems

Von der
Fakultät Architektur, Bauingenieurwesen und Umweltwissenschaften
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig

zur Erlangung des Grades eines
Doktoringenieurs (Dr.-Ing.)
genehmigte

Dissertation

von
Tobias Pohlmann
geboren am 07.06.1977
aus Mannheim

Eingereicht am	11. Mai 2010
Disputation am	06. September 2010
Berichterstatter	Prof. Dr.-Ing. Bernhard Friedrich
	Prof. Dr.-Ing. Manfred Boltze

2010

Danksagungen

Ich bedanke mich herzlich bei Prof. Dr.-Ing. Bernhard Friedrich für die Betreuung dieser Doktorarbeit. Während meiner Zeit am Institut für Verkehrswirtschaft, Straßenwesen und Städtebau der Leibniz Universität Hannover und am Institut für Verkehr und Stadtbauwesen der Technischen Universität Carolo-Wilhelmina zu Braunschweig hatte ich die Möglichkeit, in zahlreichen fachlichen Gesprächen von seinem Wissen zu profitieren. Zudem hat er von Anfang großes Vertrauen in mich gesetzt und mir bei der Bearbeitung diverser Forschungsprojekte viele Freiräume gelassen. Des Weiteren bedanke ich mich bei Prof. Dr.-Ing. Manfred Boltze für die Übernahme des Koreferats. Ihm verdanke ich außerdem meine Tätigkeit im Bereich des Verkehrswesens, mit dem ich im Rahmen meines Bauingenieurstudiums an der TU Darmstadt erstmalig in seinen Vorlesungen in Berührung kam.

Bei meinen ehemaligen Kollegen an beiden Instituten bedanke ich mich für die kollegiale und angenehme Atmosphäre und die zahlreichen Gespräche, die mir wertvolle Anregungen lieferten. Großen Dank schulde ich auch Anike Biener, die mir als studentische Hilfskraft mit viel Einsatz und Engagement bei der Durchführung zahlreicher Simulationsstudien, der Auswertung von Daten, der Anfertigung von Diagrammen und diversen weiteren Aufgaben eine große Hilfe war.

Weiterer Dank gebührt Luisa Cuppone sowie Ruth und Klaus-Peter Schäfer, die sich die Zeit nahmen, die fertige Arbeit in sprachlicher Hinsicht zu lesen und zu korrigieren. Die Verantwortung für verbleibende Germanismen und sonstige Fehler liegt allein bei mir.

Meinen Eltern danke ich, dass sie es mir durch mein Studium ermöglichten, den Weg hin zur Promotion einzuschlagen. Ein besonderer Dank geht an meine Verlobte Alexandra Kretz für die große Unterstützung insbesondere in der stressigsten Zeit kurz vor Abgabe der Arbeit sowie für das Durchlesen und Auffinden weiterer Fehler.

Der Deutschen Forschungsgemeinschaft (DFG) danke ich für die finanzielle Unterstützung des Forschungsvorhabens mit der Fördernummer FR 1670/4-1. Dieses Projekt finanzierte in den Jahren 2008 und 2009 meine Drittmittelstelle an der TU Braunschweig und führte zu der vorliegenden Doktorarbeit. Dr.-Ing. Klaus Meffert und allen weiteren Entwicklern des Java Genetic Algorithms Package danke ich für die Bereitstellung dieser kostenlosen Java-Bibliothek, die in dieser Arbeit erfolgreich zum Einsatz kam.

Abstract

Traffic signal control is one of the key factors in urban traffic control. It directly influences the quality of traffic flow. Besides fixed time and vehicle-actuated control, so called Adaptive Traffic Control Systems (ATCS) have been developed in the past. These systems control a set of connected intersections in a network and aim at optimizing traffic signal control in real-time by continuously adapting the signalization at the intersections to the currently detected or estimated traffic demand in the network. Improvements of traffic modeling techniques and of computer power promote further development of such sophisticated systems. In this thesis a new ATCS prototype has been developed and evaluated. The development has been motivated by recent research.

Based on a comprehensive overview of the state-of-the-art of science and technology of traffic signal control the remaining need for research has been derived, followed by an overview of the conceptual design of the new ATCS prototype. The prototype employs a centralized concept and uses an optimization interval of 15 minutes, i.e. every quarter of an hour signal timings of all signalized intersections are optimized on a central computer and sent to the local controllers where the traffic signals are controlled accordingly.

The first major task that has to be performed at each interval is to estimate the traffic demand of the next optimization interval. It is assumed that all lanes at all signalized intersections are equipped with vehicle detectors. Based on the detector counts of previous time intervals and on reference traffic demand patterns, a forecasting module estimates the detector counts of the next interval. In a next step, these counts are used as constraints for the estimation of the overall traffic demand in the whole network. This demand comprises Origin-Destination flows, traffic volumes on different routes and on all links of the network. The demand estimation module is based on previous research which built on information theory. The method has been implemented and further generalized to be applicable in the framework of an ATCS. After application of the two modules, an estimate of the upcoming traffic demand in the whole network is available which can be used for optimization of signal settings,

The next two modules of the ATCS perform an adjustment of cycle length and phase durations and an optimization of offsets. The first is done by implementing classic formulas for the calculation of fixed time signal plans. The module calculates a network-wide common cycle length in order to enable coordination of the intersections.

The main focus of the adjustment of signal settings to the currently estimated traffic demand has been on the model-based offset optimization. Offset optimization aims at establishing a good coordination of adjacent intersections in such a way that vehicles do not have to stop at each intersection but can travel in so called green bands. A macroscopic traffic flow model has been used to evaluate the effects of different offset combinations in terms of total delay. For each offset combination to be tested, a single run of the model has to be performed. Different optimization algorithms have been implemented, thereof two based on heuristic Genetic Algorithms. A third, deterministic algorithm has been developed in addition.

The last major object of research of this thesis was signal plan transition. At the beginning of each time interval, the new signal timings have to be implemented at each intersection. This requires application of an appropriate technique of signal plan transition that does not induce major disturbances of traffic flow. Based on the state-of-the-art of signal plan transition and on an additional simulation study a rather smooth transition technique has been identified which has been implemented in the framework of the ATCS prototype. It has been incorporated into the model-based offset optimization in order to consider the effects of signal plan transition directly during the optimization process.

Finally, a comprehensive microsimulation study has been performed to evaluate the performance of the ATCS prototype. The prototype has been applied to two networks in the city of Hanover, Germany. The results revealed that the prototype of the newly developed ATCS has some potential to improve travel times in a sub-network compared to an optimized fixed time signal control. However, the degree of this improvement depends on the network.

Zusammenfassung

In der städtischen Verkehrssteuerung spielt die Steuerung von Lichtsignalanlagen (LSA) eine entscheidende Rolle. Sie beeinflusst direkt die Qualität des Verkehrsablaufs. Neben Festzeit- und lokalen regelbasierten Steuerungen wurden in der Vergangenheit diverse Netzsteuerungsverfahren entwickelt. Diese steuern alle LSA in einem (Teil-)Netz und zielen darauf ab, die Signalisierung kontinuierlich an die aktuell geschätzte Verkehrsnachfrage anzupassen. Fortschritte in der Verkehrsmodellierung und eine Verbesserung der Computerleistungsfähigkeit ermöglichen die stetige Weiterentwicklung solcher Verfahren. In dieser Arbeit wurde ein neues Netzsteuerungsverfahren prototypisch entwickelt und evaluiert. Motiviert wurde diese Arbeit durch aktuelle Forschungsergebnisse.

Anhand einer umfassenden Literaturanalyse zum Stand der Wissenschaft und Technik bezüglich netzweiter LSA-Steuerungen wurde der weitere Forschungsbedarf abgeleitet, gefolgt von einem Überblick über das Grundkonzept der neuen Netzsteuerung. Der Prototyp optimiert die Signalpläne der einzelnen LSA zentral alle 15 Minuten und sendet sie an die einzelnen Steuergeräte, wo sie entsprechend ausgeführt werden.

Die erste Hauptaufgabe, die alle 15 Minuten durchgeführt werden muss, ist eine Schätzung der Verkehrsnachfrage im nächsten Optimierungsintervall. Das Verfahren setzt voraus, dass sich auf den Fahrstreifen in den Zufahrten der einzelnen Knotenpunkte Fahrzeugdetektoren befinden. Basierend auf den Detektorzählwerten der vier vorangegangenen Zeitintervalle und auf Referenznachfragemustern schätzt das Prognosemodul die Detektorzählwerte des nächsten Zeitintervalls. Im nächsten Schritt werden diese geschätzten Zählwerte als Randbedingungen für eine netzweite Gesamtnachfrageschätzung verwendet. Die Gesamtnachfrage umfasst dabei die Verkehrsstärken verschiedener Quelle-Ziel-Beziehungen sowie die Verkehrsbelastungen auf unterschiedlichen Routen und auf allen Kanten im Netz. Das Nachfrageschätzmodul basiert auf vorangegangenen Forschungsarbeiten, die sich auf die Informationstheorie stützen. Die existierende Methode wurde umgesetzt und verallgemeinert, um sie im Rahmen der neuen Netzsteuerung verwenden zu können. Nach Ausführung der beiden Module liegt eine Schätzung der zu erwartenden netzweiten Verkehrsnachfrage im nächsten Zeitintervall vor, die für die Optimierung der Signalpläne genutzt werden kann.

Die nächsten beiden Module der Netzsteuerung passen die Umlaufzeiten, die Dauern der einzelnen LSA-Phasen und die Versatzzeiten der LSA an die aktuell geschätzte Verkehrsnachfrage an. Umlaufzeiten und Phasendauern werden über einen klassischen Ansatz zur Berechnung

von Festzeitsignalprogrammen justiert. Hierbei wird eine netzweit einheitliche Umlaufzeit gewählt, um eine nachfolgende Koordinierung der LSA zu ermöglichen.

Das Hauptaugenmerk der Signalprogrammanpassung lag auf der Versatzzeitoptimierung. Versatzzeiten werden dazu genutzt, die Signalprogramme der LSA eines Netzes so gegeneinander zu verschieben, dass der Verkehr das Netz möglichst ohne Halte passieren kann. Es wurde ein makroskopisches Verkehrsflussmodell verwendet, um die Auswirkungen unterschiedlicher Versatzzeitkombinationen auf die Wartzeiten im Verkehr abzuschätzen. Jede zu testende Versatzzeitkombination erfordert einen Durchlauf des Modells. Für die Optimierung wurden verschiedene Algorithmen umgesetzt. Zwei dieser Algorithmen basieren auf heuristischen Genetischen Algorithmen. Zusätzlich wurde ein drittes, deterministisches Verfahren entwickelt.

Ein letzter Schwerpunkt der Arbeit lag auf der Berücksichtigung von Signalplanumschaltungen. Zu Beginn jedes neuen Zeitintervalls müssen die neu optimierten Signalpläne an allen LSA umgesetzt werden. Hierzu muss ein Umschaltverfahren zum Einsatz kommen, das möglichst keine Störungen im Verkehrsfluss verursacht. Basierend auf einer Literaturanalyse und einer Simulationsstudie wurde ein störungsarmes Umschaltverfahren identifiziert. Um Umschaltverluste direkt bei der Optimierung berücksichtigen zu können, wurde dieses Verfahren direkt in die Versatzzeitoptimierung integriert.

Schließlich wurde eine umfassende Mikrosimulationsstudie durchgeführt, um die Leistungsfähigkeit des neuen Netzsteuerungsverfahrens zu bewerten. Der Prototyp wurde in zwei Teilnetzen der Stadt Hannover angewandt. Die Ergebnisse zeigen, dass das Verfahren die Reisezeiten im Vergleich zu einer optimierten Festzeitsteuerung reduzieren kann. Jedoch hängt der Grad der erreichten Verbesserung von den Randbedingungen des jeweiligen Netzes ab.

Contents

1 Introduction

1.1 Background

The quality of urban traffic depends on different factors such as design of the network, traffic demand pattern, public transport services, urban traffic control and others. Increasing traffic volumes over the last decades contribute to a reduction of the quality of traffic flow. During peak-hours the situation is often at its worst.

Traffic signal control is one of the key factors in urban traffic control. It directly influences the quality of traffic flow and affects all user groups, i.e. individual motorized traffic, public transport, cyclists and pedestrians. A good system of traffic signal control is therefore of major importance. Classical approaches of signalization include fixed time control with preplanned fixed time signal plans that do not react to stochastic variations of traffic flow. Different fixed times signal plans can be used at different times of the day to consider changes of traffic demand over one day. Another very common technique is vehicle-actuated control which reacts directly to stochastic variations of traffic demand. Green times of underlying preplanned framework signal plans can be lengthened, shortened or altered in some other ways to a certain extent in order to adapt signalization to the current detection of vehicles on the approaching lanes of an intersection. Each controller can react directly to detected vehicles.

During the last 30 years a number of so called Adaptive Traffic Control Systems (ATCS) have been developed that go even a step further. These systems control a set of connected intersections in a network and aim at optimizing traffic signal control in real-time by continuously adapting signalization at the intersections to the currently detected or estimated traffic demand in the network. The major goal is in general to optimize a certain performance index (PI) such as minimization of overall delay, number of stops or fuel consumption or maximization of throughput or combinations thereof. ATCS rely on certain models which try to estimate the effects of different signal timings on traffic flow. The signal timings which perform best are then temporarily implemented at the intersections in real-time until new updated signal timings are available from the ATCS. Either arterials or (sub-)networks comprising several signalized intersections are controlled by one system. Well known ATCS (among others) are SCOOT and SCATS, which are well established nowadays. In Germany, mainly MOTION and BALANCE share the market. Many other ATCS with different philosophies and modes of operation exist.

1.2 Objectives

Improvement of traffic modeling techniques and increasing computing power promote continuous enhancement and further development of ATCS. The aim of this thesis was therefore to develop and evaluate a prototype of a new ATCS based on recent research results.

The main motivation for this project was drawn from the work by ALMASRI (2006) who dealt with model-based offset optimization for signalized networks. Offset optimization is used to establish a good coordination of adjacent signalized intersections in a network in such a way that platoons of vehicles can pass the network without having to stop at each intersection. This strategy does not only reduce the number of stops, which contributes significantly to a reduction of exhaust emissions, but also delays experienced by the drivers. ALMASRI (2006) used a macroscopic traffic flow model called Cell Transmission Model (CTM) by DAGANZO (1994, 1995) to estimate the delay of vehicles subject to different offset combinations. For optimization of offsets he applied a Genetic Algorithm (GA). A GA is a heuristic and stochastic optimization technique. His concept had been designed to optimize offsets of fixed time signal plans offline for a given static traffic demand. His work showed promising results. Other research projects, too, used the CTM in combination with optimization techniques such as GA. However, no reports on real-time applications of this concept in the framework of an ATCS have been available so far. Building on the work by ALMASRI (2006), this gap has been closed in this thesis by developing, implementing and evaluating the prototypical ATCS. The original method has been transformed into an online capable version that continuously adapts signal timings of related intersections at constant intervals of 15 minutes. In order to be more flexible, an additional adjustment of other signal timing parameters has been included into the ATCS prototype as well. However, these adjustments are not model-based.

The CTM has some limitations in modeling urban intersections. While ALMASRI (2006) used the original version of the CTM, which was primarily designed for highway applications, another aim of this thesis was to further adapt the CTM in order to enable better modeling of urban intersections with and without traffic signal control.

The original approach of offset optimization by ALMASRI (2006) used a given static traffic demand for the optimization of offsets. If used in the framework of an ATCS, a reliable estimation of the current traffic demand is needed. This thesis therefore also built on another recent research project by WANG (2008) who investigated on an improvement of the estimation of so called Origin-Destination (OD) matrices of a network by means of the Information Minimization (IM) model by VAN ZUYLEN/WILLUMSEN (1980). WANG (2008) improved the quality of estimation by considering and eliminating the effects of redundant constraints of the estimation process. Constraints are detector counts of vehicle flows at different locations in the network which have to be reproduced by the estimated OD matrix after according assignment to the different routes of the network. She furthermore incorporated different traffic assignment techniques into the estimation process.

Since a suitable demand estimation was needed in the framework of the new ATCS prototype, a further aim was to incorporate the improved method of OD matrix estimation by WANG (2008) into the prototype. The approach has been analyzed and generalized for arbitrary networks. It has also been enhanced to estimate not only traffic flows between OD pairs, but also on different routes in the network and on all links.

Demand estimation is based on detector counts. These counts have been made during previous time intervals, whereas optimization of signal timings has to be performed for the upcoming traffic demand of the next optimization interval. Therefore, a fourth aim of this thesis was to include an appropriate forecasting technique of detector counts and to investigate its influence on the general performance of the ATCS in order to decide whether such forecasting is really needed or can be neglected.

Another important objective of this thesis was to investigate the effects of signal plan transition. At the beginning of each time interval, the new optimized signal plans have to be implemented at each intersection. This implementation requires a transition period which may have a disruptive impact on traffic flow. The extent of these disruptions depends on the transition method, and therefore increases in delay or travel time may be more or less severe. The task of transitioning becomes even more complex in coordinated networks where new coordination patterns have to be established quickly without inducing major disturbances. Therefore, transition effects of different transition techniques in a meshed network with different coordinated relations have been investigated in this thesis in order to decide which of these techniques should be used in the framework of the ATCS prototype. Furthermore, an approach has been developed to include the effects of transition directly into the process of optimization.

The last objective of this thesis was to evaluate the new ATCS prototype. An extensive microsimulation study using the commercial software AIMSUN NG 5.1.8 (TSS, 2006) has been conducted. Two different networks have been modeled and the performance of the ATCS prototype has been compared to a reference case with optimized fixed time signal control.

The ATCS prototype developed in this thesis has its focus on individual motorized traffic. Public transport, cyclists and pedestrians have not been considered explicitly so far.

1.3 Outline

Chapter 2 presents the state-of-the-art of science and technology. Some fundamentals of traffic signal control will be provided, followed by a presentation of different existing ATCS. Previous research projects that used GA and/or the CTM for optimization of traffic signal control will be cited. Based on this literature review the need for further research that has been addressed in this thesis will be derived.

Some fundamentals of this thesis will be provided in chapter 3. The conceptual design of the ATCS prototype will be presented. In a next step, the setup of the simulation study that has been used to validate the implementation of the CTM and to evaluate the performance of the modules for demand estimation and the overall ATCS prototype will be described. The last part of this chapter presents the statistical indicators and the statistical test that have been used in this thesis for evaluation of the modules and the ATCS.

Chapter 4 deals with the CTM. The model will be explained, along with some extensions that have been used in this thesis. Furthermore, the chapter covers the implementation and validation of the model.

In chapter 5 the modules for demand estimation will be addressed, i.e. the techniques to forecast detector counts and to estimate OD matrices, route and link volumes. The basic principles, some adaptations, and the performance of the modules will be discussed.

The optimization of signal timings is presented in chapter 6. First, a simple approach to adjust the cycle length and phase durations will be described, before the model-based offset optimization is addressed. The concept of GA will be presented, and three different algorithms for offset optimization will be derived, thereof two based on ALMASRI (2006).

Chapter 7 deals with the problem of signal plan transition. After a short discussion of the state-of-the-art of signal plan transition, an experimental study will be described that assessed different transition techniques. Finally, implementation of the best transition technique into the framework of the new ATCS prototype will be addressed.

Chapter 8 comprises the comprehensive microsimulation study that has been conducted to assess the overall performance of the final ATCS prototype. An additional second test network will be introduced and all results will be thoroughly discussed.

The thesis closes with a summary and an outlook in chapter 9.

2 State-of-the-art of science and technology

2.1 Overview

This chapter describes the state-of-the-art of science and technology with relevance to this thesis. Paragraph 2.2 introduces some fundamental terms of traffic signal control. Paragraphs 2.3 and 2.4 give a short overview of fixed time and vehicle actuated control. The basic philosophy of ATCS, some prominent examples, and a review of studies on their performance will be described in paragraph 2.5. Paragraphs 2.6 and 2.7 comprise a literature review on the use of GA and the CTM respectively in the context of traffic signal control. Finally, the need for further research that has been addressed in this thesis is derived in paragraph 2.8.

2.2 Fundamentals of traffic signal control

Traffic signal control is an important factor of urban traffic control and management. Signalized intersections are installed for two reasons: improvement of traffic safety and improvement of the quality of traffic flow, especially at intersections where high traffic volumes occur. Traffic signal control can be applied at single isolated intersections, at several intersections along an arterial or even at intersections in a meshed network consisting of different intersecting streets and roads.

Before the different types of traffic signal control are described in the next paragraph, some fundamental terms of traffic signal control are introduced first. Figure 2-1 and figure 2-2 illustrate most of the terms.

Signal head: a single traffic signal that can display different signal indications such as green, red, or the intermediate signal indications amber and/or red-amber.

Signal group: a set of one or more traffic signal heads that display at all times the same signal indication for a single turning movement or lane respectively (left, through, right), or for a combination of these turning movements on separate or mixed lanes.

Signal plan: a set of switching times that indicate the points in time in seconds when each signal group changes its signal indication from green to red and vice versa (fixed time signal plan).

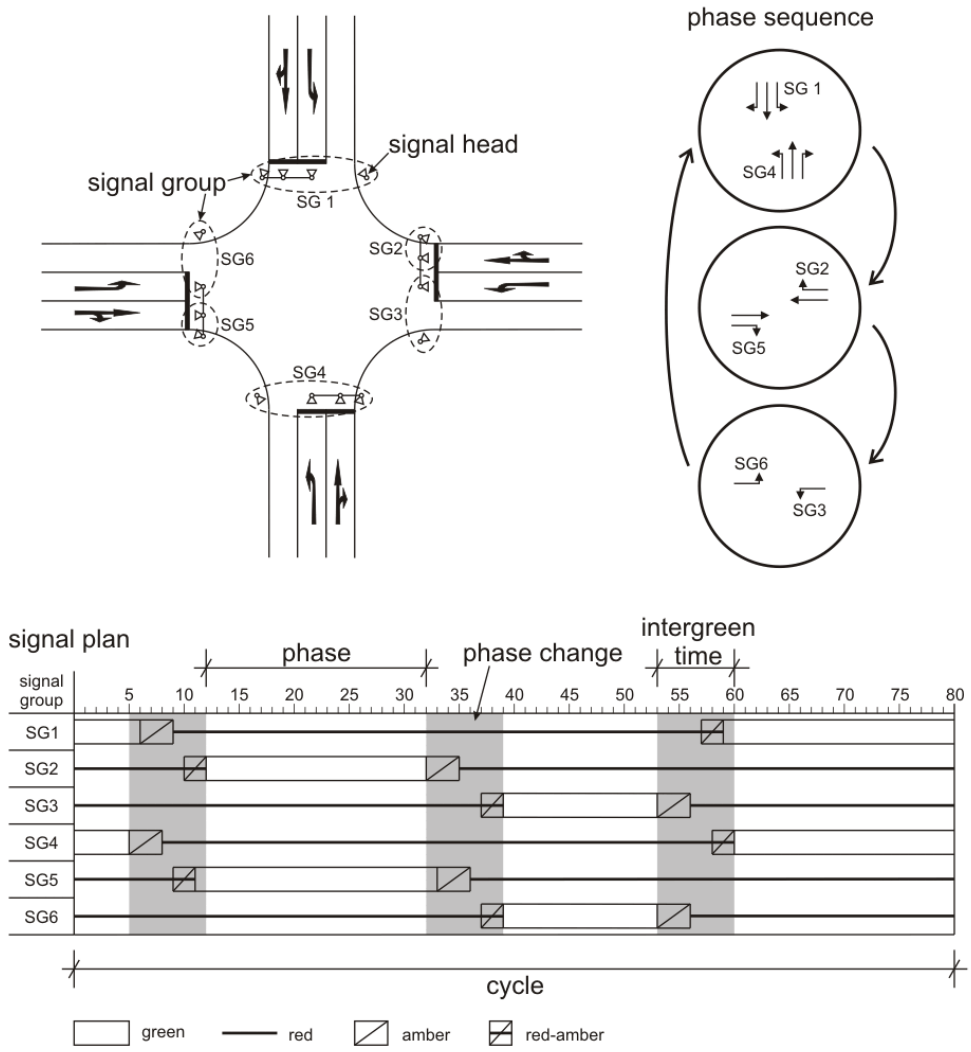


Figure 2-1: Fundamental terms of traffic signal control

Phase: a time interval of a signal plan during which the signal indication of all signal groups remains unchanged. Green times of individual signal groups may start and end at different times.

Intergreen time: the time interval between the end of the green signal indication of a signal group and the start of the green signal indication of a conflicting signal group. This time interval must be long enough to allow clearing vehicles to pass the conflict area safely before entering vehicles reach it.

Phase change: a time interval of a signal plan during which the signal indication of all affected signal groups changes from one phase to the next. The phase change contains intermediate signal indications for all signal groups whose signal indication changes between the two phases. It starts at the end of the green time of the earliest changing signal group and ends at the beginning of the green time of the latest changing signal group.

Cycle: a single repetition of a signal plan consisting of a sequence of phases and phase changes such that each signal group displays a green signal at least once per cycle.

Cycle length: duration of a complete cycle in seconds.

Phase sequence: the sequence of phases that is applied during one cycle.

Split: share of the available total green time per cycle that is assigned to a specific phase of the signal plan. The available green time is the cycle length minus the sum of the durations of all phase changes. The split directly results in corresponding phase durations and thus in green times of each signal group. The durations of phase changes however are fixed and independent from splits.

Offset: an amount of time in seconds by which the entire signal plan of an intersection is shifted in time against a reference point in time. This concept is called the absolute offset. The offset can take a value between zero and the cycle length minus one. It is used at arterials or in networks to coordinate signalized intersections in a progressive signal system ("green wave") such that (ideally) platoons of vehicles can pass a sequence of intersections in a so called green band without stopping. The difference of the absolute offsets of two adjacent intersections is called the relative offset between these intersections. It is equal to the time span between the start of the green signal indications at both intersections for the coordinated traffic stream.

Saturation flow rate: the maximum flow rate in vehicles per hour [veh/h] at which queued vehicles can discharge after start of the green signal indication.

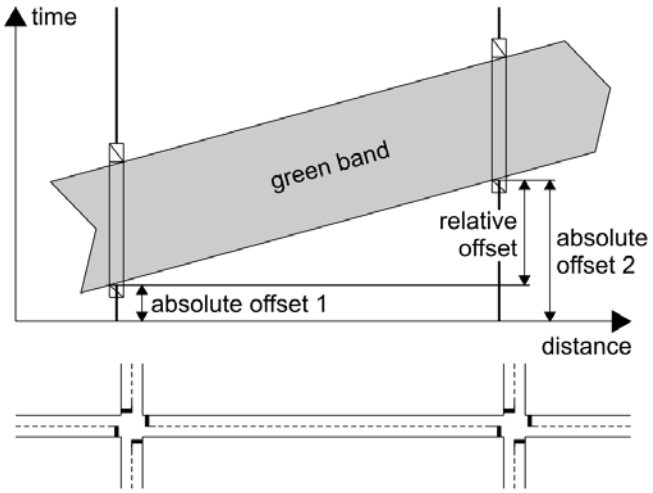


Figure 2-2: Definition of offsets

2.3 Fixed time control

The simplest method of traffic signal control is fixed time control. The same sequence of phases and phase changes is applied at each cycle with no variation of any kind. The cycle length, all phases and the splits are determined for an average traffic demand that has been measured on-site. Fixed offsets can be planned to allow for a good progression of vehicles under the assumption of a fixed cruising speed between intersections.

The control does not react to stochastic variations of traffic flow. To cope with different traffic demands that may arise at different times of the day, for instance during the morning and afternoon peak hour or at off-peak intervals, a set of different fixed-time signal plans is provided in general. The signal plans are stored in the traffic signal controllers where they are chosen either

by time-of-day or by traffic measurements at strategic detector locations that indicate a significant change in traffic demand. They are executed over time periods of several hours.

Application of fixed-time signal control is simple and inexpensive. However, it can only be used effectively if the assumption of a relatively constant traffic demand at the respective intersection is justifiable. Since traffic demand patterns may change slowly but significantly over the years, fixed time signal plans are at risk to be outdated after some time which requires constant examination and adaptation of signal plans.

Appropriate calculation rules to establish fixed time signal plans for a given demand are available from manuals and standards such as the German “Richtlinien für Lichtsignalanlagen – RiLSA” (Guidelines for Traffic Signals, FGSV, 2010) and the “Handbuch für die Bemessung von Straßenverkehrsanlagen – HBS” (Manual for the Dimensioning of Road Infrastructure, FGSV, 2001). These manuals cover manual calculation procedures for isolated intersections as well as for the coordination of intersections along an arterial.

Another technique to determine appropriate fixed-time signal plans for a given demand especially for the case of whole networks is to use adequate software tools. One of the best-known tools world-wide to design fixed time signal plans is TRANSYT (Traffic Network Study Tool, HALE, 2005) which has also been used in this thesis to determine appropriate reference signal plans for the test networks. TRANSYT is based on initial research by ROBERTSON (1969) and has evolved and still does evolve until today. It is a model-based offline optimization tool. The network is modeled as a set of links and intersections. Traffic demand has to be defined by link flows and turning portions at intersections. The structure of the phases and phase changes at each intersection has to be defined by the user. Based on these input data TRANSYT optimizes arbitrary combinations of cycle length, splits and offsets of all intersections in the modeled network. Different criteria such as total delay or quality of progression can be chosen for optimization. The effects of different signal settings are estimated by time-discrete modeling of traffic flow in the network. So called cyclic flow profiles that represent variations of traffic flow at the beginning of each link over one cycle are propagated along the links in consideration of platoon dispersion. While the original version of TRANSYT only used a hill climbing algorithm for optimization, the latest version also offers GA as an option. The optimization results in fixed time signal plans for each intersection which can then be implemented and executed on the traffic signal controllers in the real network.

2.4 Vehicle actuated control

In contrast to fixed time control, vehicle actuated control is designed to react to stochastic variations of traffic demand at an intersection. Some or all approaching lanes are equipped with inductive loops or other type of vehicle detectors. These detectors measure single vehicles as well as headways between successive vehicles, and occupancies, i.e. the portions of time that a detector is occupied by a vehicle. Based on these measurements the signal controller can effect minor and major variations of the current framework signal plan. In contrast to a fixed time signal plan, a framework signal plan does not define fixed switching times for each signal group. In general, overlapping periods during which different phases may be displayed in one cycle are defined instead. Within these constraints the local traffic-responsive controller decides when to start and end each phase subject to the current traffic situation captured by vehicle detectors on the approaching lane. Depending on the degree of freedom phases can be merely aborted or

extended or even be skipped, inserted and swapped, resulting in variable phase durations and sequences during each cycle. Simple processing logics are programmed to define the desired behavior of the controller. These logics have to make sure that minimum and maximum green and red times are maintained. While vehicle actuated control is only applied at individual intersections, coordination of adjacent intersections can be implemented by fixed offsets that shift the framework signal plans accordingly.

Recent research has shown that vehicle actuated control does not always have a clear benefit over fixed time control (WIETHOLT, 2009). Especially during periods of high traffic demands which are close to the intersection's capacity vehicle actuated control tends to act like fixed time control since all phases are extended to their maximum phase duration.

2.5 Adaptive Traffic Control Systems

2.5.1 Basic principles

While vehicle actuated control only reacts to stochastic variations of traffic flow on a local level, ATCS aim at optimizing traffic signal control of a set of intersections in a network context according to the current traffic demand. Such systems are also often referred to as Urban Traffic Control (UTC). The general idea of ATCS is to adapt the traffic signal settings of all signalized intersections in a network in such a way that they favor the current traffic demand. ATCS therefore rely on a set of detectors located at different positions in the network whose traffic counts and other measurements are used to estimate the current or upcoming traffic demand. Based on this estimation, traffic models and/or impact models are used to estimate the impact of different possible signal settings which can be applied within the next few seconds or minutes, depending on the ATCS. Traffic signal settings are optimized by different optimization techniques. In general, ATCS try to minimize or maximize a PI which can be total delay, number of stops, fuel consumption, throughput etc. or weighted combinations thereof. The traffic signal settings that produce the best PI according to the applied optimization technique are then chosen to be used during the next time interval. Different philosophies of ATCS exist. Optimization can be performed on a central level for all intersections at once, or in a decentralized way where decisions made on a local level are assumed to produce a good solution even from a network-wide point of view. Both philosophies can also be combined.

Adaptive Traffic Control is a field of science of very high interest for researchers and practitioners likewise. Correspondingly, numerous ATCS have been developed in the past years and decades, and numerous summarizing overviews have been published which shortly describe and compare the different strategies, e.g. WOOD (1993), FRIEDRICH (1997), MERTZ (2001), HOUNSELL/MCDONALD (2001), PAPAGEORGIOU ET AL. (2003, 2007), ALMASRI (2006), BRAUN (2008) and VAN KATWIJK (2008) to name only a few. In order to enable classification of the ATCS prototype developed in this thesis in the context of existing ATCS a brief description of selected ATCS is presented in the following without intending to be exhaustive and despite the already existing amount of such overviews. However, only the basic ideas and concepts behind the different strategies are described shortly, based on the according primary sources, where available, or on the aforementioned reviews. For more information the reader is referred to these sources. It has to be mentioned that some of the sources do not allow for detailed com-

prehension of the respective strategy, intentionally or unintentionally, and only some fundamental remarks can be made.

2.5.2 Examples of existing ATCS

2.5.2.1 SCOOT

SCOOT (Split, Cycle and Offset Optimization Technique) has been developed in the early eighties (HUNT ET AL., 1981) and is one of the most renowned ATCS worldwide with currently more than 200 installations in over 14 countries (SCOOT, 2010). It is a centralized system, i.e. all decisions are computed on a central computer. The computer transfers the respective commands to switch signals to the individual controllers at each intersection where they are merely executed without any further intervention.

SCOOT uses a traffic model that is similar to the one used in the offline planning tool TRANSYT. Therefore SCOOT is often referred to as the online version of TRANSYT. The traffic model relies on so called cyclic flow profiles. These flow profiles consist of traffic flows and occupancies measured by inductive loops at the upstream end of each approach of an intersection, i.e. the detectors are located directly at the exits of each intersection where the risk of congestion is low. The flows are measured for relatively small fractions of a complete cycle so that the temporal resolution of the profile is high and variations of the flow rate over one cycle are visible. The traffic model propagates the measured flow profiles in downstream direction in order to make a short-term prediction of queue lengths, delays and stops at each intersection. An average cruising speed as well as the effects of platoon dispersion are taken into account. A configurable saturation rate is assumed for vehicles discharging at green.

SCOOT tries to minimize a PI which is a weighted combination of the sum of queue lengths, delays and stops. This minimization is done by frequent small iterations of the settings of an underlying fixed time signal plan. Before each change of a signal phase the signal optimizer decides whether to reduce or to lengthen the current phase by a maximum of four seconds or to leave the phase duration as it is. Such a temporary change of the green split leads to a permanent change of the phase duration of one second during the next cycle. Once per cycle a decision is made whether to slightly change the offsets of individual intersections by minus or plus four seconds in order to improve traffic progression between a pair of intersections. The common cycle time in a sub-area of grouped intersections is adapted at intervals of 2½ minutes or more. Again, increments are small. The aim of this cycle length adaption is to keep the degree of saturation at all approaches of all intersection in the sub-region below 90 percent. Different strategies are used during congestion when the detectors are occupied by stationary queues.

Later developments of SCOOT include active bus priority (BOWEN/BRETHERTON, 1996), estimation of vehicle emissions and enhancements of the cycle length adjustment by incorporating average flows taken from a database with historic traffic flow records (BRETHERTON ET AL., 1998).

2.5.2.2 SCATS

The development of the Australian SCATS (Sydney Coordinated Adaptive Traffic System) began in the early seventies (LOWRIE, 1982). Together with SCOOT it is one of the most widely used ATCS worldwide controlling more than 30,000 intersections with most applications in

Australia, Asia and South America according to SCATS (2010). SCATS has a hierarchical structure. Regional computers control up to 120 intersections which are grouped into independent systems and sub-systems, the latter comprising one to ten intersections. Adjacent sub-systems can “marry” if traffic conditions favor such a linkage.

In contrast to SCOOT, SCATS does not make use of a traffic model. Green splits, offsets and cycle lengths are continuously adapted by more or less simple algorithms instead. SIMS/DOBINSON (1980) state that the underlying strategic options of the adjustments are minimum delays, minimum stops, or maximum throughput. The decisions of the regional computer rely on data from detectors that, unlike SCOOT, are located immediately in front of the stop lines. They measure the number of passing vehicles and the total time that a detector is unoccupied during green.

In a first step, SCATS adjusts cycle lengths and green splits. According to SIMS/DOBINSON (1980) this task has to be performed first, before any benefits from coordination can be attained. A common cycle length is determined for each sub-system. Based on the detector measurements SCATS determines the share of effectively used green time for each phase of each intersection of the sub-system, referred to as degree of saturation (DS) by LOWRIE (1982). A formula that takes into account the currently highest DS in the sub-system is used to continuously adjust the cycle length at each cycle. The maximum increment is six seconds. The cycle length is constrained by a minimum and maximum value.

For green split adjustment SCATS uses fixed background green split plans for each intersection. These plans define how to distribute the available total green time of one cycle among the phases. Each cycle a green split plan is voted for that would produce the lowest DS at the respective intersection given the current detector measurements. If a green split plan gets two votes within three cycles, it is finally applied.

A similar vote is performed to select a background plan for internal offsets of a sub-system. This vote is based on the approaches with the currently highest traffic flows. The plan that receives four votes within five cycles is selected. The offsets can be suited to the current cycle length by a given formula. The decision on whether to link adjacent sub-systems by an appropriate external offset plan is made by a similar procedure of consecutive votes. A positive vote is given if cycle lengths of the two sub-systems are close to each other. In the case of linkage, the higher of the cycle lengths of each sub-system is used for both.

All the above mentioned rules lead to control settings that are sent to the local controllers where additional vehicle actuation is applied on a tactical level based on measurements of the same stop line detectors mentioned above.

2.5.2.3 UTOPIA

UTOPIA (Urban Traffic Optimization by Integrated Automation), developed in the eighties in Italy, is a hierarchical-decentralized ATCS with an area control level and a local level (DONATI ET AL., 1984, MAURO/DI TARANTO, 1990, MIZAR AUTOMAZIONE, 1993). The first installation was in Turin, Italy, but there are other installations in Europe by now (PEEK, 2010). UTOPIA aims at minimizing delays for private vehicles while giving priority to selected public transport vehicles at intersections. Optimization of signalization is subject to phase sequences and minimum and maximum phase durations. Upstream detectors for public transport and private vehicles are

used to detect every single vehicle. Additional public transport detectors are located at bus stops.

On the area control level UTOPIA makes use of a macroscopic model of private traffic in the whole network. The street network is divided into interconnected storage units. Based on a prediction of OD flows from detector counts the model calculates the number of vehicles in these storage units and the flows in between for consecutive discrete time intervals by applying a vehicle law of motion and a vehicle propagation law. The model is used to generate optimized control rules which minimize a cost function describing the overall network performance. The main indicator is total travel time. Optimization uses a rolling horizon of 30 minutes and is executed every time that a so called observer predicts changes in traffic conditions. No details on the optimization technique are available. The control rules contain reference or framework signal plans and weights for the local cost function and are transmitted to the local level. A public transport model is used to forecast section travel times and arrivals at intersections long before the vehicle finally gets there. Each public transport vehicle is modeled individually, generated according to the time table. Current travel times of the vehicles are modeled stochastically and consider free flow travel time, waiting time at stops and lost time at signalized intersections. The predicted public transport arrivals are intended to be used for public transport priority without major disturbances of private vehicles.

On the local level an open-loop feedback strategy is applied at each intersection. A microscopic model of private traffic regards individual intersections as a set of links connecting it to its adjacent intersections. Based on vehicles detected at the upstream end of a link the vector of arrivals at the intersection is estimated in time steps of 3 seconds by a deterministic propagation law. Important parameters are turning percentages, average speeds and saturation flows. The local controller is programmed to minimize at each decision instant a weighted combination of delays, stops, and queues by using a so called branch-and-bound algorithm in combination with a heuristic optimization method. Optimization also includes information from adjacent intersections to enable a concept of “strong interaction” between intersections. The horizon for the optimization covers 120 seconds, but optimization itself is repeated every 3 seconds and the resulting signal settings are in operation during this time step only before being readjusted again. Public transport vehicles with priority constrain the local optimization.

2.5.2.4 OPAC

OPAC (Optimization Policies for Adaptive Control) is an American ATCS initially developed in the early eighties (GARTNER, 1982, 1983, GARTNER ET AL., 1983). It has been refined through the years (GARTNER ET AL., 2001). Some few installations exist in the United States.

The initial test version of OPAC, called OPAC I, is entirely decentralized, i.e. each intersection is controlled independently from all other intersections. OPAC I does not make use of the concept of cycles, green splits and offsets. A sequence of switching times is generated instead which specifies when to change from one phase to the next. The aim is to reduce the total delay experienced over the optimization horizon. The horizon is divided into time intervals of 5 seconds. For each interval a decision on phase changes has to be made, i.e. the phase of the previous interval is either kept or terminated in favor of the next phase. The arrival rates of vehicles at each approach of the intersection during each interval are assumed to be known exactly in OPAC I. Discharge rates at green signals are set to two vehicles per interval. Queuing delays

due to signal control are estimated based on the arrivals and discharges at the signal. A dynamic programming approach determines an optimized phase sequence for the entire horizon that reduces the total delay. This technique guarantees an optimal solution but turned out to be too slow for real-time optimization. OPAC II has therefore been modified in such a way that only one to three phase changes are allowed in a time period of 50 to 100 seconds. Delays are assessed for all feasible solutions, i.e. a complete enumeration is applied. To make OPAC operational for real applications, arrival rates are derived by short-term predictions based on upstream detector measurements in OPAC III. The arrival rates of time intervals within the optimization horizon that are not covered by such predictions are estimated by smoothed average flows. In addition, a rolling horizon concept has been added.

To allow for coordination of intersections, which is not possible without a synchronized network-wide cycle time, a later version called Virtual-Fixed-Cycle OPAC (VFC-OPAC) has been introduced (GARTNER ET AL., 2001). Every few minutes a synchronization layer determines a network-wide cycle length based on the needs of the critical intersection or the majority of intersections (VAN KATWIJK, 2008). On the coordination layer offsets are optimized once per cycle for each intersection based on the current inflows and outflows of intersections. On the local layer, OPAC III optimizes the phase changes as described above subject to the constraints from the two upper layers.

2.5.2.5 MOTION

MOTION (Method for the Optimization of Traffic Signals in Online Controlled Networks) is available since the early nineties and is one of the two major German ATCS. It is distributed by SIEMENS. Most systems are installed in German cities and some other European countries (SIEMENS, 2010).

BUSCH/KRUSE (1993) and BIELEFELDT/BUSCH (1994) gave a first overview of the system. Later, KRUSE/BUSCH (2002) further updated some information. MOTION is a phase and model based ATCS. A network-wide common cycle length, green splits, offsets and phase sequences are optimized every 5, 10 or 15 minutes on a strategic level. The according framework signal plans are then sent to the traffic-responsive controllers which are free to react to stochastic variations of traffic flow on a second-to-second basis within the limits of these framework signal plans.

MOTION needs strategic detectors at all entries and exits of the network as well as on approaches of intersections. All of these detectors should be located ideally in sections with a low risk of congestion. Additional detectors for the local level should be located about 40 meters in front of the stop lines. Aggregated detector counts are used to estimate flows of all turnings at intersections and of traffic streams in the whole network. This task is performed by the Path-Flow-Estimator by BELL ET AL. (1997). The estimates of turning flows are used to calculate a common cycle length and suitable phase durations by means of simple formulas based on the degrees of saturation. The estimates of flows on network traffic streams are sorted from high to low volumes to derive an adequate sequence for the coordination of intersections. The original version of MOTION used the heuristic SIGMA method by GARBEN ET AL. (1988) to coordinate intersections. The two respective optimization criteria are delays and stops, expressed as a normalized index KAPPA which is derived by means of a queuing and traffic flow model. A different source (BUSCH/KRUSE, n.d.) claims that the VERO method by BÖTTGER (1972) was used instead, but this is rather unlikely.

To ease the disruptions caused by signal plan transition every 5 to 15 minutes, a special smooth transition plan procedure is used, but no details on this procedure have been published. However, changed framework signal plans are only sent to the local controllers if significant improvements of the optimization criteria are to be expected. Besides vehicle-actuation, the traffic responsive local controllers consider public transport priority by influencing the current phase sequence subject to the current framework signal plan. MOTION also disposes of an automatic congestion and incident detection and management module that is based on comparisons of occupancies and flows at adjacent detectors.

MÜCK (2008a, 2008b) reports on latest developments in MOTION. A new method for the estimation of turning portions at intersections has been implemented. Furthermore, a GA has been implemented to replace the former phase sequence and offset optimization with SIGMA. Details on this adaptation will be discussed in paragraph 2.6.5.

2.5.2.6 BALANCE

BALANCE (Balancing Adaptive Network Control Method) has been developed in the nineties. It is the second of the two major German ATCS. However, it is not as widespread as MOTION. Its distributor speaks of more than 100 signalized intersections in several German cities (GEVAS, 2010). BALANCE has a similar philosophy as MOTION, i.e. framework signal plans are optimized every 5 to 15 minutes on a network level and sent to the individual traffic responsive controllers that can run any local control strategy that observes the framework signal plans.

The initial version of BALANCE has been described by FRIEDRICH (1997, 2000a, 2000b). BALANCE optimizes the network-wide common cycle length, green splits, phase sequences and offsets. Aggregated detector counts are used as constraints to estimate flow volumes on turnings at intersections and on different routes in the network. The estimation employs a dynamic approach based on correlation analysis of inflow and outflow profiles at intersections (KELLER/PLOSS, 1987) and the method of entropy maximization (VAN ZUYLEN/WILLUMSEN, 1980). Based on the queuing model by KIMBER/HOLLIS (1979) or Markov chains BALANCE estimates queue lengths and delays to optimize the cycle length, which is dictated by the critical intersection, and green splits. Offset optimization in BALANCE is described by FRIEDRICH (1997) as a sequential procedure which handles the offsets of intersections one after another according to the highest estimated network flows and based on a mesoscopic platoon model similar to the one used by SCOOT. BRAUN ET AL. (2008b, 2009) specify that the mesoscopic model is used to calculate deterministic performance indicators whereas the macroscopic queuing model calculates the effects of stochastic variations of traffic flow. The model estimates delays, stops, and queue lengths.

While FRIEDRICH (1997) developed a model-based local control method called MicroBALANCE to be used on the local controllers, a different local control method is used in general in the BALANCE framework. The most common option is the TRENDS kernel (GEVAS, 2005), a traffic-actuated control method which is also distributed by GEVAS. It incorporates public transport priority if desired. Signal plan transition in TRENDS is based on a dwell concept in order to coordinate traffic signals according to the new offsets, i.e. the main phase is held for a certain time period before the new signal plan starts. Thus, no smooth transition method as in MOTION is applied.

While the original BALANCE made use of a hill-climbing algorithm for optimization, the latest version has been modified by a GA that is able to optimize cycle length, green splits, phase sequences and offsets at once (BRAUN ET AL., 2008a, 2008b). A closer discussion of this adaptation will follow in paragraph 2.6.5.

2.5.2.7 TUC

TUC (Traffic-responsive Urban Control) is a relatively new ATCS developed at the Technical University of Crete by DIAKAKI ET AL. (2002, 2003). TUC is in operation in Chania, Greece, and Macaé, Brazil. Some test implementations have been investigated in the UK and Germany.

TUC relies on phase specifications and sequences that are planned offline. The initial version of TUC only optimized green splits subject to minimum and maximum green times. The network-wide cycle length and the offsets of intersections were fixed. TUC employs a store-and-forward modeling technique. The street network is modeled as a directed graph with links and nodes, i.e. intersections. The state of a link at a specific time step is expressed by the current number of vehicles on that link. The duration of a time step equals the duration of the control interval. The current states of all links are combined in a network state vector which is used to set up a system state equation that describes the evolution of the link states from the current to the next control interval. The current states of links are updated by adding all inflows and subtracting all outflows subject to signal control. The system state equation contains a matrix B that reflects network topology, phases, cycle length, saturation flows and turning rates. In order to reduce the risk of oversaturation and queue spillbacks TUC attempts to minimize and balance the occupancies of all links, i.e. their portions of current number of vehicles to maximum number of vehicles that can be stored on the link. Based on the system state equation a linear-quadratic feedback control law is derived. The related control matrix has to be determined only once per network, which is time-consuming but can be done offline. Application of the control law, however, can be performed in real-time. Instead of measuring flows directly to update the number of vehicles on a link, TUC uses detectors located in the middle of each link to measure occupancies which are transformed into estimated numbers of vehicles by a non-linear function. Based on the regulator, the optimized green splits are determined. In case of oversaturation on specific links the regulator prevents more vehicles from entering the link by reducing the according green times at the upstream intersection.

Further extensions of TUC comprise cycle length and offset optimization as well as public transport priority. A feedback control law (P-regulator) adjusts the network-wide cycle length such that the maximum degree of saturation on a pre-specified percentage of links is limited to a certain level. Offsets are modified consecutively for different arterials in a pre-specified order which reflects the different importance of the arterials. Based on a feedback control law TUC performs a decentralized optimization of offsets by consideration of couples of intersections along the arterials. Based on free-flow mean speeds and current estimates of queue lengths TUC aims at adjusting offsets such that vehicles coming from the upstream end of a link arrive at the end of the queue at the moment when the last vehicle in the queue starts moving. Single transient cycles are used to implement the new offsets. Public transport priority is included either during green split optimization by appropriate weighting of detector measurements or by local modification of the network-wide signal settings. For all optimization tasks, different control intervals can be chosen.

2.5.2.8 Decentralized self-control of traffic lights

Even though not yet implemented in a real urban network, the decentralized self-control approach recently developed by LÄMMER (2007) is worth mentioning. A good overview of the operating mode of this method that has not yet been baptized is available by LÄMMER ET AL. (2009). Just like for instance SCOOT and OPAC the method applies a decentralized concept in which each signalized intersection of a network is controlled entirely on its own. Signal control is acyclic to ascertain a maximum of flexibility subject to security related aspects of signalization such as intergreen times, minimum green and red times etc. It is assumed that even without the concept of fixed cycles, phase sequences and offset optimization the system will adjust itself to dynamic traffic signal progression if it is beneficial for the current traffic demand.

The method requires upstream detectors at the beginning of each approach of an intersection and downstream detectors at the stop line. Vehicles detected at the upstream detector are projected to the downstream end of the approach of an intersection to allow for a short-term forecasting of the arrivals within the next few seconds. The vehicles are assigned to different turning lanes according to fixed turning percentages. An anticipation module evaluates the delays that arise if arriving platoons are served immediately or later. This estimation is performed for each signalized turning on all approaches of the intersection. An optimization module selects the currently best signal group combination subject to the current demand on the different approaches. But instead of simulating any feasible switching sequence, the module decides which traffic streams have to be served first. Thus, the intersection eases the “highest pressure”. Since the optimization process might neglect side-street traffic with low volumes, a stabilization process assures that maximum red times are not exceeded and that all queues are discharged regularly.

Since the method is purely academic so far, only evaluations based on simulation exist. LÄMMER (2007) tested individual intersection, arterial and network scenarios. He compared the results against a fixed time control strategy of unkown quality and reported some benefits of his method. Simulation of a real network in Dresden (LÄMMER ET AL., 2009) with 13 signalized intersections resulted in delay savings of 9 percent for private traffic, 56 percent for public transport and 36 percent for pedestrians and cyclists.

2.5.3 Performance of existing ATCS

Many field trials have been conducted to assess the performance of the aforementioned ATCS. Table 2-1 compiles the reported results from different selected studies and gives a general idea of the potential of ATCS. However, such studies always have to be interpreted with care. The improvements achieved by ATCS depend on the quality of the reference case which might be an updated or sometimes also outdated fixed time signal control. Furthermore collection of data to assess the quality of traffic flow might not be statistically sound. The results might only cover main street traffic, neglecting effects on side-street traffic. And finally, the performance of an ATCS always highly depends on the specific constraints of a network which makes it impossible to directly compare different ATCS assessed in different studies. Table 2-1 can thus provide no more than a general overview. It has to be noted that all results apply to private vehicles only. In cases where no values of the percentage reduction of delays, stops or travel times are given, this data was not available from the respective publication.

Table 2-1: Performance of different existing ATCS (continuation on next page)

ATCS	year of study	city	intersections	reference case	percentage reduction in… …delay	…stops	…travel time	source
SCOOT	1989	Beijing	40	FT-nc	-32.3 (am) -19 (pm)	-32.7 (am) -29 (pm)	-15 (am) -1.8 (pm)	PECK ET AL. (2007)
	n/a	Glasgow	n/a	FT-Tr			+2 (am) -10 (pm)	MCDONALD/ HOUNSELL (1991)
	n/a	Coventry, Foleshill	n/a	FT-Tr			-23 (am) -22 (pm)	
	n/a	Coventry, Spon End	n/a	FT-Tr			-8 (am) -4 (pm)	
	n/a	Worcester	n/a	FT-Tr			-7 (am) -19 (pm)	
				VA-c			-31 (am) -20 (pm)	
	n/a	Southamp-ton	n/a	VA-nc			-40 (am) -48 (pm)	
	1993	Toronto	75	FT	-6 to -26	-10 to-31	-6 to -11	GREENOUGH/ KELMAN (1998)
	1996	Nijmegen	n/a	FT	-58		-25	TAALE ET AL. (1998)
	1992	Sao Paulo, Rio Branco	10	FT-Tr	-40 (am) ±0 (pm)			MAZZAMATTI ET AL. (1998)
	1992	Sao Paulo, Alvarenga	8	FT-Tr	-41 (am) ±0 (pm)			
SCATS	1974	Sydney	n/a	FT-c			-39.5 (am) -32.8 (pm)	SIMS/DOBINSON (1980)
	2007	Gresham (US)	11	FT-c			-13.8 to +9.7 (am) -15.8 to -15.5 (pm)	PETERS ET AL. (2007)
	2007	Park City (US)	12	VA-c	-20		-7.6 (am) -3.9 (pm)	STEVANOVIC ET AL. (2008)
UTOPIA	n/a	Turin	n/a				-10 to -15	BARNHART/ LAPORTE (2007)
OPAC	1996	New Brunswick	15	VA-c		±0 (am) -55.4 (pm)	±0 (am) -25.6 (pm)	ANDREWS ET AL. (1997)
MOTION	n/a	Piraeus	22	FT			-8 to -14	KRUSE/BUSCH (2002)
	2007	Münster	20	FT			-3 to -7 (am) ±0 to -12 (pm)	MÜCK (2008a, 2008b)
	2008	Münster	24	FT		-17	-11	BRILON ET AL. (2009)
				VA		-15	-9	
BALANCE	n/a	London	n/a	MOVA	-4			FRIEDRICH (2000b)
	n/a	Munich-Riem	n/a	FT-c	-38 to -39	-30 to -38		GANSER (2003)
	2004	Hamburg	13	FT-c	-13.6	-14	-6,4	CITY OF HAM-BURG (2005), KOCH (2006)
				VA-c	-27.4		-10.8	

Table 2-1: Performance of different existing ATCS (continuation of previous page)

ATCS	year of study	city	intersections	reference case	percentage reduction in… …delay	…stops	…travel time	source
BALANCE	2006 2008	Ingolstadt	46	VA	-19 (am) -32 (pm)	-9 (am) -32 (pm)		BRAUN ET AL. (2009)
TUC	2001	Chania	2	TASS			-11	DINOPOULO ET AL. (2006)
	2003	Chania	23	TASS			-5 to -25	BIELEFELDT ET AL. (2004), KOSMATOPOULOS ET AL. (2006)
	2003	Southampton, City Center	35	SCOOT			+10 (am) +19 (pm)	
	2003	Southampton, Biterne	18	SCOOT			-30 (am) -5 (pm)	
	2004	Munich	25	BALANCE			-2 to -15.4	

FT = fixed time control, VA = vehicle actuated control, c = coordination, nc = no coordination, Tr = Transyt, am = morning hours, pm = afternoon/evening hours, n/a = not available

2.6 Genetic Algorithms in traffic signal control

2.6.1 Basic principles

A GA is a heuristic and stochastic optimization technique that is designed to search a highly irregular non-convex solution space in order to find solutions close to the global optimum. The development of GA has been inspired by the principle of evolution. GA start from a given set of solutions which are created randomly and may include solutions known to be of good quality or found by other optimization techniques. The performance or fitness of each solution is evaluated by a fitness function. Then, a number of solutions are selected from this initial set of solutions. Solutions with better fitness have a higher probability to be selected. The selected solutions are combined and modified according to certain rules to form new solutions which are evaluated again. These new solutions form a new pool for selection. The process is repeated again and again until a stop criterion is met. The final solution is the best one within the final pool of solutions. Further details of the concept will be presented in paragraph 6.3.2.

During the last two decades, GA became more and more popular in traffic signal control. The following paragraphs summarize relevant publications that applied GA to solve the problem of optimizing traffic signal timings.

2.6.2 Optimization of signal timings

Presumably the first to experiment with GA in order to optimize signal timings were FOY ET AL. (1992). Using a very simplistic microscopic and time-discrete traffic flow model to evaluate the fitness of different signal settings in terms of total delay, they optimized green splits at four intersections of a simple grid network. They highlighted that the purpose of the study was to assess the general potential of GA in traffic signal optimization, and that their approach was not ready to be used in a real network. No comparison between the model and real traffic has been made.

HADI/WALLACE (1993) proposed two different ways to incorporate a GA into TRANSYT. In the first alternative the GA optimized cycle length and phase sequences of all intersections in a network, whereas offsets and green splits for each solution of one generation were calculated by the traditional hill-climbing algorithm of TRANSYT. In the second alternative the GA was allowed to additionally optimize offsets, too. In both cases, a TRANSYT specific measure of progression was used as fitness value of a potential solution. The promising results of this study lead to the inclusion of GA optimization in TRANSYT as an alternative to the hill-climbing technique (cf. paragraph 2.3).

ABU-LEBDEH/BENEKOHAL (1997) used GA for traffic signal control and queue management along an oversaturated one-directional arterial. Their aim was to identify optimal green times and offsets over several cycles of variable length in order to maximize throughput. For evaluation of different signal settings they developed a dynamic mathematical formulation that takes into account changing states of the links, i.e. changing numbers of vehicles on each link per cycle. A state-equation updates queue lengths on all links for each cycle based on arriving and discharging vehicles during the previous cycle. The formulation has been set up in such a way that traffic which is discharged at an upstream intersection reaches the end of the queue on the next link when the latter starts moving. The necessary speed of acceleration wave is a parameter of the formulation. In addition, the concept tries to prevent de-facto red, i.e. green periods that cannot be used due to spill-backs. This requires a stop wave speed as a second important parameter.

The mathematical formulation was used as fitness function which was executed 20,000 times during one GA run for an arterial with five intersections, no turnings, given initial queue lengths and a constant average inflow at crossing approaches. The authors found that after some initial cycles the GA chooses a similar cycle length for all intersections. During the first cycles the relative offsets between intersections are negative because of queue clearance and become positive during the last cycles to ensure good signal progression. Both cycle lengths and offsets reveal that a normal forward progression of vehicles is obtained after the existing queues have dispersed. No tests of the resulting signal settings in a simulation environment have been conducted.

PARK ET AL (1999, 2000) tested GA-based optimization of signal timings at oversaturated intersections. They used different objective functions for throughput maximization and delay minimization, incorporating a penalty function for intersection turning movements with volume-to-capacity ratios greater than 0.9. A mesoscopic simulator has been used to evaluate different signal settings within the GA framework. The simulator included binomial arrival distribution of vehicles, a saturation flow rate model, a queue evolution model and three fixed turning percentage vectors per intersection approach which depend on the upstream movement, i.e. on the direction from which vehicles entered the approach. The authors optimized cycle length, offsets, green splits and phase sequences simultaneously and used an artificial arterial network with four intersections. The test cases covered different levels of traffic demand. The microsimulator CORSIM was used to evaluate the optimized signal plans and to compare them to a reference case created with TRANSYT. The latter performed worse. Computation time was about 23 minutes for one optimization run, but it has to be taken into account that computer technology evolved tremendously since then.

BRAUN/WEICHENMEIER (2005a, 2005b) used the traffic flow model of BALANCE to evaluate different signal settings in the framework of a GA. The GA optimized cycle length, phase se-

quences, offsets and green splits at once. The procedure has been designed for offline application and has been tested in a real network in Regensburg, Germany, with six intersections. The optimized signal plans have been implemented on the real controllers on-site. Travel times on three different routes have been measured via floating car data and vehicle re-identification. The results were inconsistent. Both improvements and deteriorations have been observed, depending on the time of day and the routes. Nevertheless, the authors conclude that all in all the algorithm was able to improve the performance of the traffic signal settings.

2.6.3 Combination of traffic assignment and optimization of signal control

LEE/MACHEMEHL (1998) dealt with the problem of combined traffic assignment and signal control. They tried to identify optimal green splits of intersections in a meshed network for a given OD matrix assuming User Equilibrium assignment. Travel times on links which are needed for traffic assignment consisted of a constant free flow travel time and additional delays estimated by means of the Webster delay formula (WEBSTER, 1958). The measure of performance of a solution was chosen to be the total network travel time. The authors tested different optimization techniques: a GA, a gradient-based local search technique and an iterative procedure of repeated traffic assignment and signal optimization. They applied all three algorithms to different sample networks and concluded that the GA reduced the risk of finding a suboptimal solution whereas the other two techniques converged much faster.

CEYLAN/BELL (2004, 2005) applied a more sophisticated approach to the same problem. They developed a method called GATRANSPFE which optimizes a network-wide cycle length, green splits and offsets in a meshed network for a given OD matrix. For each solution, i.e. each combination of signal settings that has to be evaluated during a GA optimization, two steps are taken. First, the Path Flow Estimator is used to determine the according link flows subject to Stochastic User Equilibrium (SUE). Then the TRANSYT traffic flow model is used to evaluate the performance of the current signal settings subject to the estimated link flows. GATRANSPFE took 18.4 h to find a solution for a meshed network with six signalized intersections. The authors found their approach to be efficient and much simpler than other heuristic algorithms.

TEKLU ET AL. (2007) built on the work by CEYLAN/BELL (2004, 2005), but replaced TRANSYT and the Path Flow Estimator by the combined simulation-assignment modeling software package SATURN (VAN VLIET, 1982). They applied the method to a real network in the city of Chester, UK, with a total of 75 signalized intersections. They found that superiority of the GA based method over other optimization techniques was more pronounced at higher congestion rates, whereas no significant differences were observed for lower traffic demands.

2.6.4 Microsimulation-based approaches

YUN/PARK (2006) used the CORSIM microsimulator to calculate the fitness values of different signal settings in the framework of a GA. Besides the traditional signal control settings they also included the option to optimize traffic controller settings such as minimum green times, vehicle extension times etc. and also detector locations. Each evaluation of such settings required five CORSIM simulation runs. The median value of either queue time or control delay has been used as measure of effectiveness of each tested solution, i.e. as the corresponding fitness value. Several test networks have been created artificially or chosen from real sites to evaluate the method. The optimized signal settings achieved significant reductions in delay compared to

other optimization tools. Even though no remark on the runtime of the optimization process could be found in the report, the algorithm must be considered to be very slow due to the generally long runtimes of microsimulators compared to macroscopic models.

STEVANOVIC ET AL. (2007) pursued a similar approach. They combined a GA with the microsimulator VISSIM to optimize cycle length, green splits, phase sequences and offsets in a network. VISSIM was used to evaluate different signal setting combinations. Different fitness values are offered, including total delay, total travel time, number of stops and throughput. For each possible solution, the results of five VISSIM simulation runs have been averaged. Tests have been conducted using a real arterial network in Park City, Utah. Since microsimulation is very time consuming, a single optimization run which was executed in parallel on 10 computers took about 90 hours. Depending on the scenario, the optimized signal plans produced 5 to 15 percent less delay than those optimized with other techniques.

2.6.5 Online control

The studies mentioned so far investigated on offline optimization, i.e. signal settings of varying kind have been optimized for a given demand. BRAUN (2008) and BRAUN ET AL. (2008a, 2008b) respectively were among the first to report on implementation of GA into a real ATCS. They replaced the hill-climbing algorithm previously used in BALANCE by a GA to optimize signal settings in real-time. This modified version of BALANCE has been tested in a real network in Ingolstadt, Germany, with 46 signalized intersections which have been grouped into three different sub-networks. The authors stated that delays could be further reduced by an average of 10 percent over the whole day compared to the original BALANCE, as a field test with floating car data and vehicle re-identification revealed. Compared to the original vehicle-actuated control strategy used in Ingolstadt the reduction amounted to 21 percent. Even though the algorithm had been designed to optimize cycle length, green splits, phase sequences and offsets, the latter has been excluded from optimization during the field trials due to the assumption that major changes of the coordination pattern from one to the next optimization interval might result in significant disturbances during transition. BRAUN (2008) suggests further research on the inclusion of signal plan transition into the optimization process in order to prevent the GA from finding offset combinations that require unfavorable transition periods at the beginning of the next optimization interval.

MOTION, too, has been enhanced by a GA (MÜCK, 2008a, 2008b). It is thus the second ATCS that incorporates this optimization technique. In contrast to the approach by BRAUN (2008) only phase sequences and offsets are optimized by the GA, whereas cycle length and green splits are optimized before in the way described in paragraph 2.5.2.5. For evaluation of different offsets and phase sequences a link-based mesoscopic traffic flow model is used to estimate delays and stops which are transformed into a weighted performance index. Modeled vehicles arriving at a red signal are stored in a vertical queue, i.e. the physical queue length is neglected. MÜCK (2008a, 2008b) states that the model takes about 50 milliseconds for a single run. However, no details on the modeled network are mentioned. The runtime of the model presumably refers to the arterial test network in Münster, Germany, with 24 signalized intersections which is presented later in the papers. The optimization of offsets can be constrained to values between pre-specified upper and lower limits in order to avoid major changes of signal settings from one to the next optimization interval.

Field tests in Münster revealed some problems arising from the GA. Since the GA searches for a system optimum, it does not always implement a well functioning progressive signal system along the whole stretch of the arterial. However, such “green waves” are often asked for by the authorities that are in charge of the city’s traffic signal systems. In general manually planned coordination favors traffic flows to the city center during the morning hours and to the suburbs in the afternoon. This strategy might not correspond to the real system optimum but is expected by drivers along the arterial.

The second problem that has been observed has its origin in the highly irregular solution space that contains a lot of local optima of similar performance. Therefore the GA tends to change signal settings significantly between optimization intervals despite the aforementioned constraints. This lead to the unwanted negative transition effects described by BRAUN (2008). Because of these observations a second, deterministic approach for offset optimization has been developed by MÜCK (2008a, 2008b) that will be offered as an alternative to the GA in MOTION. No details on this approach are available yet.

2.7 Cell Transmission Model in traffic signal control

The CTM is a macroscopic time and space discrete traffic flow model that can be used to model highway and urban traffic flow subject to signal control. A detailed presentation of the model will be given in chapter 4.

This thesis is not the first research to make use of the CTM for optimization of traffic signal control. The next paragraphs will give an overview of relevant previous work. Most researchers combined the CTM with a GA where the CTM was used to evaluate the effects of different possible solutions, i.e. different signal settings.

2.7.1 Optimization of signal timings

One of the first to use the CTM for modeling signalized intersections was LO (1999, 2001) who emphasized that the CTM can also model oversaturated conditions. This case is in general not covered by other offline optimization tools. He transformed the model equations of the CTM into a mixed integer programming approach to minimize total network delay. He optimized durations of signal phases and applied the method to an artificial, simple network with two intersections and only one-way streets without any turnings. Even though the method worked well, the solution time was very long, impeding its application on larger networks. LIN/WANG (2004) extended the approach by including the number of stops into optimization which they derived from an approximate formula. They also included optimization of cycle length. The same small test network has been used.

Due to the long solution time mentioned above, LO ET AL. (2001) tried to use the CTM as fitness function of a GA, evaluating the performance of signal settings by repeated CTM simulation runs. They optimized green splits for a real one-way street network in Hong Kong with three intersections, this time including turning movements with fixed turning percentages. They compared the resulting signal timings to those produced by TRANSYT by simulating both with the CTM and concluded that the performance of their method was better. LO ET AL. (2004) further extended the method by including different options for optimization which allow for different combinations of fixed and variable cycle lengths and green splits over several cycles. The bene-

fits of entirely variable signal settings turned out to be more relevant during periods of high congestion. The authors also concluded that even though the duration of an optimization run could be reduced compared to the mixed integer programming technique, the approach was still not ready to be used in a real-time system. None of the papers by LO ET AL. reports on an assessment of the finally optimized signal plans by unbiased tools such as microsimulation.

The fundamental work of ALMASRI (2006) has been the initial starting point for this thesis. He used the CTM to assess the fitness of different offset combinations of signalized intersections. Cycle length, phase durations and phase sequences have been kept fixed. He, too, included the CTM as fitness function into a GA framework. He proposed a parallel GA (PGA) that optimizes all offsets in a network simultaneously, and a serial GA (SGA) that optimizes the offsets of groups of intersections consecutively. The groups are formed according to the heaviest loaded routes which are derived from network inflows and turning percentages at intersections. The first group covers all intersections along the heaviest loaded route, the second group all remaining intersections along the second heaviest loaded route and so on until all intersections have been considered. ALMASRI (2006) used different artificial and real test networks with three to twelve intersections and evaluated the GA based optimized offsets by means of microsimulation with AIMSUN. Reference cases have been developed using other optimization techniques including TRANSYT, but they were all outperformed by both GA variations. SGA outperformed PGA both in reduction of delay and computation time.

Based on the CTM and GA ZHANG ET AL. (2010) developed a method to design robust fixed-time signal settings along urban arterials under demand uncertainty. The aim was to design optimized fixed time signal plans in such a way that they perform well under a variety of different traffic demand patterns. The criterion for optimization was total system delay. To incorporate variations of traffic demand different demand scenarios with varying inflows at the entries of the network have been defined with given probabilities of occurrence. The GA has been designed in such a way that it tries to find a signal plan that works fine for a pre-specified percentage α of possible demand outcomes. The probability α can be interpreted as a level of confidence of a signal plan. Each feasible signal plan that is evaluated during a GA run is evaluated by several simulation runs of the CTM, one for each defined demand scenario. Based on the probabilities of occurrence of each scenario, a total delay is computed that will not be exceeded with probability α. This total delay is used as fitness value of the respective signal plan. ZHANG ET AL. (2010) tested the algorithm in an artificial arterial network with five signalized intersections and compared the results to nominal signal plans optimized for an average demand. Comparison has been done by CORSIM based microsimulation. 20 traffic demands generated randomly subject to the probabilities of occurrence of different traffic demands have been used. The robust signal plans produced less delay than the nominal signal plans on average.

2.7.2 Enhancements of the CTM

FELDMAN/MAHER (2002a, 2002b) applied the CTM in combination with GA to optimize offsets at signalized roundabouts. They wanted to make use of physical queue lengths which are explicitly modeled by the CTM and which must not be neglected during optimization, especially if vehicle storage capacity of links is limited which is the case at roundabouts. Since the original CTM does not incorporate platoon dispersion, they tested different approximations of the fundamental diagram besides the one used in the original CTM (cf. paragraph 4.2.1) to overcome this defi-

ciency. The approach enabled modeling different degrees of platoon dispersion. They finally applied their method to a signalized roundabout with four approaches and found that the GA performed slightly better than the hill-climbing algorithm.

ROHDE ET AL. (2008) built on the work by ALMASRI (2006) and developed a planning tool that can be used for offline optimization of offsets in street network for a given traffic demand. They introduced concepts to model arbitrary intersection layouts and permitted left-turns.

LI/CHANG (2010) also developed a GA-based optimization of different signal settings in arterial networks. They focused on an enhancement of the CTM to cover the impacts of lane blockage on adjacent lanes, i.e. blocked left-turn lanes impeding through traffic on adjacent lanes and vice versa. However, the original CTM already provides model equations for diverging lanes. These equations cover the effect of blocked lanes. Consequently, the problem addressed by LI/CHANG (2010) does not arise in the opinion of the author of this thesis if all lanes are modeled as individual links.

Another enhancement of the CTM in the context of signal control has been proposed by CHOW ET AL. (2010) who simulated traffic-actuated controllers. The microscopic data that is needed for actuated-control, i.e. vehicles headways and time gaps, is derived from the macroscopically modeled traffic flows of the CTM by assumption of exponentially distributed inter-vehicle spacing following the Poisson distribution. The authors conclude that travel times in networks with vehicle-actuated controllers can be estimated reasonably well by this approach.

2.8 Remaining need for research and methodology

As described in paragraph 2.6, GA have not been widely used in the framework of ATCS. Until today, only BALANCE and MOTION make use of GA for online optimization of traffic signal settings. Paragraph 2.7 further showed that none of the cited publications reports on online applications of GA in combination with the CTM. To the knowledge of the author of this thesis it has never been tried to create a new ATCS prototype which optimizes signal settings based on a GA and the CTM. MÜCK (2008a, 2008b) refers to FRIEDRICH/ALMASRI (2005) to substantiate his doubts that the CTM could be used for online application due to the reported long runtime of the model.

One of the aims of this thesis was to show that the CTM can be used for real-time application, given a sufficiently fast implementation of the model. To achieve this aim, an ATCS prototype which adopts a similar philosophy as BALANCE and MOTION has been conceived. Its conceptual design will be presented in chapter 3. The CTM and its implementation which include some enhancements that are partly and loosely based on ROHDE ET AL. (2008) will be discussed in chapter 4. It will be shown that this version of the CTM is sufficiently fast for online application.

In all of the above-mentioned research studies on application of the CTM for traffic signal control, optimization was performed based on a given demand. This is perfectly acceptable for offline optimization of signal settings. In the framework of an ATCS, however, the current demand that is used for optimization must be estimated. The ATCS that have been described in paragraph 2.5.2 apply different techniques to estimate the current traffic demand which can then be used for optimization. Such a technique also has to be incorporated into the new ATCS prototype. To do so, recent research conducted by WANG (2008) has been adopted. The approach is not equal but similar to the OD matrix estimation used in BALANCE. However, recent

findings of WANG (2008) help to improve the quality of OD matrix estimation. These findings have been analyzed, slightly altered and implemented in this work. A detailed presentation of the method and the alterations will be given in chapter 5.

BALANCE and MOTION optimize signal settings for time intervals of 5 to 15 minutes. BALANCE does not use any forecasting technique. The estimation of the upcoming traffic demand is merely based on detector counts of the last time interval. Even though MOTION has the capability to forecast detector counts, to the knowledge of the author of this thesis this forecasting is not in use. Therefore, a second question that has been dealt with in this work in the context of demand estimation was whether a suitable technique to forecast detector counts can improve the estimation of the upcoming traffic demand and thus the performance of the ATCS prototype. An according forecasting technique and its investigation will be described in chapter 5, too.

A simple approach to adapt cycle length and green splits to the currently estimated traffic demand will be presented in chapter 6. The main focus of the optimization has been set to the model-based optimization of offsets as proposed by ALMASRI (2006). This part of optimization will also be described in chapter 6. Both PGA and SGA have been implemented, along with a third deterministic alternative. SGA and the deterministic algorithm need to know the heaviest loaded routes as input data. In contrast to ALMASRI (2006), who derived this information from fixed input flows and turning portions, the results from the afore-mentioned demand estimation have been used in this work. In comparison with the PGA, which does not need information on traffic volumes on routes, it could then be evaluated whether inclusion of such information can improve the quality of optimization as stated by ALMASRI (2006).

An obviously significant problem of GA optimization that has been assumed by BRAUN (2008) and described by MÜCK (2008a, 2008b) is signal plan transition. Both authors state that the random behavior of GA can result in major changes of signal settings from one optimization interval to the next. This implies the need for signal plan transition and may therefore induce major disturbances of traffic flow. Even though MOTION is reported to use a smooth transition technique, no details on this technique are available. Therefore, this thesis also includes a detailed evaluation of different transition techniques and presents a possible way to include the effects of transition into the optimization of signal settings. Both aspects will be described in chapter 7.

Previous studies that used the CTM for optimization of signal settings either did not use microsimulation at all to evaluate the performance of the optimized signal plans, or only a single time interval with the same demand as the one used for optimization was simulated. Since an ATCS prototype is developed in this thesis which is intended to adapt signal settings to changing traffic demands, a much more comprehensive simulation study had to be performed to evaluate the performance of the ATCS. The setup of the simulation is presented in chapter 3. On the one hand, the simulation is used to assess the different modules of the ATCS. On the other hand, chapter 8 presents the results of a detailed evaluation of the entire ATCS prototype based on this simulation. These results will reveal the potential and also some weaknesses of the ATCS prototype. The derived information can be generalized to some extent and may therefore also be relevant to other ATCS.

3 Fundamentals of this thesis

3.1 Overview

This chapter provides some fundamentals which are relevant for the remaining parts of this thesis. Paragraph 3.2 gives an overview of the conceptual design of the developed ATCS prototype and its modules. Paragraph 3.3 presents the setup of a simulation study conducted with the microsimulator AIMSUN. This simulation study will be used to assess the performance of the different modules of the ATCS prototype in the following chapters and of the entire prototype as a whole. Finally, paragraph 3.4 describes the statistical indicators and the statistical test which are used in this thesis to evaluate the performance of different modules and of the overall ATCS prototype.

3.2 Conceptual design of the new ATCS prototype

3.2.1 Basic principles

The following overview of the general structure of the developed ATCS prototype and its different modules enables a basic understanding of how the overall system works (cf. figure 3-1). The following chapters will explain the modules in detail. The ATCS prototype has been implemented as an object-oriented program written in Java.

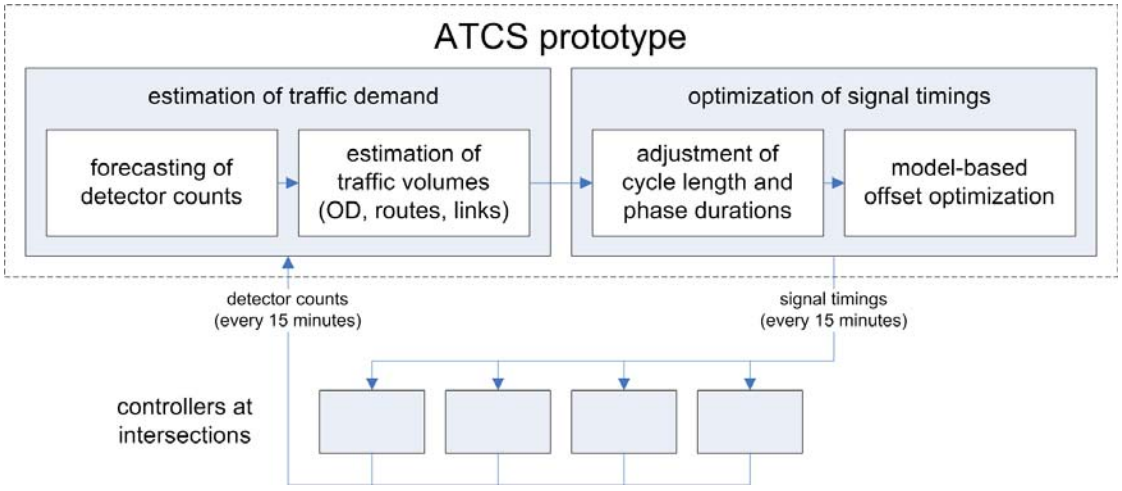


Figure 3-1: Conceptual design of the ATCS prototype

The prototype of the ATCS is designed for use in urban sub-networks containing several signalized intersections. It can explicitly handle not only mere arterials but also arbitrary networks with intersecting or even meshed roads and competing traffic streams. Based on an estimation of the upcoming traffic demand, the strategy optimizes a network-wide common cycle length, phase durations and offsets. The signal groups of each phase, phase sequences and phase changes are fixed and have to be configured offline. Optimization is performed for consecutive time intervals of 15 minutes, i.e. every 15 minutes a new optimization process is initiated. The duration of 15 minutes has been chosen for several reasons. On the one hand, the ATCS prototype must be flexible enough to react quickly to changing traffic demand patterns. Therefore, the optimization intervals must not be too long. On the other hand, each change of signal timings also requires a certain period of signal plan transition that can take several minutes and may induce some disturbances in the network. If the duration of the optimization interval is very short, the network might end up in a state of constant transition and the desired coordination patterns are almost never in operation. A duration of 15 minutes has therefore been regarded as a good compromise.

The ATCS prototype operates on a strategic, i.e. central level. The local controllers at the intersections send vehicle counts from their detectors to the central computer where the ATCS prototype is executed. Detector counts are sent at regular intervals or, more precisely, at the end of each interval of 15 minutes. The ATCS uses this data for traffic demand forecasting and estimation. Based on this estimation, the aforementioned signal timings are adjusted or optimized in order to minimize total network delay. After some computational time, the new signal timings are sent back to the controllers where they will be used from the beginning of the next time interval. The new signal settings include switching sequences for the period of transition in order to transform the previous signal plans to the new ones.

The centralized concept which optimizes signal timings for the entire network under the control of the ATCS prototype for consecutive optimization intervals is comparable to the basic concepts of BALANCE and MOTION.

3.2.2 Demand estimation

A preferably precise estimation of the current or forthcoming traffic demand in a network is crucial for optimization of signal plans. The optimization algorithm needs the traffic demand that the signal settings shall be optimized for as input data. For demand estimation the network is represented internally as a directed graph with separate links not only for sections connecting the intersections but also for all internal turning movements at all intersections (i.e. intersections are not represented as single nodes only). Current or expected traffic volumes on all alternative and reasonable routes in the network and, especially with regard to the traffic model, i.e. the CTM, traffic volumes on all links in the sub-network are needed.

The task of traffic demand estimation is divided into two major steps. In a first step the expected detector counts of the next optimization interval have to be forecasted. In this thesis, detector counts at signalized intersections are assumed to be the only available information in order to estimate traffic demand. Special detector locations are not required. However, preferable detector locations are those that allow measuring turning flows at intersections directly if possible, i.e. if these turns have a separate lane. More details on that issue will be given later. In any case, at

least some of the links in the network have to be equipped with detectors measuring the flow on these links.

The forecasting module is based on an approach by FÖRSTER (2008). It uses so-called space-time-patterns of detector counts. When a new time interval starts, all available detector counts from the previous four time intervals are used to forecast the detector counts of the next optimization interval. The optimization starts right after the previous interval has ended. Since the optimization consumes some computational time in the range of several minutes during which a portion of the next time interval elapses simultaneously the optimization has to be done for the subsequent time interval (cf. figure 3-2). Therefore, based on the counts of the four preceding time intervals, the forecasting module has to look two intervals ahead instead of only one.

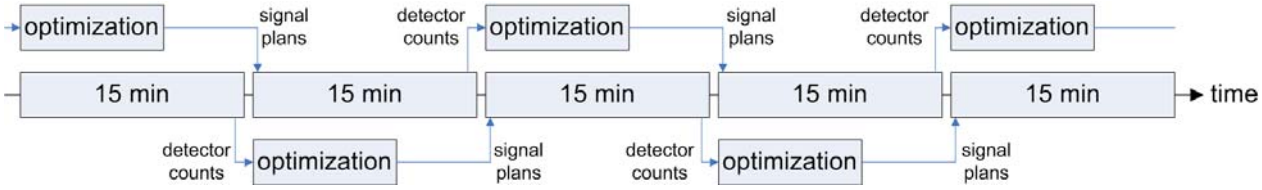


Figure 3-2: Time flow of optimization

Once counts for all links equipped with a detector are forecasted, another module that combines OD matrix estimation and traffic assignment uses these counts as constraints in a second step. The module is based on the work by WANG (2008) and FRIEDRICH/WANG (2006, 2008) respectively who base their research on VAN ZUYLEN/WILLUMSEN (1980), i.e. the IM model is used. While the method is originally designed to estimate OD matrices, it also allows deriving traffic volumes on routes and links as needed in the framework of the ATCS prototype presented in this thesis. The method has been adopted and enhanced with regard to further modification concerning consistency and improvement of constraints. Some rules proposed by WANG (2008) to eliminate redundant constraints have been generalized as well. The method results in consistent traffic volumes on all links in the network, and thus, an estimate of all source inflows into the network and of all turning percentages at all intersections during the next optimization interval is available and can be used as the desired input data for the optimization algorithms. The traffic demand during each interval is assumed to be static.

3.2.3 Optimization of signal timings

The next module adapts the cycle length and phase durations to the forecasted traffic demand using a simple non-model-based approach. This approach is based on the German HBS (FGSV, 2001) and RiLSA (FGSV, 2010). The module determines a common cycle length for all intersections. This cycle length depends on the most heavily loaded intersection.

The module for offset optimization is more sophisticated. It is based on the work by ALMASRI (2006). The optimization process performs a model-based offset optimization in order to identify the currently best possible coordination pattern. The CTM in combination with two different GA variations is applied. A third, deterministic optimization algorithm has been developed as well for comparison. Great importance has also been attached to the effects of signal plan transition which are also considered in the module.

The resulting signal plans are used as continuously updated fixed time signal plans. They are sent to the controllers at the end of each time interval or at the beginning of the next interval respectively. It is also conceivable to use the fixed time signal plans as framework signal plans in vehicle-actuated controllers. However, this option has not been implemented in this thesis.

3.3 Setup of simulation environment

3.3.1 General information

An extensive microsimulation study has been conducted to test the modules and the overall ATCS. A test network has been modeled in AIMSUN NG 5.1.8, a sophisticated microsimulation software distributed by Transport Simulation Systems (TSS, 2006).

AIMSUN has been used for different purposes in this study. First, it served as reference to validate the CTM (cf. paragraph 4.6). Second, artificial detector data as well as route and link volumes from different simulation runs have been logged to test the forecasting and traffic demand estimation modules (see paragraphs 5.2.2, 5.3.9 and 5.4). And third, AIMSUN and the developed ATCS have been connected via an Application Programming Interface (API) in C++ provided by AIMSUN (TSS, 2007) to test the new ATCS in a realistic simulation environment (cf. chapter 8).

In anticipation of detailed descriptions of the modules that comprise the ATCS prototype, the setup of the simulation study will be described first in the following paragraphs. On the one hand, an example network helps to better understand the correct application of the modules. On the other hand, the simulation and the test network will be referred to when the CTM and the forecasting and demand estimation modules are presented. Therefore, the test network and the creation of artificial test data are introduced first.

3.3.2 Specifics of the network

The chosen test area is an urban sub-network in the List district of Hanover, Germany, with eight signalized intersections, two pedestrian lights and two non-signalized intersections. The AIMSUN model of the network is shown in figure 3-3. The numbers in the colored circles correspond to the numbers of traffic signal systems as assigned by the authority in Hanover which is in charge of signalization. A total of 55 detectors are located on the lanes in front of traffic signals. However, not all turning flows are detected directly because of mixed lanes. At intersection 508 the right turning traffic coming from the south is excluded from signalization (free right turn) and thus not detected at all. At some of the signalized intersections, left turns from certain directions are restricted.

The network is not far from the city center in the southwest and connects it to suburban districts. It is thus important for commuters, but also for other traffic. All streets of the network have one lane per direction with some additional turning lanes at signalized intersections. Most of the streets have a speed limit of 50 km/h except for the area bordered by intersections 509, 510 and 520 where only 30 km/h are allowed. On the road passing intersections 520, 523, 508 and 505 public transport buses operate in both directions. The tramway travels entirely underground in the area.

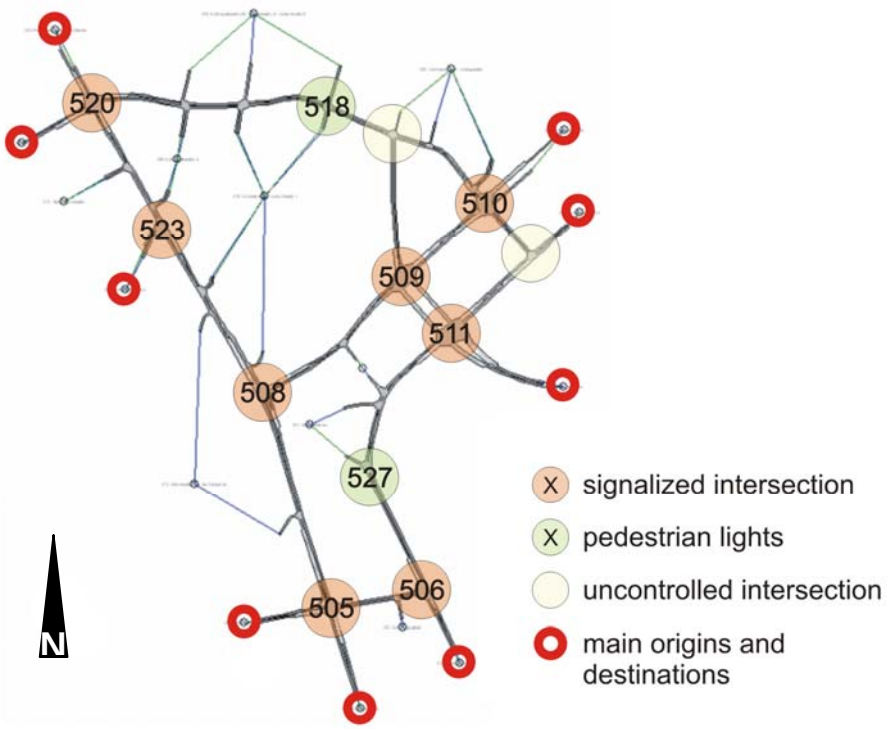


Figure 3-3: Test network in Hanover List district

3.3.3 Traffic demand

The traffic demand that was used in the study is semi-fictitious. It has been deduced from real measurements on-site that have been taken on July 8, 2008. These measurements include loop detector counts over the whole day, aggregated in 5-minute-intervals. The counts have been kindly provided by the city of Hanover. Furthermore data from six radar detectors that have been additionally installed at strategic positions on that day was available. The data from the two types of detectors was not entirely consistent due to measuring errors and sporadic failure of single loops. At intersection 508, which happens to be the most heavily loaded intersection in the sub-network, no detector data had been logged at all due to malfunction. And, as has been said, for some turnings no data is available at all. Therefore, some assumptions had to be made to derive the desired traffic demand for the whole network.

AIMSUN enables two different ways of defining traffic demand. The first way is to define source inflows [veh/h] and turning percentages. This type of demand has been used for validation of the CTM. Two traffic demands for the morning and afternoon peak interval of 15 minutes have been defined based on the available detector counts. Inconsistent counts have been balanced manually. Missing information on several turnings has been added by assumption. Only cars and trucks have been considered.

The second type of traffic demand definition requires input of OD matrices. It can also be specified how vehicles that travel between a specific OD pair are distributed among the available routes of that relation. This can be done either automatically by AIMSUN or manually by the user. This second type of demand definition has been used for a simulation that comprises the whole time period between 6 am and 8 pm on a regular weekday, i.e. 14 hours are simulated in

a row. Separate OD matrices have been defined for each 15-minute-interval, comprising a total of 56 matrices. Within this time period two peak hours from 8 to 9 am and from 5 to 6 pm occur with periods of lower traffic demand before, between and after those peak hours. The chosen demand reflects changes of traffic demand over time that the ATCS prototype is supposed to react to. Again, only cars and trucks were included.

The available data collected on-site contained only traffic flows at detector locations. Therefore, the data had to be transformed into OD matrices. 108 plausible alternative routes between 71 reasonable OD relations connecting 9 origins and destinations have been selected. These origins and destinations are highlighted in figure 3-3 by red circles. It was planned initially to define also traffic starting and ending at minor side streets. However, these trips could not be estimated from the available data. In order not to inflate complexity of the simulation, only the main origins and destinations were used in the end.

The number of available routes between different OD pairs varies between one and four, depending on the relation. Only reasonable alternatives have been considered, i.e. possible routes representing major detours have been excluded. The route choice probabilities were set manually in such a way that they appeared reasonable and at the same time reproduced the real measurements at detector locations as well as possible when assigning the OD matrices to the network accordingly.

To facilitate the process of OD matrix generation, only one matrix for the morning peak interval of 15 minutes and one for the afternoon peak interval of the same duration have been deduced from the measurements in the first place. The OD matrices for all other intervals are linear combinations of these two peak interval matrices and reflect the overall profile of traffic demand over time as observed on-site in an acceptable way. Figure 3-4 shows the profile of the overall traffic demand that has been finally used for simulation. It displays the average number of vehicles entering the network during each of the 15-minute-intervals. The curve has been derived from 30 simulation runs in AIMSUN with different random seed numbers and shows the average demand at each interval.

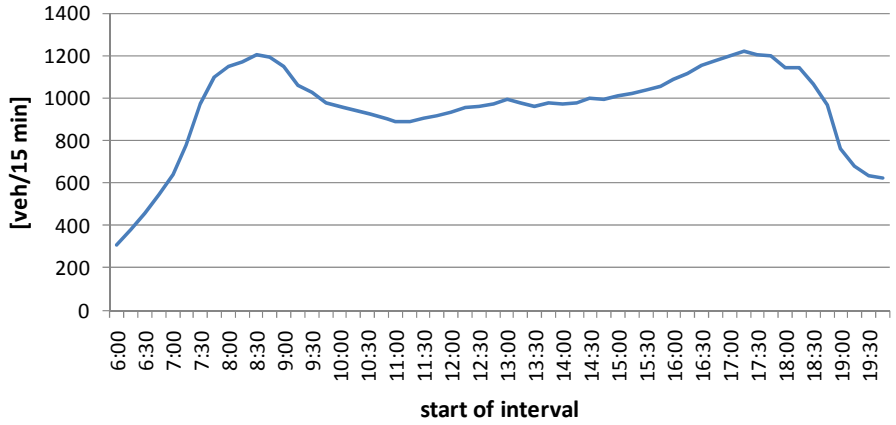


Figure 3-4: Overall traffic demand used in the simulation study

The resulting main streams during morning peak hour are those going to the city center (i.e. relations North-South and East-South) whereas traffic during the afternoon peak hour rather tends to leave the city center. However, the respective opposite directions also have traffic volumes at comparable level during both peak hours.

3.3.4 Signalization

Traffic signal control was based on the real fixed time signal plans in the first two cases, where the simulation either serves as reference for the CTM (peak intervals only) or is used for logging data to test the forecasting and demand estimation techniques (whole day). The fixed time signal plans were made available by the respective authority. It has to be highlighted that the real signal system in the test area operates in vehicle-actuated mode. The fixed time signal plans are used as fallback plans in case of detector failures or other types of malfunction. However, in the study only fixed time control has been used with some minor adaptations of the original signal plans. These adaptations were necessary because of the semi-fictitious traffic demand that obviously does not exactly correspond to the real demand on-site in all details.

The original signalization comprises only two different signal plans at each intersection for the time period covered by the simulation. The morning signal plans operate from 6 am to 1 pm and the afternoon signal plans from 1 pm to 8 pm. This concept has been adopted in this study. All signal plans have a cycle length of 90 seconds.

In order to test the new ATCS prototype, the developed system must be able to communicate directly with AIMSUN. The API of AIMSUN in C++ allows the user to implement and test own control strategies. Among other things it offers functionalities to read detector data at regular intervals and to directly change the current state of any signal group. Therefore, the prototype could be integrated into the simulation environment. A special module has been programmed in C++ that handles communication between the main application of the new ATCS, which is written in Java as mentioned before, and AIMSUN. Therefore, signalization in this simulation depends entirely on the ATCS prototype.

3.4 Statistical evaluation

3.4.1 Assessment of the quality of estimation

In this work, the quality of some estimated performance indicators, for instance total delay and travel times, which are modeled by the CTM has to be assessed by comparing the modeled to the “real” values derived from AIMSUN. Furthermore, the quality of estimation of the two modules for demand estimation, i.e. detector count forecasting and estimation of OD flows, route and link volumes has to be evaluated in a similar way. Three different statistical indicators are used in this thesis for this purpose. They are the correlation coefficient r_{xy}, the root mean square error (*RMSE*), and the relative root mean square error (*RRMSE*).

The correlation coefficient r_{xy} is a statistical dimensionless indicator for linear relationship between the paired values of two sets of data with n values each. It is determined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \qquad (3\text{-}1)$$

where:

x_i real value [arbitrary unit]

$\bar{x}$ arithmetic mean of the real values [arbitrary unit]

y_i estimate of the real value x_i [arbitrary unit]

$\bar{y}$ arithmetic mean of the estimated values [arbitrary unit]

n number of pairs of real and estimated values [-]

The correlation coefficient can take values between -1.0 and 1.0. These two extreme values indicate a perfect negative or positive linear relationship. A correlation coefficient of 0.0 means that there is no linear relationship between the two sets of values at all. In the context of this thesis, a correlation coefficient of 1.0 would correspond to a perfect estimation without any errors.

The RMSE describes the overall deviation between the estimated and the real values. It has the same unit as the compared values. The calculation rule is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}{n}} \qquad (3\text{-}2)$$

The smaller the *RMSE*, the better is the quality of estimation. An error-free estimation would results in an *RMSE* of 0. Since the difference between real and estimated values is squared, larger single deviations have a higher influence on the RMSE.

The *RRMSE* considers relative instead of absolute estimation errors and is thus dimensionless. These relative errors are weighted by the real values x_i. The RRMSE is determined by the following equation.

$$RRMSE = \sqrt{\frac{\sum_{i=1}^{n} x_i \cdot \left(\frac{x_i - y_i}{x_i}\right)^2}{\sum_{i=1}^{n} x_i}} \qquad (3\text{-}3)$$

Again, smaller values indicate a better quality of the estimation.

3.4.2 Comparison of different control strategies

For evaluation of the final ATCS prototype, multiple AIMSUN simulation runs have been conducted for different control strategies and parameter settings of the prototype. This evaluation will be described in chapter 8. In order to asses whether differences between the results of theses scenarios are significant or not, a Student's t-test has been applied as described, for instance, by HERZ ET AL. (1976).

Different performance indicators such as vehicle delays and number of stops are derived from n AIMSUN runs. These indicators are then compared to the according values from m runs with different settings. (In general, $n = m$ with only some few exceptions as explained in chapter 8.) Based on the n and m runs respectively, the arithmetic means $\bar{x}$ and $\bar{y}$ of the performance indicators and the according standard deviations s_x and s_y can be derived. In a next step, a null hypothesis H_0 has to be formulated. In this thesis, H_0 conforms to the assumption that the expected values of both scenarios are the same ($\mu_x = \mu_y$), i.e. all AIMSUN runs belong to the same statistical population with no significant difference. The Student's t-test is then used to try to

reject H_0 in order to show that the n and m AIMSUN runs belong to different statistical populations and that the derived average performance indicators differ significantly from each other.

Based on $\bar{x}$, $\bar{y}$, s_x, and s_y a value $\hat{t}$ has to be determined:

$$\hat{t} = \sqrt{\frac{n \cdot m \cdot (n+m-2)}{(n+m) \cdot \left[(n-1) \cdot s_x^2 + (m-1) \cdot s_y^2\right]}} \cdot (\bar{x} - \bar{y}) \qquad (3\text{-}4)$$

where:

n	number of AIMSUN runs of the first scenario [-]
m	number of AIMSUN runs of the second scenario [-]
$\bar{x}$	arithmetic mean of the performance indicator of the first scenario [arbitrary unit]
$\bar{y}$	arithmetic mean of the performance indicator of the second scenario [arbitrary unit]
s_x	standard deviation of the performance indicator of the first scenario [arbitrary unit]
s_y	standard deviation of the performance indicator of the second scenario [arbitrary unit]

This value $\hat{t}$ has to be compared to a critical value t_{crit} from the Student's t-distribution. This distribution approximates the normal distribution for large sample sizes. Since it cannot be said in advance whether potential differences between the average performance indicators will be positive or negative, a two-tailed t-test has to be used. The critical value $t_{crit} = t(1-\alpha/2, n+m-2)$ can be taken from according tables or calculated directly by means of a spreadsheet program. It depends on the degree of freedom, which is derived from the sample sizes n and m, and on the significance level α. The latter is the probability of mistakenly rejecting H_0. In this work a significance level of $\alpha = 0.05$ is chosen. H_0 is rejected if $|t| > t_{crit}$. In this case, differences of the average performance indicators are statistically significant.

4 Macroscopic traffic flow model

4.1 Overview

Building on the approach by ALMASRI (2006) the CTM is used in this work to model traffic in urban networks. DAGANZO (1994, 1995) initially proposed the CTM for highway applications. However, it has been successfully applied to urban networks in several research projects (cf. paragraph 2.7). While DAGANZO (1994, 1995) and ALMASRI (2006) explain the model in detail, further insights are also presented in the next paragraphs for the sake of completeness. Moreover, these insights provide a basis for some extensions and adaptations of the model that will be presented afterwards, followed by some explanations on how to derive total delay and travel times from the model. The last two subsections will present the implementation and validation of the model.

4.2 Original Cell Transmission Model

This section describes the CTM as presented by DAGANZO (1994, 1995) and ALMASRI (2006). Some variable names have been altered in order to comply with the notation used in this thesis.

4.2.1 Basic model characteristics

The CTM is a space- and time-discrete variation of the well-known Lighthill-Whitham-Richards (LWR) model (cf. LIGHTHILL/WITHAM, 1955, RICHARDS, 1956), a macroscopic traffic flow model derived by analogy to hydrodynamics. That is to say, the model does not consider individual vehicles the way a microscopic model would, but treats traffic flow as a fluid.

The CTM is based on a trapezoidal simplification of the fundamental diagram (cf. figure 4-1), which describes the relationship between traffic flow q [veh/s] and density k [veh/m] (note that all units are given on the basis of meters and seconds which is more suitable for the CTM than hours and kilometers). Some model parameters have to be set either globally or individually for each road or road section. The free flow speed v_f [m/s] applies to the traffic state where individual vehicles drive freely without being influenced by other vehicles. This state corresponds to the rising line in figure 4-1. Traffic flow q grows with increasing density (and vice versa), with v_f being the slope of the line. This growth is limited by the capacity Q [veh/s] of the road section. Once capacity is reached, the traffic flow stays at a constant level over a range of densities (horizontal line in figure 4-1). This corresponds to the state of turn-over between free flow and

forced or congested flow. Finally, the traffic flow decreases again with increasing density until jam density k_{jam} [veh/m] is reached and the road is totally blocked. The gradient $-w$ [m/s] of this falling line, which corresponds to the state of congested flow, is the backward wave speed. Disturbances in traffic flow propagate backward with this speed during congestion.

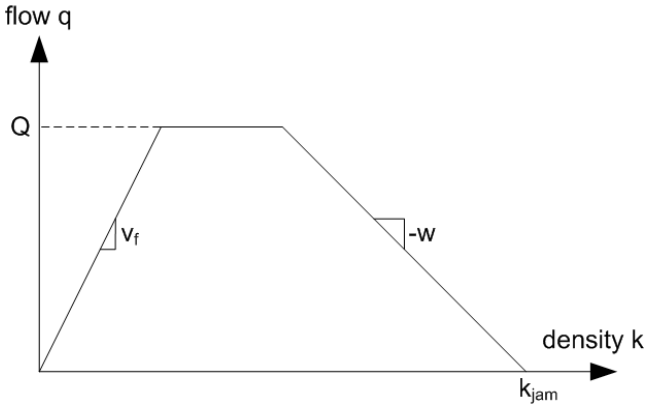


Figure 4-1: Simplified fundamental diagram

Since the CTM is a time-discrete model, all calculations are done repeatedly for consecutive time steps of constant duration T. Since it is also a space-discrete model, a road section is divided into a finite number of cells which can contain a limited number of vehicles. The length of a cell is the ratio of free flow speed v_f and duration T of a simulation step. This guarantees that no vehicle can pass more than one cell during one simulation step, which facilitates the necessary calculations. The maximum number N of vehicles a cell can contain is determined by the product of cell length and jam density.

4.2.2 Basic model equations

Figure 4-2 shows a series of connected cells. At the beginning of each time step t each cell i contains a specific real number $n_i(t)$ of vehicles that lies between zero and N_i. During each time step t a flow $q_i(t)$ of vehicles leaves cell i and another flow $q_{i-1}(t)$ of vehicles enters cell i. So, for each time step t two calculations have to be performed: first, all flows $q_i(t)$ between cells during time step t have to be determined, and second, the numbers $n_i(t+1)$ of vehicles in each cell at the beginning of the next time step $t+1$ have to be updated.

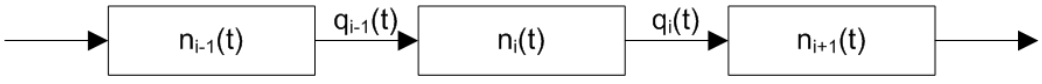


Figure 4-2: Series of connected cells

For the first step DAGANZO (1994) defined the following equation (note that from now on flow $q_i(t)$ is regarded as the number of vehicles during one time step and is thus given in vehicles instead of vehicles per second):

$$q_i(t) = \min\left\{ n_i(t), Q_i, \frac{w}{v_f} \cdot \left(N_{i+1} - n_{i+1}(t) \right) \right\} \qquad (4\text{-}1)$$

where:

$q_i(t)$ number of vehicles leaving cell i during time step t [veh]

Q_i	flow capacity of cell *i* [veh]
$n_i(t)$	number of vehicles in cell *i* at the beginning of time step *t* [veh]
N_i	holding capacity of cell *i* [veh]
w	backward wave speed [m/s]
v_f	free flow speed [m/s]

Thus, the flow $q_i(t)$ is the smallest of the three terms in the curly braces. Equation 4-1 reflects the trapezoidal fundamental diagram displayed in figure 4-1. Under free flow conditions all vehicles that are in cell *i* at the beginning of time step *t* can leave the cell during the time step (first term). However, if the flow exceeds the cells outflow capacity Q_i, it is limited to the latter (second term). Under congested conditions, the third term applies and the outflow $q_i(t)$ is limited to the available space in the following cell *i*+1. This available space is expressed by the term in round brackets, which is the maximum capacity N_{i+1} of vehicles in cell *i*+1 minus the current number $n_{i+1}(t)$ of vehicles in that cell. The ratio of backward wave speed *w* and free flow speed v_f incorporates the desired speed of propagation of disturbances. If this ratio is left aside, propagation occurs at free flow speed. In this case the rising and falling lines in figure 4-1 for free flow and congested conditions have the same slope but with different algebraic signs.

When all flows $q_i(t)$ have been determined for the current time step *t*, the number of vehicles in each cell at the beginning of the next time step *t*+1 can be updated according to the following equation:

$$n_i(t+1) = n_i(t) + q_{i-1}(t) - q_i(t) \qquad (4\text{-}2)$$

Equation 4-2 assures flow conservation. The number of vehicles in cell *i* at the beginning of a time step simply equals the number of vehicles in that cell at the beginning of the previous time step plus all vehicles entering and minus all vehicles leaving the cell during the previous time step. If $n_i(t)$ is subtracted from both sides, equation 4-2 becomes a discretized form of the differential flow conservation equation of the hydrodynamic LWR model:

$$\frac{\delta k}{\delta t} = -\frac{\delta q}{\delta x} \qquad (4\text{-}3)$$

where:

k	density [veh/m]
q	traffic flow [veh/s]
t	time [s]
x	space [m]

Equation 4-3 states that the change of traffic density over time equals the negative change of traffic flow over space, which is quite evident when regarding the discretized form described by equation 4-2.

To further simplify calculations, equation (4-1) can be split into three equations (4-4 to 4-6).

$$S_i(t) = \min\{n_i(t), Q_i\} \qquad (4\text{-}4)$$

$$R_i(t) = \min\left\{Q_i, \frac{w}{v_f} \cdot \left(N_{i+1} - n_{i+1}(t)\right)\right\} \qquad (4\text{-}5)$$

$$q_i(t) = \min\left\{S_i(t), R_{i+1}(t)\right\} \quad (4\text{-}6)$$

where:

$S_i(t)$ maximum flow that can be sent by cell i during time interval t [veh]

$R_i(t)$ maximum flow that can be received by cell i during time interval t [veh]

Thus, the following steps have to be executed for each time step t:

Step 1: Calculate $S_i(t)$ and $R_i(t)$ for each cell i (equation 4-4 and 4-5).

Step 2: Determine the flows between all cells by choosing the smaller of both $S_i(t)$ of the respective preceding cell and $R_{i+1}(t)$ of the respective following cell (equation 4-6).

Step 3: Update the number $n_i(t+1)$ of vehicles in all cells i at the beginning of the next time step $t+1$ (equation 4-2).

4.2.3 Diverges

The discussion so far has only covered ordinary cells with one preceding and one following cell. In order to model network topologies other than mere unbranched road sections, DAGANZO (1994) proposed two special cell types, the first covering diverging cells. He only considered diverges from one into two cells (cf. figure 4-3) to model off-ramps on highways.

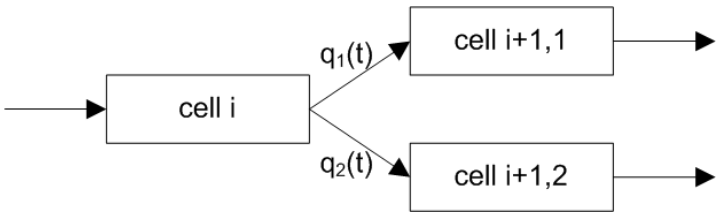


Figure 4-3: Diverging cell

Two flows $q_1(t)$ and $q_2(t)$ leaving cell i have to be determined in this case instead of only one single flow $q_i(t)$, subject to the conditions $q_i(t) = q_1(t) + q_2(t)$, $q_1(t) = q_i(t)\cdot\beta_1$ and $q_2(t) = q_i(t)\cdot\beta_2$. The factors β_1 and β_2 are fixed turning percentages which have to sum up to 1.

In a first step $S_i(t)$, $R_{i+1,1}(t)$ and $R_{i+1,2}(t)$ have to be calculated according to equations 4-4 and 4-5. After this, three cases have to be considered. The total flow $q_i(t)$ cannot be greater than $S_i(t)$, i.e. $q_i(t) \le S_i(t)$. Moreover, none of the two flows $q_1(t)$ and $q_2(t)$ can be greater than the respective $R_{i+1,1}(t)$ and $R_{i+1,2}(t)$, i.e. $q_i(t)\cdot\beta_1 \le R_{i+1,1}(t)$ and $q_i(t)\cdot\beta_2 \le R_{i+1,2}(t)$, or $q_i(t) \le R_{i+1,1}(t)/\beta_1$ and $q_i(t) \le R_{i+1,2}(t)/\beta_2$. This leads to the following three equations to derive $q_1(t)$ and $q_2(t)$:

$$q_i(t) = \min\left\{S_i(t), R_{i+1,1}(t)/\beta_1, R_{i+1,2}(t)/\beta_2\right\} \quad (4\text{-}7)$$

$$q_1(t) = q_i(t)\cdot\beta_1 \quad (4\text{-}8)$$

$$q_2(t) = q_i(t)\cdot\beta_2 \quad (4\text{-}9)$$

where:

$\beta_{1/2}$ turning percentages of flows 1 and 2 respectivley [-]

Equations 4-7 to 4-9 account for blockages of cells *i*+1,1 or *i*+1,2. If one of the two last terms in equation 4-7 prevails, this means that flow $q_1(t)$ or $q_2(t)$ will fill up the available space in the following cell completely. No further vehicles can thus leave cell i, regardless of whether they want to proceed to cell *i*+1,1 or *i*+1,2. A strict first-in-first-out scheme is applied to cell *i*.

4.2.4 Merges

DAGANZO (1994) proposed another cell type to model merges at on-ramps, again only connecting two cells to one cell (cf. figure 4-4).

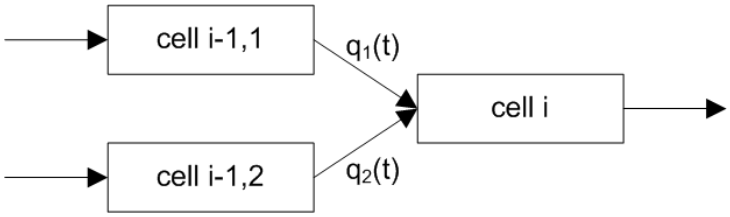


Figure 4-4: Merging cell

Contrary to the case of diverging, the case of merging is less trivial. As before, equations 4-4 and 4-5 have to be applied, this time to calculate $S_{i\text{-}1,1}(t)$, $S_{i\text{-}1,2}(t)$ and $R_i(t)$. Since two different cells try to release vehicles into cell *i* simultaneously, a sort of priority rule has to be expressed by introducing two fixed priority percentages p_1 and p_2 that sum up to 1. Their role will become evident in the next paragraphs.

The first case to consider is simple. If $S_{i\text{-}1,1}(t) + S_{i\text{-}1,2}(t) \leq R_i(t)$, i.e. the total of both sending volumes of the preceding cells does not exceed $R_i(t)$, each cell can send its entire volume $S_{i\text{-}1,1}(t)$ and $S_{i\text{-}1,2}(t)$ respectively to cell *i*. If however the total of both sending volumes exceeds $R_i(t)$, some more cases have to be differentiated (cf. figure 4-5). Since cell *i* can receive no more than $R_i(t)$ vehicles, the final solution, i.e. a feasible combination of $q_1(t)$ and $q_2(t)$, must inevitably lie on the line defined by the linear equation $q_2(t) = R_i(t) - q_1(t)$. By means of the two priority portions a second line can be defined so that $q_1(t):q_2(t) = p_1:p_2$. The according linear equation is $q_2(t) = p_2/p_1 \cdot q_1(t)$. For case (a) depicted in figure 4-5 the solution can be found at the intersection of both lines. It is $q_1(t) = p_1 \cdot R_i(t)$ and $q_2(t) = p_2 \cdot R_i(t)$. Both cells share $R_i(t)$ according to their priority portions.

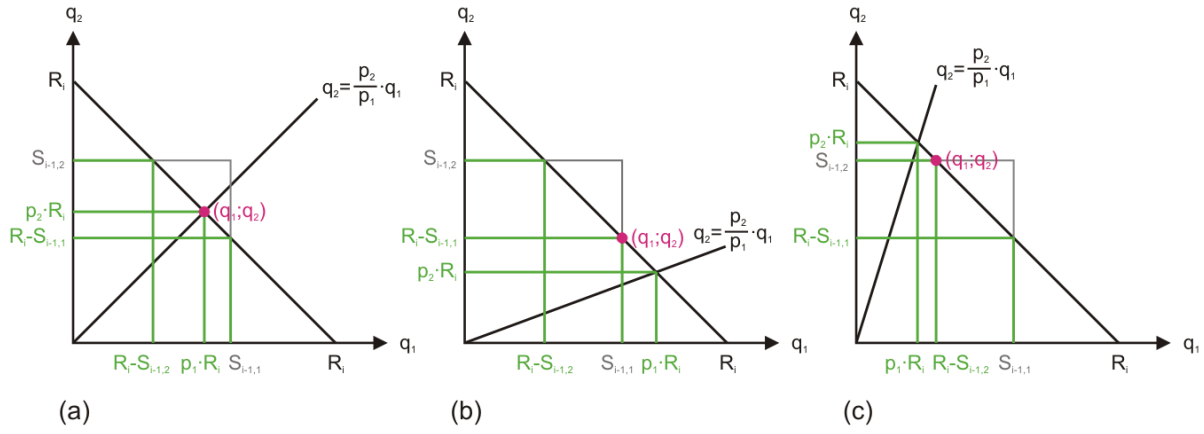


Figure 4-5: Three cases of merging

It is also evident that a feasible solution must lie within a rectangle defined by the two equations $q_1(t) = S_{i\text{-}1,1}(t)$ and $q_2(t) = S_{i\text{-}1,2}(t)$, because the flow originating from one cell cannot be greater than the respective maximum flow of that cell. In cases (b) and (c) the intersection of both lines lies outside of this rectangle, i.e. $p_1 \cdot R_i(t) > S_{i\text{-}1,1}(t)$ or $p_2 \cdot R_i(t) > S_{i\text{-}1,2}(t)$, and $q_1(t)$ has to be limited to $S_{i\text{-}1,1}(t)$ or $q_2(t)$ to $S_{i\text{-}1,2}(t)$. The remainder of $R_i(t)$ can be assigned to the other flow, i.e. $q_2(t) = R_i(t) - S_{i\text{-}1,1}(t)$ or $q_1(t) = R_i(t) - S_{i\text{-}1,2}(t)$.

A comparison of all three cases depicted in figure 4-5 reveals that the final solution always corresponds to the middle of the three values $S_{i\text{-}1,1}(t)$, $p_1 \cdot R_i(t)$ and $R_i(t) - S_{i\text{-}1,2}(t)$ for $q_1(t)$ and $S_{i\text{-}1,2}(t)$, $p_2 \cdot R_i(t)$ and $R_i(t) - S_{i\text{-}1,1}(t)$ for $q_2(t)$. Thus, all the aforementioned considerations can be converted into four simple equations:

$$q_1(t) = S_{i-1,1}(t) \qquad \text{if } S_{i-1,1}(t) + S_{i-1,2}(t) \leq R_i(t) \qquad (4\text{-}10)$$

$$q_2(t) = S_{i-1,2}(t) \qquad (4\text{-}11)$$

$$q_1(t) = \text{mid}\left\{S_{i-1,1}(t), p_1 \cdot R_i(t), R_i(t) - S_{i-1,2}(t)\right\} \qquad \text{otherwise} \qquad (4\text{-}12)$$

$$q_2(t) = \text{mid}\left\{S_{i-1,2}(t), p_2 \cdot R_i(t), R_i(t) - S_{i-1,1}(t)\right\} \qquad (4\text{-}13)$$

where:

$p_{1/2}$ priority percentages of flows 1 and 2 respectively [-]

4.2.5 Sources and sinks

DAGANZO (1994) introduced input and output cells to model sources and sinks where vehicles can enter and leave the network. For modeling sources he proposed to use a cell pair consisting of a source cell and a gate cell. The number $n_i(0)$ of vehicles in the source cell at time step $t = 0$ is set to infinity. The source cell is followed by an empty gate cell with a holding capacity N_i set to infinity and a flow capacity Q_i that equals the desired source inflow. The gate cell serves as a metering device that allows traffic to flow into the network at the desired rate while holding vehicles that are currently unable to enter due to spillbacks.

In this work this metering functionality has been combined into one single source cell. This cell also has an infinite holding capacity N_i, but the same capacity Q_i as the cells of the following link. Its inflow $q_{i\text{-}1}(t)$ is set to a constant value that corresponds to the desired source inflow and does not change over time. If vehicles can discharge freely, they will do so at the desired rate. If the following link is congested, vehicles will be held back in the source cell. They discharge at flow capacity rate Q_i as soon as congestion is over.

Output cells or sinks are modeled as proposed by DAGANZO (1994). They have the same flow capacity Q_i as the cells of the preceding link and an infinite holding capacity N_i which enables them to store any amount of vehicles that leave the network.

4.3 Extensions of the Cell Transmission Model

Some extensions of the CTM are necessary or at least desirable in order to better model signalized urban networks. These extensions include enhanced merging and diverging to model

complex urban intersections, modeling of traffic signals and consideration of permitted left-turns and entirely uncontrolled intersections.

4.3.1 Enhancement of merging and diverging

At signalized intersections it is advisable to model each turning lane separately. Based on the two merge and diverge cell types of the original CTM, diverges from one into two lanes and merges from two into one lane can be handled easily. However, diverges from one into three or more lanes and respective merges can only be dealt with by using intermediate cells (cf. figure 4-6). This approach has been used by ALMASRI (2006). However, it may lead to a distorted topology, and therefore a more generalized connectivity between cells is desirable.

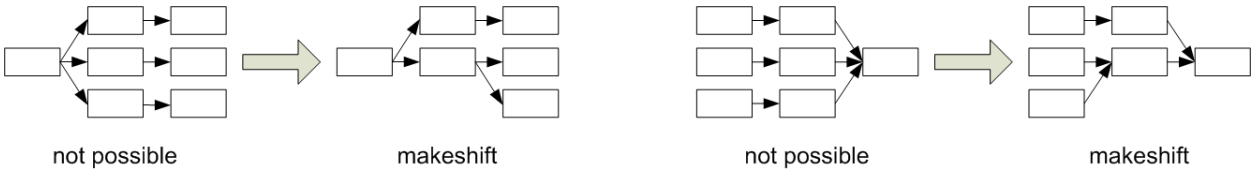


Figure 4-6: Intermediate cells to model arbitrary diverges and merges

FLÖTTERÖD/NAGEL (2005) presented an algorithm that allows connecting any number of predecessor cells to any number of successor cells (*n*:*n*-connectivity). In addition to cells they introduced connectors as auxiliary network elements. Each connector has 1 to *n* predecessor cells and 1 to *n* successor cells. It is responsible for flow calculations between these cells. To solve this task, a sub-process is executed for each connector at every time step *t*.

The basic idea of the approach by FLÖTTERÖD/NAGEL (2005) is to regard all available upstream vehicles $S_i(t)$ of all predecessor cells and the available spaces $R_i(t)$ in all successor cells as resources which are consumed at individual and constant rates during one time step *t*. The rates of consumption depend on the turning percentages β_i and the priority percentages p_i. Each time step *t* is sub-divided into several calculation steps *k* with varying duration so that at least one of the resources is exhausted at the end of each step *k*. The algorithm stops as soon as either all $S_i(t)$ or $R_i(t)$ or a combination of both are consumed entirely so that no further flow is possible. The total amounts of consumption of each $S_i(t)$ or $R_i(t)$ at the end of the last calculation step *k*, i.e. at the end of time step *t*, correspond to the final outflows of predecessor cells and inflows of successor cells during that time step. FLÖTTERÖD/NAGEL (2005) showed that for all three types of connectivity covered by the original CTM (1:1, 1:2, and 2:1) the algorithm produces the same flows as the original CTM. These types are thus special cases of the general algorithm.

Figure 4-7 illustrates two possible ways of how to model complex intersections if arbitrary connectivity is allowed. In case (a) 1:3-connections are used to model diverges into three lanes at the upper and lower approach. No intermediate cells are necessary. If an intersection requires diverges and merges with even more than three lanes, 1:*n*- and *n*:1-connections can be used. However, *n*:*n*-connections are avoided in case (a), i.e. all turning movements are modeled in detail even at the conflict area. An advantage of this type of modeling is its higher accuracy. First, it allows considering adapted free flow speeds v_f for different turning movements. Second, it does not neglect those fractions of travel time that occur while vehicles traverse the intersection. This is an important detail when it comes to optimization of offsets between intersections

based on the CTM. And third, explicit modeling of turning movements facilitates modeling of permitted left-turns (cf. paragraph 4.3.3). A major disadvantage of this type of modeling is the increased effort when building the model, which is crucial in the case of large networks that contain many different intersections.

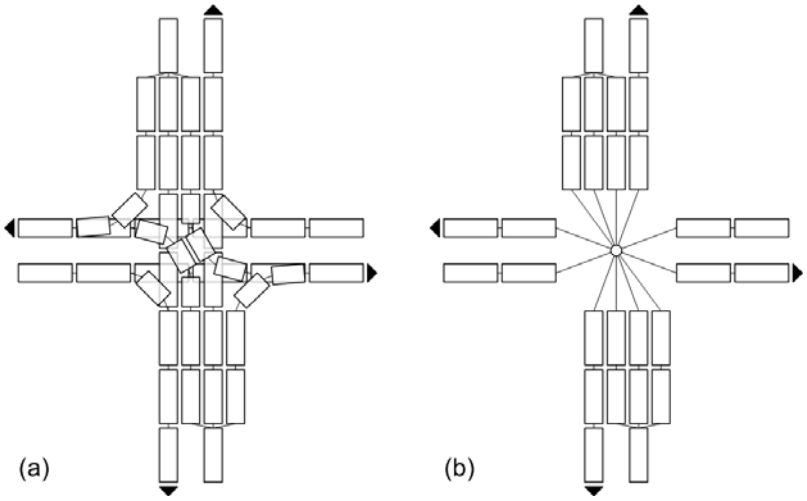


Figure 4-7: Detailed and simplified modeling of an intersection

Modeling type (b) does not only include 1:*n*- and *n*:1-connections for diverging and merging lanes. It also employs a general *n*:*n*-connector that deals with all traffic flows traversing the conflict area simultaneously. When creating models of larger networks the reduced complexity of intersections is very convenient. However, the disadvantages of this modeling type are obvious. Accuracy is reduced and important fractions of travel time are entirely neglected. Vehicles are "teleported" through the intersection rather than driving through it. And when considering permitted left-turns, some further problems arise (cf. paragraph 4.3.3). Therefore, modeling type (a) is applied in this work in order to avoid the important disadvantages of type (b) especially in view of offset optimization.

The algorithm proposed by FLÖTTERÖD/NAGEL (2005) must be considered to be relatively time consuming. Since the application of the CTM described in this thesis has to operate in real-time, a short computing time is a crucial requirement. Therefore, only a modest extension of the CTM allowing 1:3-diverges and 3:1-merges has been chosen. In most cases this will be enough to model even complex intersections. Necessity for diverges and merges of higher order at ordinary intersections with no more than four approaches is rare.

Where applicable, the ordinary equations of the CTM are used. For the above-mentioned modest extension some more equations are needed. Extending the case of diverges does not pose any difficulties. The necessary equations can be derived by analogy to equations 4-7 to 4-9.

$$q_i(t) = \min\left\{S_i(t), R_{i+1,1}(t)/\beta_1, R_{i+1,2}(t)/\beta_2, R_{i+1,3}(t)/\beta_3\right\} \quad (4\text{-}14)$$

$$q_1(t) = q_i(t) \cdot \beta_1 \quad (4\text{-}15)$$

$$q_2(t) = q_i(t) \cdot \beta_2 \quad (4\text{-}16)$$

$$q_3(t) = q_i(t) \cdot \beta_3 \qquad (4\text{-}17)$$

The necessary equations for extending the merging case can be deduced by expanding the two-dimensional case depicted in figure 4-5 to the third dimension (cf. figure 4-8) so that three instead of only two flows from preceding cells can be determined.

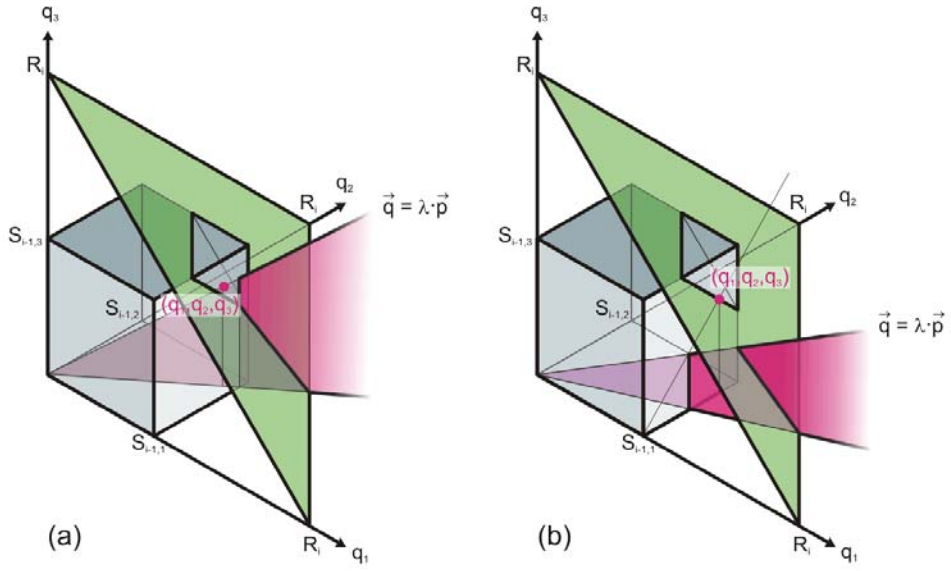


Figure 4-8: Extension of merging to three dimensions

In the three-dimensional case, the final solution must lie inside a cuboid that is bounded by the three planes $q_1(t) = S_{i\text{-}1,1}(t)$, $q_2(t) = S_{i\text{-}1,2}(t)$ and $q_3(t) = S_{i\text{-}1,3}(t)$. The first case to consider corresponds to equations 4-10 and 4-11 of the two-dimensional case. All three cells can send their respective volumes $S_{i\text{-}1,1}(t)$, $S_{i\text{-}1,2}(t)$ and $S_{i\text{-}1,3}(t)$ according to equations 4-18 to 4-20.

$$q_1(t) = S_{i-1,1}(t) \qquad (4\text{-}18)$$

$$q_2(t) = S_{i-1,2}(t) \qquad \text{if } S_{i-1,1}(t) + S_{i-1,2}(t) + S_{i-1,3}(t) \le R_i(t) \qquad (4\text{-}19)$$

$$q_3(t) = S_{i-1,3}(t) \qquad (4\text{-}20)$$

In all other cases, $R_i(t)$ of the merge cell will be used up completely and the solution lies on the green plane shown in figure 4-8. The relation between all three traffic flows $q_1(t)$, $q_2(t)$ and $q_3(t)$ leads to the straight line $\vec{q} = \lambda \cdot \vec{p}$. The vector $\vec{p}$ contains all priority percentages (summing up to 1) and λ is a scalar. In figure 4-8 this line is the upper edge of the red area. It has to be determined whether this line intersects with the green plan inside or outside of the cuboid (cases (a) and (b) in figure 4-8). Therefore the value λ_{min} has to be determined as follows:

$$\lambda_{min} = \min\left\{\frac{S_{i-1,1}(t)}{p_1}, \frac{S_{i-1,2}(t)}{p_2}, \frac{S_{i-1,3}(t)}{p_3}\right\} \qquad (4\text{-}21)$$

If λ_{min} is greater than $R_i(t)$, the intersection lies inside of the cuboid and the final solution is:

$$q_1(t) = p_1 \cdot R_i(t) \qquad (4\text{-}22)$$

$$q_2(t) = p_2 \cdot R_i(t) \qquad \text{if } \lambda_{min} \geq R_i(t) \qquad (4\text{-}23)$$

$$q_3(t) = p_3 \cdot R_i(t) \qquad (4\text{-}24)$$

If λ_{min} is smaller than $R_i(t)$, this implies that the line intersects with one of the three sides of the cuboid before intersecting with the green plane. To generalize the three possible outcomes the index of the respective side is named *j*. The other two indices are *k* and *l*. Flow $q_j(t)$ equals $S_{i\text{-}1,j}(t)$. The other two flows *k* and *l* can now be determined as in the two-dimensional case. Equations 4-12 and 4-13 have to be adapted slightly. The final solution is:

$$q_j(t) = S_{i-1,j}(t) \qquad (4\text{-}25)$$

$$q_k(t) = \text{mid}\left\{\begin{array}{l} S_{i-1,k}(t), \\ \left(R_i(t) - S_{i-1,j}(t)\right) \cdot \dfrac{p_k}{p_k + p_l}, \\ R_i(t) - S_{i-1,j}(t) - S_{i-1,l}(t) \end{array}\right\} \qquad \text{if } \lambda_{min} < R_i(t) \qquad (4\text{-}26)$$

$$q_l(t) = \text{mid}\left\{\begin{array}{l} S_{i-1,l}(t), \\ \left(R_i(t) - S_{i-1,j}(t)\right) \cdot \dfrac{p_l}{p_k + p_l}, \\ R_i(t) - S_{i-1,j}(t) - S_{i-1,k}(t) \end{array}\right\} \qquad (4\text{-}27)$$

4.3.2 Traffic signals

A special cell type has to be conceived to model traffic signals. ALMASRI (2006) and others proposed to set the outflows $q_i(t)$ of these cells to zero if the according traffic signal is red during time step *t*. If the signal is green, the regular model equations are applied. In this work, $S_i(t)$ instead of $q_i(t)$ is set to zero in the case of a red signal. This facilitates modeling of permitted left-turns. Amber and red-amber signal indications belong to the red phase, i.e. traffic can only flow if the signal shows green.

4.3.3 Permitted left-turns

According to the regular CTM equations, flows between cells are only influenced by the involved preceding and following cells. At merges priority rules are expressed by priority percentages. Situations in which permitted turns have to give way to opposing flows are not covered.

Figure 4-9 shows the concept used in this work to model such permitted turns. An appropriate superior cell *i* belonging to the opposing flow with right of way and a related subordinate cell *j* belonging to the permitted turn are chosen. First, $S_i(t)$ and $R_j(t)$ are determined as usual. Afterwards, $R_j(t)$ can be further reduced to $R_{j,red}(t)$ subject to $S_i(t)$ if $R_{j,red}(t) < R_j(t)$. This implies that vehicles are held in the cell preceding cell *j* if S_i is too high. This preceding cell defines the end

of an intersection internal storage area for turning vehicles. It would not be possible to model such an area if a general connector were used (cf. figure 4-7b) because all approaches to the intersection would end at the stop lines.

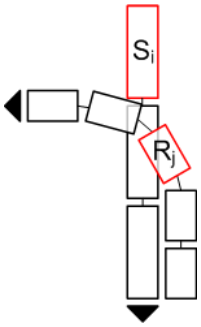


Figure 4-9: Handling of permitted left turns

Choosing $S_i(t)$ to reduce $R_j(t)$ allows modeling discharging vehicles during extended green time or intergreen times. If the signal that controls cell *i* turns red, $S_i(t)$ becomes zero and $R_j(t)$ is no longer reduced. The remaining permitted turning vehicles can discharge freely.

ROHDE ET AL. (2008) proposed to use equation 4-28 taken from the German HBS (FGSV, 2001) for proper reduction of $R_j(t)$.

$$C_D = \frac{3600 \cdot f}{t_f} \cdot e^{-q \cdot (t_g - t_f/2)/(3600 \cdot f)} \qquad (4\text{-}28)$$

where:

C_D	capacity of permitted left turn [veh/h]
q	traffic flow of opposing direction [veh/h]
f	portion of green time during one cycle [-]
t_g	smallest headway accepted by leading vehicle in queue [s]
t_f	additional headway accepted by following vehicles [s]

They adapted the equation in order to calculate the reduced capacity of a permitted left turn at the current time step *t*.

It has to be emphasized that equation 4-28 is designed to calculate the average capacity of a permitted left turn during one hour. It incorporates two components: the probability of different headways in an opposing flow *q* and the number of permitted left-turning vehicles accepting a specific headway. By integrating the product of both components over time, equation 4-28 is obtained. Time steps *t* of the CTM however are rather short. Their duration *T* ranges from one to a few seconds. Thus all $S_i(t)$ are calculated for a short time interval only. It can be argued whether it makes sense to consider probabilities of different headways for such a short interval. If $S_i(t)$ of the next time step *t* is known, the average headway during this time step under free flow conditions will be the reciprocal value of $S_i(t)$, and variations of that headway are small.

In this work, this reciprocal value is used as average headway of the opposing flow during time step *t*. The number *n* of permitted turning vehicles that accept this average headway is given by the Siegloch formula (SIEGLOCH, 1973):

$$n = \frac{\Delta t - t_0}{t_f} \qquad (4\text{-}29)$$

where:

Δt available (average) headway $T/S_i(t)$ of the opposing flow [s]

t_0 t_g - $t_f/2$ [s]

The final equation to determine $R_{j,red}$ is:

$$R_{j,red}(t) = n \cdot \frac{T}{\Delta t} = \frac{1}{t_f} \cdot \left(T - t_0 \cdot S_i(t)\right) \qquad (4\text{-}30)$$

Even more than one superior cell can be used. In this case the respective $S_i(t)$ of all superior cells have to be added up before applying equation 4-30. Moreover, this extension can also be used to model mere priority junctions without any traffic signals at all.

4.4 Estimation of performance indicators

4.4.1 Total delay

In this work, the main purpose of the CTM is to estimate the total delay imposed on the traffic by different signal settings or, more precisely, different offsets that create a certain coordination pattern.

According to DAGANZO (1994) and ALMASRI (2006) the total delay in vehicle seconds [veh·s] that is experienced by vehicles in one cell i during one simulation step t of duration T is:

$$d_i(t) = T \cdot \left(n_i(t) - q_i(t)\right) \qquad (4\text{-}31)$$

The term in brackets is equivalent to the number of vehicles that have already been in the cell at the beginning of time step t but could not leave it during the step. These vehicles are held in the cell during the time step and experience a delay of T seconds. The total network delay is the sum of all $d_i(t)$ over all time steps of a model run and over all cells of the modeled network.

4.4.2 Travel times

For traffic demand estimation a good estimate of travel times on different alternative routes is needed. Therefore, a simple approach to derive travel times from the CTM has been developed. By summing up the delays $d_i(t)$ and inflows $q_{i-1}(t)$ of each cell over all time steps t of one model run the total delay D_i and total inflow Q_{i-1} of each cell is obtained. The average delay of a vehicle having entered a specific cell i during the model run is thus $D_{avg,i} = D_i / Q_{i-1}$. The travel time on a link of the network can then be estimated by adding the sum of average delays of all its cells to its free flow travel time which corresponds to its number of cells multiplied by the duration T of a time step. Travel times for routes are obtained by adding up the travel times of all corresponding links.

4.5 Implementation

The CTM has been implemented in Java. The program builds CTM networks in a semi-automatic process. It can read a text file exported from AIMSUN. The file contains information on the topology of the network as modeled in the microsimulator. The program creates road sections and the according cells automatically. Turnings at intersections and traffic signals have to be added manually. A graphical user interface (GUI) supports the user. It also enables further changes and adjustments of the network, as well as input of source flows, turning and priority percentages and fixed time signal plans. The final CTM network can be stored in an XML file.

The implementation of the CTM has been optimized for short runtimes. All cells are stored in different lists depending on their type. These lists are not exclusive. A cell can be in more than one list. The different lists comprise origin cells, destination cells, ordinary cells with one successor cell, diverging cells with two successors, diverging cells with three successors, merging cells with two predecessors, merging cells with three predecessors, and cells that are subordinate to other cells (permitted turns). An additional list contains all signal groups. These are related to the cells they control.

At each time step t the program executes the following steps:

Step 1: Calculate all $S_i(t)$ and $R_i(t)$.

Step 2: Update all signal groups and set $S_i(t)$ of controlled cells to 0 if the signal is red.

Step 3: Reduce $R_i(t)$ of subordinate cells according to the sum of $S_i(t)$ of all superior cells.

Step 4: Run through all ordinary cells and determine $q_i(t)$.

Step 5: Run through both lists of diverge cells and determine $q_i(t)$ and the sub-flows.

Step 6: Run through both lists of merge cells and determine $q_{i-1}(t)$ and the sub-flows.

Step 7: Update number of vehicles $n_i(t+1)$, total inflow Q_i and total delay D_i of all cells.

At the end of the model run, the total delay experienced in the network and travel times of all routes are calculated. The CTM has been implemented in such a way that it can start with an empty or preloaded network. It can also consider a warm-up period that will not be included into the calculations of total delay and travel times.

4.6 Validation

The test network in List district, Hanover, has been modeled as CTM. The duration T of time steps has been set to one second. This is a good choice for urban networks. It enables a quite precise modeling of urban network topology since the spatial resolution that is directly related to the temporal resolution will also be high. Furthermore, signal plans can be reproduced precisely because switching commands are given on a second-to-second basis. With v_f = 13.8 m/s (50 km/h) and T = 1 s, each cell has a length of 13.8 meters. Increasing the duration T would reduce the spatial resolution, and important details of the network might be distorted. Long cells are only acceptable for modeling highways. A further reduction of T below one second would improve the accuracy of the modeled topology, but at the expense of more and thus longer computation time which is a crucial factor for an online application. The final network is shown in figure 4-10.

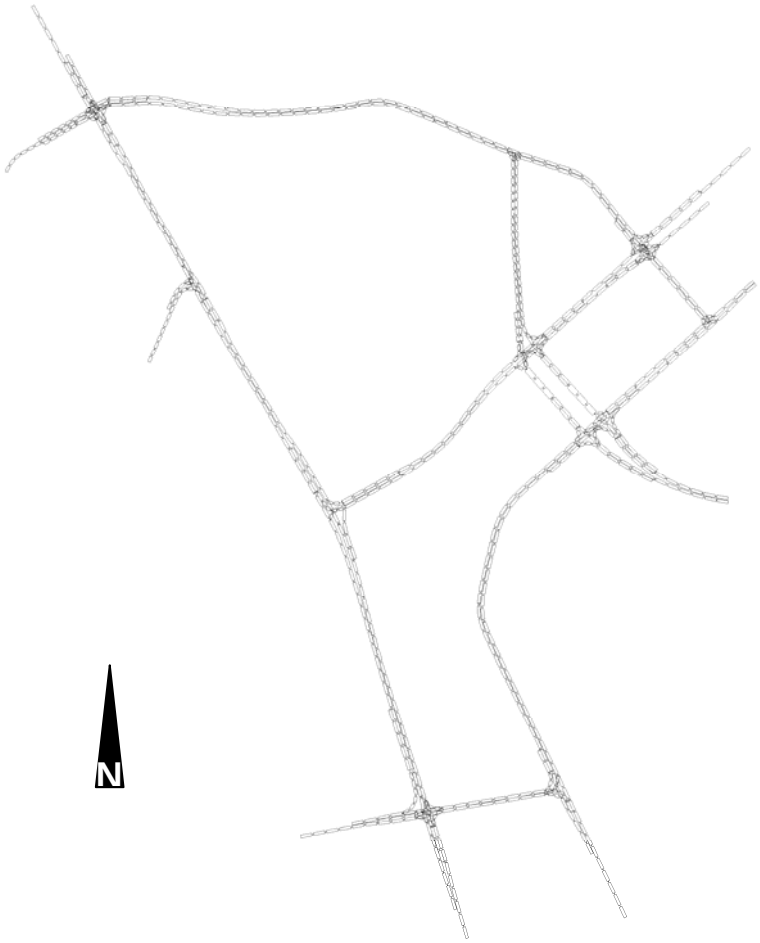


Figure 4-10: CTM representation of the test network in Hanover List district

The network consists of 841 cells. Simulating a time period of 15 minutes takes 52 milliseconds on average on an Intel Core i7-920 quad-core processor with 2.67 GHz using only one processor. The prototypical implementation has not been designed for parallel computing on two or more processors. Since the optimization interval has a duration of 15 minutes, almost the same amount of time can be used for optimization of signal settings. This implies that about 17,000 runs of the CTM can be executed.

Besides computation rate, accuracy of the model is of paramount importance. To investigate the accuracy of the CTM a comparison between AIMSUN and the CTM has been conducted, where AIMSUN is considered to be an acknowledged reference. Several test scenarios have been simulated both with AIMSUN and the CTM. Different traffic demands on the basis of the real measurements taken on-site have been used (cf. paragraph 3.3.3). All demands are expressed by source inflows at the boundaries of the network and turning percentages at all intersections. The tests comprised both morning and afternoon peak intervals. For both cases three scenarios with 100 percent, 75 percent, and 50 percent demand have been simulated. Each simulation had a duration of 15 minutes. The respective real fixed time signal plans of the morning and afternoon period have been used in the scenarios.

For comparison of both models the following measures have been used: traffic flows [veh/h] at the locations of the 55 loop detectors, delays [veh·s] on 50 sections of the network and travel

times [s] on the 108 predefined routes. The CTM as deterministic model must run only once per scenario. AIMSUN however uses a stochastic model. The outcome of each simulation run depends on the respective random seed number of the run. Therefore, average values resulting from 30 different runs are taken for comparison to the CTM results.

Different CTM parameter settings have been tested. The finally chosen parameters that produced the best results are:

free flow speed:	v_f =	13.89 m/s (50km/h) and 8.33 m/s (30 km/h) respectively depending on streets, several adapted speeds for different turnings
backward wave speed:	w =	v_f
flow capacity:	Q_i =	0.5 veh/s (1800 veh/h) for single lane cells, multiple lane cells adapted accordingly
jam density:	k_{jam} =	0.16667 veh/m (166.67 veh/km □ 6 m/veh)
smallest accepted headway:	t_g =	4 s
additional accepted headway:	t_f =	2 s

Table 4-1 outlines the findings on accuracy. The according diagrams are shown in appendix A and oppose flows, delays and travel times resulting from AIMSUN to the results of the CTM.

The modeled flows achieve the best accuracies. But even delays produced by both models correspond well. Estimation of travel times performs only slightly worse but is still of good quality. The results are in accordance with similar tests performed by ALMASRI (2006) who achieved comparable results for flows and delays. It can be concluded that the CTM is appropriate for modeling the effects of signalization in urban networks.

Table 4-1: Results of CTM validation

		morning peak			afternoon peak		
	demand	100%	75%	50%	100%	75%	50%
flow	r_{xy} [-]	0.999	0.999	0.998	0.999	0.999	0.997
	RMSE [veh/h]	7.352	5.964	4.700	9.005	7.277	6.369
	RRMSE [-]	0.037	0.037	0.046	0.040	0.040	0.050
delay	r_{xy} [-]	0.979	0.992	0.994	0.988	0.989	0.994
	RMSE [veh·s]	265.499	107.412	61.991	275.759	150.857	71.490
	RRMSE [-]	0.172	0.129	0.111	0.150	0.138	0.133
travel time	r_{xy} [-]	0.947	0.937	0.914	0.940	0.929	0.907
	RMSE [s]	24.430	22.090	24.727	22.550	20.739	22.474
	RRMSE [-]	0.143	0.139	0.160	0.136	0.134	0.151

5 Demand estimation

5.1 Overview

This chapter describes the demand estimation that is used in the framework of the ATCS prototype. An estimated demand is needed for each optimization interval in order to adapt the signal settings accordingly.

Paragraph 5.2 describes the forecasting technique that is used in this thesis to forecast mere detector counts of the next optimization interval, as well as its evaluation based on simulated data taken from the AIMSUN simulation as described in paragraph 3.3. Paragraph 5.3 deals with the estimation of OD matrices and traffic volumes on routes and links based on given traffic flows on links that are equipped with a detector. Again, results of the evaluation of this technique are presented.

Both modules can be combined, so that the traffic demand estimation is based on forecasted detector counts. An evaluation of the performance of this combination of forecasting and demand estimation is presented in paragraph 5.4.

5.2 Forecasting of detector counts

The first module of traffic demand estimation executes a forecast of detector counts for the next optimization interval. These counts will then be used in the next module as constraints for the subsequent estimation of an OD matrix and resulting route and link volumes. Detectors may be of any type as long as they provide vehicle counts. The currently still most common detectors in urban networks are loop detectors, but other technology becomes more and more widespread as well.

The forecasting technique employed in this work is based on a method proposed by FÖRSTER (2008) using current and reference space-time-patterns of detector counts. The method has been chosen due to its simple but obviously effective approach. It has been investigated, further enhanced and slightly adapted.

5.2.1 Concept

FÖRSTER (2008) proposed this new approach to forecast traffic counts for a relatively short prognosis horizon of 20 minutes. The method forecasts traffic counts of all j_{max} detectors in a

sub-network simultaneously. It uses two-dimensional reference space-time-patterns of traffic counts which are derived from available data from the past. Over a certain period of time of several weeks or months, counts $q^{i,j}$ of every detector j in the sub-network are collected for all time intervals i of a day (i_{max} = 72 in the case of 20-minute-intervals). The data is clustered into several relevant groups (e.g. different weekdays, Sundays, holidays), each containing data of K days. The reference pattern of each group consists of reference values $q_{ref}^{i,j}$ that are the averages of all counted values of a specific time interval i for a specific detector j, so that:

$$q_{ref}^{i,j} = \frac{1}{K} \sum_{k=1}^{K} q_k^{i,j} \qquad (5\text{-}1)$$

where:

K number of days in the cluster

$q_k^{i,j}$ count of detector j at time interval i of day k [veh/interval]

Each reference pattern thus contains i_{max} times j_{max} reference values. Once the patterns are derived, they can be used in the future to forecast detector counts. In order to forecast the counts of the next time interval to come, the real detector counts $q_{real}^{i,j}$ that have been observed during the currently ended time interval $i = T_0$ and the previous N time intervals T_{-1} to T_{-N} are taken as current traffic count pattern with N+1 times j_{max} values that characterize the traffic demand that has currently passed the sub-network. Then, all reference patterns are scanned from the beginning to the end for a traffic count pattern of the same size that matches best the current count pattern. For identifying the best matching sub-pattern within the reference patterns, two approaches have been proposed by FÖRSTER (2008): either the correlation coefficient r_{xy} as given in equation 5-2 or the mean squared error MSE as given in equation 5-3 is used. The first has to be preferably high, the second preferably low.

$$r_{xy} = \frac{\sum_{j=1}^{j_{max}} \sum_{i=T_0}^{T_{-N}} \left(q_{real}^{ij} - \overline{q}_{real}\right) \cdot \left(q_{ref}^{ij} - \overline{q}_{ref}\right)}{\sqrt{\sum_{j=1}^{j_{max}} \sum_{i=T_0}^{T_{-N}} \left(q_{real}^{ij} - \overline{q}_{real}\right)^2} \cdot \sqrt{\sum_{j=1}^{j_{max}} \sum_{i=T_0}^{T_{-N}} \left(q_{ref}^{ij} - \overline{q}_{ref}\right)^2}} \qquad (5\text{-}2)$$

$$MSE = \frac{1}{(N+1) \cdot j_{max}} \cdot \sum_{j=1}^{j_{max}} \sum_{i=T_0}^{T_{-N}} \left(q_{real}^{ij} - q_{ref}^{ij}\right)^2 \qquad (5\text{-}3)$$

where:

$q_{real}^{i,j}$ real observed count of detector j at interval i [veh/interval]

$\overline{q}_{real}$ arithmetic mean of current traffic count pattern [veh/interval]

$q_{ref}^{i,j}$ reference count of detector j at interval i [veh/interval]

$\overline{q}_{ref}$ arithmetic mean of reference sub-pattern [veh/interval]

N number of intervals prior to current interval [-]

j_{max} number of detectors [-]

Once the best reference sub-pattern has been found (comprising the time intervals T'_0 to T'_{-N}), the values $q_{ref}^{T'_{+1},j}$ of the next time interval following the sub-pattern are taken and the forecasted values $q^{T_{+1},j}$ are determined as follows:

$$q^{T_{+1},j} = \frac{\sum_{i=T_0}^{T_{-N}} \sum_{j=1}^{j_{max}} q_{real}^{i,j}}{\sum_{i=T'_0}^{T'_{-N}} \sum_{j=1}^{j_{max}} q_{ref}^{i,j}} \cdot q_{ref}^{T'_{+1},j} \qquad (5\text{-}4)$$

It is obvious from equation 5-4 that FÖRSTER (2008) does not simply use the plain reference counts of the next interval that follows the best matching reference sub-pattern as forecasted counts. In fact he rather proposes to adjust the reference values according to the relative difference between the observed count pattern and the reference sub-pattern. This difference is expressed by the fraction in equation 5-4.

5.2.2 Evaluation

FÖRSTER (2008) used real data taken from 52 inductive loops in a sub-network of Nuremberg, Germany. The data comprised a total of 21 weeks. Best results were achieved, if the correlation coefficient according to equation 5-2 and traffic count patterns comprising a total of four intervals (current interval plus $N = 3$ preceding intervals) are used to identify the best matching reference sub-pattern. For the evaluation, only traffic volumes of more than 20 veh/20 min were considered. FÖRSTER (2008) found that about 75 percent of all predicted counts in his scenario had a relative deviation of less than 10 percent from the real counts. The mean absolute percentage error *MAPE* was at about 16 percent. Systematic over- or underestimation has not been observed, but the data has only been analyzed as a total and not for each time interval separately. Thus, systematic over- or underestimation within single intervals cannot be excluded.

The original method has also been tested in this study using the artificial data logged from the whole-day-simulation described in paragraph 3.3. Given optimization intervals of 15 minutes, according time intervals of 15 minutes have been used instead of 20 minutes to test the forecasting.

30 replications (i.e. runs) of the AIMSUN simulation with different random seed numbers have been run. For every time interval of every replication, all detector counts produced by AIMSUN have been logged. Even though the same defined traffic demand has been used in every replication, detector counts vary in a certain range because of the random behavior of the microsimulator. The logged detector counts can thus be interpreted as $K = 30$ samples of the daily traffic demand of a specific weekday. Using these samples, a reference pattern containing the average counts for every interval and detector has been created using equation 5-1. Figure 5-1 visualizes the pattern using different colors for different values of counted vehicles.

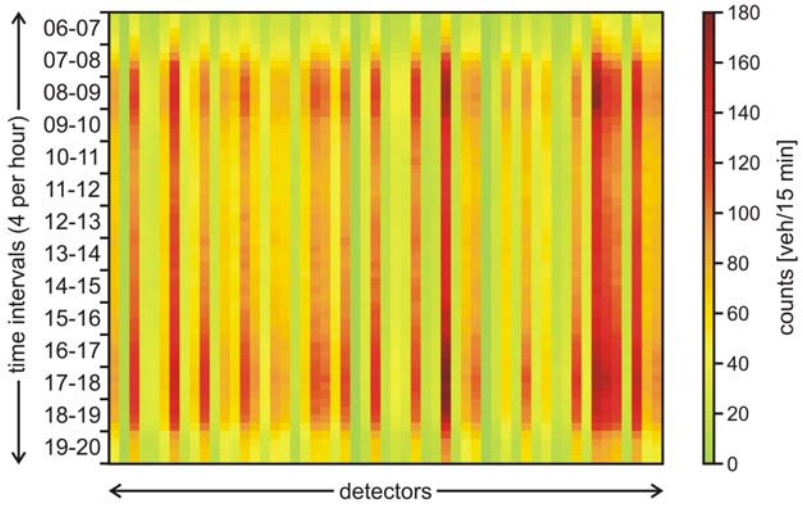


Figure 5-1: Reference count pattern derived from simulated data

This pattern has been used to analyze how well the “real” detector counts from every individual replication can be forecasted. As has been reasoned earlier in paragraph 3.1 the ATCS prototype has to optimize signal plans of the time interval that is two intervals ahead of the last complete interval. Therefore, not only forecasted counts of interval T_{+1} but also of interval T_{+2} were of note in this study. Additionally, the achieved quality of the forecasted counts has been examined not only as a total but for each time interval separately.

Two simplifications have to be highlighted. First, the tests in this study comprise only one single reference pattern, since only data for one single daily demand was available for simulation. While the original method scans all available reference patterns for the currently observed detector count pattern, in this study only one reference pattern is scanned which happens to be the correct one. However, this does not corrupt the basic functionality of the method. Moreover, the correct pattern would have been scanned anyway among others, and the finally chosen sub-pattern is likely to be part of this pattern which is closest to the current demand. And finally, considering only one reference pattern revealed an effect that would have been more difficult to discover if more than one reference pattern had been used. This effect will be explained in the following paragraphs.

Second, the reference pattern has been derived from the same set of replications whose individual detector counts are used to test the forecasting. However, by using 30 replications it has been made sure that the reference pattern can be used without any problems. The number of replications is sufficiently high to level out all variations of detector counts of specific replications in the reference pattern. Using 30 other replications would have resulted in the same reference pattern with only insignificant differences.

In a first step, the approach has been used as recommended, i.e. $N = 3$ and the correlation coefficient is used to identify the best matching sub-pattern. Figure 5-2 shows the result for an arbitrarily chosen replication. It comprises the forecasted counts of all 55 detectors and 52 intervals. Even though the complete simulation comprises 56 intervals of 15 minutes, detector counts of the first four intervals cannot be forecasted because the method needs detector

counts from four intervals in a row to scan the reference pattern. Consequently, detector counts of the fifth interval are the first that can be forecasted.

Figure 5-2a shows a comparison between the forecasted counts and their corresponding “real” counts. A sometimes occurring systematic over- or underestimation at single intervals can be suspected but is not clearly visible. The way the reference pattern is derived (cf. equation 5-1) leads to the fact that the method tends to forecast average counts by nature. Therefore figure 5-2b compares the forecasted counts of each interval to the corresponding average counts of the same interval which are obtained by averaging detector counts over all 30 replications. The faint lines differing upwards or downwards from the diagonal clearly indicate several intervals whose counts are systematically over- or underestimated.

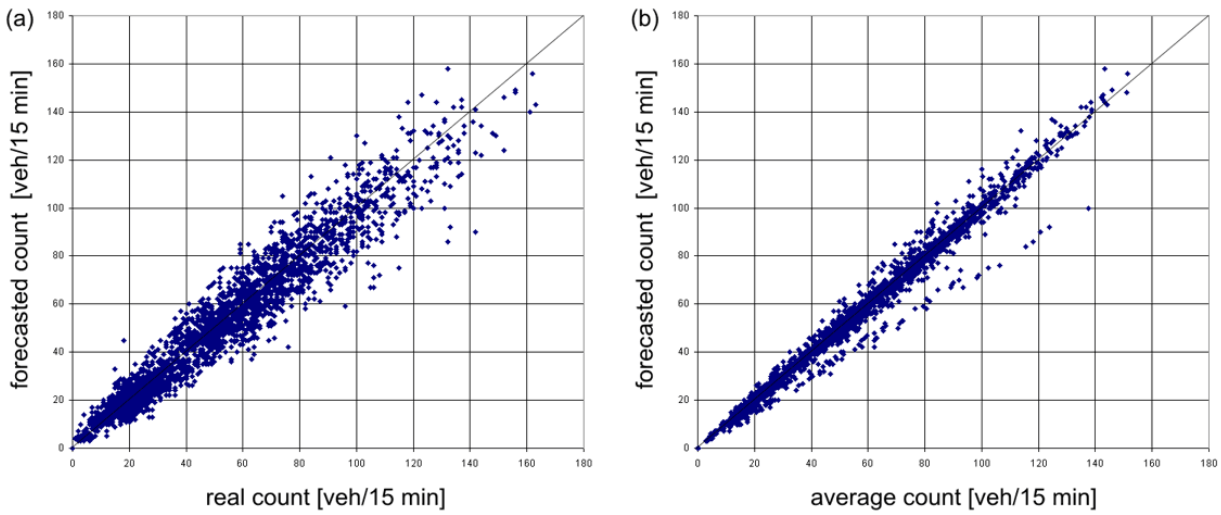


Figure 5-2: Example of quality of forecasted counts (original method)

Table 5-1 gives an overview of the quality of forecasting for all time intervals together. For each interval of each replication (52·30 = 1,560 intervals) the quality criteria r_{xy}, root mean square error *RMSE* and relative root mean square error *RRMSE* have been determined based on the 55 forecasted values. Again, both real counts and average counts have been used for comparison (see column “reference”), and the forecasting has been done one and two intervals ahead.

Table 5-1: Overall quality of forecasted counts (original method)

reference	intervals ahead	r_{xy}	*RMSE*			*RRMSE*		
		avg [-]	avg [veh/h]	max [veh/h]	>10 veh/h [%]	avg [-]	max [-]	> 0.2 [%]
real counts	1	0.98	6.92	22.96	3.33	0.14	0.45	4.23
	2	0.98	7.28	30.91	6.21	0.15	0.51	6.54
average counts	1	1.00	1.88	18.72	1.03	0.04	0.33	0.96
	2	1.00	2.32	28.79	3.27	0.04	0.44	2.75

Good average r_{xy}, *RMSE* and *RRMSE* have been achieved over all 1,560 intervals. The respective maximum values of *RMSE* and *RRMSE* however show that the forecasted counts of some

individual intervals suffer from an important loss of accuracy. The portion of intervals that exceed an *RMSE* of 10 veh/h ranges from 1.03 to 6.21 percent depending on the case. An *RRMSE* of 0.2 is exceeded in 0.96 to 6.54 percent of the intervals.

5.2.3 Discussion and modification

A close examination of the reference sub-patterns that have been identified by the scan algorithm to be the best matching sub-patterns revealed the cause of the sometimes occurring systematic over- or underestimation of traffic counts. These over- or underestimations originate from the fact that a high r_{xy} does not guarantee that a reference sub-pattern is identified which covers exactly the same time intervals T_{-N} to T_0 as the current time pattern of the real counts. According to FÖRSTER (2008) the method is not intended to always find a sub-pattern that is temporally congruent to the current pattern. The method is much more based upon the assumption that specific typical traffic patterns can be observed at different times of the day or even at completely different days. However, identifying inappropriate sub-patterns can have a negative impact, whether they are temporally congruent or not.

The case of a reference sub-pattern that covers another sequence of intervals than the current pattern is illustrated in figure 5-3. Note that for simplification only the sum of all detector counts of an interval is displayed in the illustration. In this example, a reference sub-pattern from an earlier time of day is chosen because the current pattern is highly proportional to the first, resulting in a high r_{xy}. The current pattern correlates less well to the sub-pattern that really corresponds to the currently elapsed time intervals.

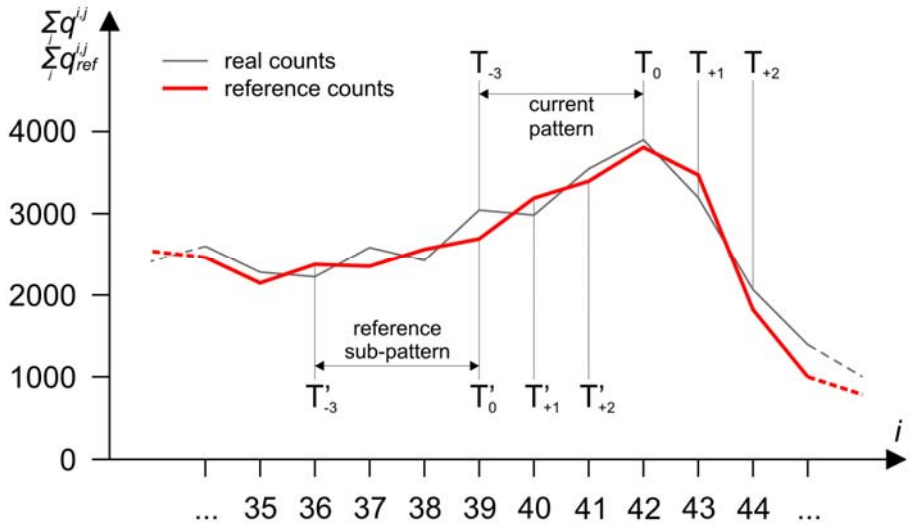


Figure 5-3: Cause for over- or underestimation of forecasted traffic counts

Equation 5-4 is meant to handle this effect. The reference counts of time interval T'_{+1} or T'_{+2} are adapted according to the difference between current pattern and reference sub-pattern. However, this adaptation assumes that the development of traffic demand in the intervals following the real time intervals T_{-N} to T_0 is more or less proportional to the reference traffic demand in the intervals following the correlating time intervals T'_{-N} to T'_0. This is not necessarily true, as figure 5-3 also shows. The traffic demands in the reference pattern at time intervals T'_{+1} and T'_{+2} are rising, while the real demands at time intervals T_{+1} and T_{+2} are decreasing. Equation 5-4

cannot cope with this case. In fact, it will even deteriorate the quality of forecasting in the example. The sum of the real counts in the current pattern is much higher than the sum of the reference counts, so that the reference values of the intervals T'_{+1} and T'_{+2} will be further increased accordingly, whereas only a moderate increase in the case of T_{+1} and even a reduction in the case of T_{+2} would be appropriate to produce a good forecast for these two intervals.

To overcome the problem of partly occurring over- or underestimation, the two ways proposed by FÖRSTER (2008) to identify the best matching reference sub-pattern have been combined in this work. In a first step, not only one, but three reference sub-patterns are identified according to the three highest r_{xy}. Among these three candidates, the one with the lowest *MSE* is chosen as the final reference pattern. A low *MSE* indicates that the reference counts are close to the real counts of the current pattern.

Figure 5-4 in comparison with figure 5-2 shows that the quality of the forecasted counts could be improved in the case of the chosen sample replication. However, the adapted version still cannot avoid systematic over- or underestimation completely. The data from some other replications still leads to these unwanted effects.

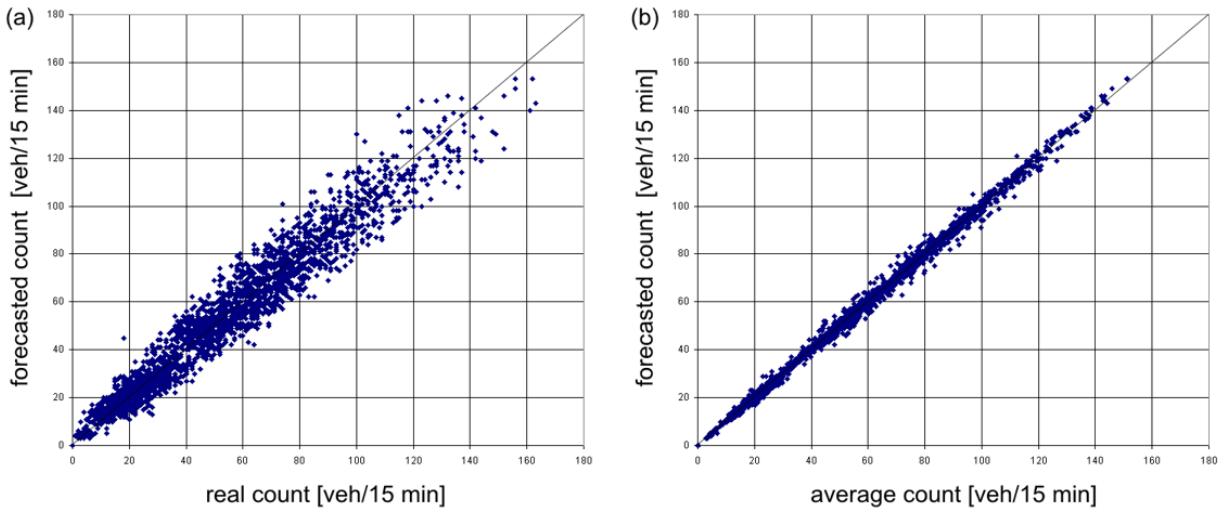


Figure 5-4: Example of quality of forecasted counts (modified method)

Table 5-2 shows the overall quality of the modified forecasting method.

Table 5-2: Overall quality of forecasted counts (modified method)

reference	intervals ahead	r_{xy}	*RMSE*			*RRMSE*		
		avg [-]	avg [veh/h]	max [veh/h]	>10 veh/h [%]	avg [-]	max [-]	> 0.2 [%]
real counts	1	0.98	6.76	16.77	2.24	0.14	0.35	2.95
	2	0.98	6.88	26.87	3.01	0.14	0.39	3.66
average counts	1	1.00	1.42	15.93	0.38	0.03	0.22	0.26
	2	1.00	1.55	26.04	1.18	0.03	0.38	1.05

All quality criteria (except for r_{xy}) have improved, but still some few intervals with a reduced quality of forecasted counts remain. Especially the maximum *RMSE* and *RRMSE* have decreased which suggests a reduction of systematic over- and underestimation. The modified version as described is used in the framework of the ATCS prototype.

5.3 OD matrix, route and link volume estimation

Once the detector counts of the next optimization interval have been forecasted, they can be used to estimate the upcoming overall traffic demand, i.e. OD flows, route and link volumes of the respective interval. The module that handles this task is based on the work by WANG (2008) and FRIEDRICH/WANG (2006) who employ the IM model presented by VAN ZUYLEN/WILLUMSEN (1980). The work of WANG (2008) has been analyzed and implemented. Some enhancements have been developed and added to further improve the estimation process.

5.3.1 Concept

Before the overall traffic demand can be estimated, the network must be transformed into a directed graph that consists of links and nodes. Paragraph 5.3.2 will give some details on the creation and certain specifics of this graph.

The basic feature of the module for OD flow, route and link volume estimation is the IM model by VAN ZUYLEN/WILLUMSEN (1980). It has been developed initially to estimate mere OD flows by using available traffic counts on several links of the network as constraints for the estimation. A precise knowledge on how trips of a specific OD pair are assigned to different alternative routes (or more precisely the corresponding links) is a precondition of the model. The IM model will be presented in paragraph 5.3.3.

For best estimation results the constraints should be as detailed and precise as possible. The most detailed constraints are traffic counts on all links that represent turnings (left, through, right) at intersections. However, as has been mentioned before, not all turnings at intersections can be equipped with detectors. A simple iterative algorithm is proposed in paragraph 5.3.4 that tries to enhance the information obtained from available detector counts with regard to the desired turning flows. The algorithm derives as many missing counts on links without detectors as possible.

The detector counts and the derived link volumes cannot be expected to be entirely consistent due to faulty measurements and imprecise forecasting. Paragraph 5.3.5 will present a method proposed by VAN ZUYLEN/BRANSTON (1982) to overcome inconsistencies.

WANG (2008) dealt with the problem of a reduced quality of estimation of the IM model if redundant information is used as constraints. This drawback had been described earlier by VAN ZUYLEN (1981). WANG (2008) and FRIEDRICH/WANG (2006) defined several rules in order to eliminate redundant information while maintaining the most precise constraints as possible. In this thesis a general algorithm covering all of these rules at once has been developed. Paragraph 5.3.6 will discuss the problem of redundant constraints and its solution.

Another problem that WANG (2008) addressed is the aforementioned precondition of the IM model. For each link the model needs to know the exact portions of trips of a specific OD relation that pass this link. In other words, a static *a priori* assignment is assumed that is independent of the final OD matrix. Such portions however are generally not known precisely in practice.

WANG (2008) tried different traffic assignment techniques to estimate these portions. Paragraph 5.3.7 will show how such techniques can be added to the estimation process.

Paragraph 5.3.8 presents the final estimation process that combines all of the aforementioned aspects. The final module integrates the IM model and its modifications into an iterative procedure as proposed by WANG (2008). It will also be shown how route and link volumes can be derived from the estimation. Finally, paragraph 5.3.9 will show some results of the performance of the OD matrix and route and link volume estimation.

5.3.2 Graph representation of the network

The module for traffic demand estimation works on a graph representation of the network. This graph consists of a number of links which are connected by nodes. A directed graph is used, i.e. each link represents only one driving direction of a road.

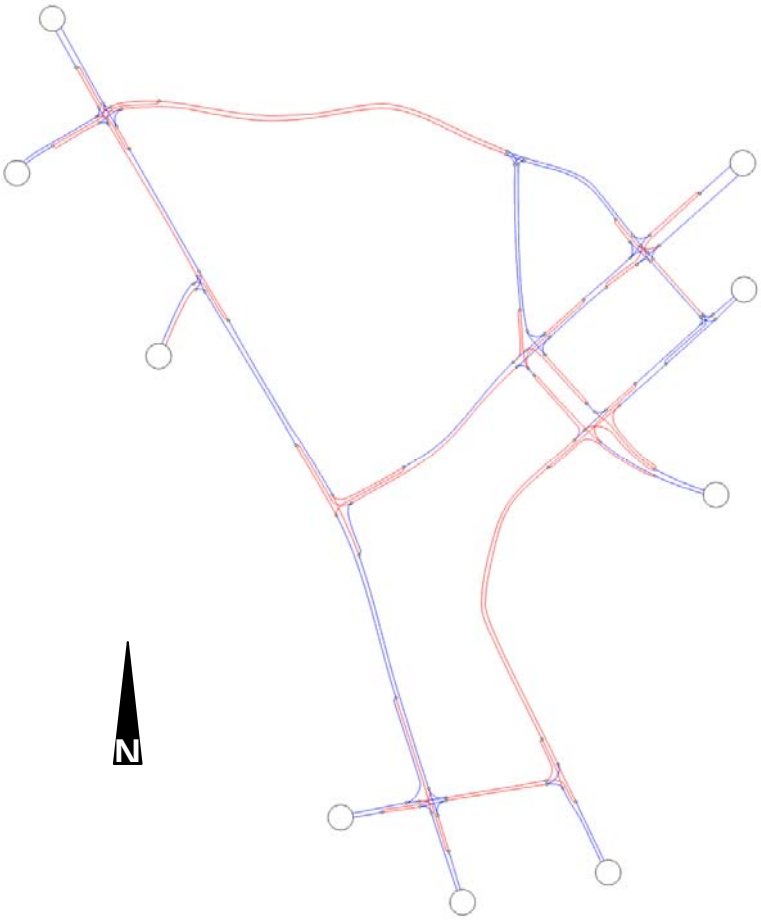


Figure 5-5: Graph representation of the test network in Hanover List district

To facilitate the creation of the graph a Java module has been programmed that transforms the CTM representation of the network (cf. paragraph 4.6) into a graph. This process is fully automatic. Every sequence of ordinary cells with one predecessor and one successor cell is transformed into a link. All diverges and merges are transformed into nodes. This procedure results in a graph where nodes can have either one incoming link and two or three outgoing links or vice versa. Nodes with multiple incoming and outgoing links are not possible because no gen-

eral *n*:*n*-connectors have been used in the CTM. Another consequence is that all turning lanes at intersections or in the case of mixed lanes the mere turnings are modeled by separate links. Each link that represents a lane equipped with a detector can be related to this detector, i.e. the forecasted traffic counts can be assigned to these links accordingly.

Figure 5-5 shows the final graph of the List network. Red links are related to a detector, i.e. they represent lanes that pass a traffic signal.

5.3.3 Information Minimization model

The IM model has been presented by VAN ZUYLEN/WILLUMSEN (1980), even though it has been developed by VAN ZUYLEN alone. WILLUMSEN developed a similar model based on entropy maximization which is presented in the same paper.

The problem that both models try to solve can be described as follows. A network represented as a directed graph with several links and nodes has a set of origins *i* and destinations *j* where vehicles can enter and exit the network. The flows f_{ij} between different OD pairs are in general unknown and have to be estimated. Direct observation of these flows is impossible unless all vehicles are tracked. This is not entirely unimaginable with modern GPS and navigation technology, but yet infeasible in practice. What can be observed instead are flows q_a on some of the links *a* in the network, either by manual counts or by automatic detection. These flows or counts q_a can be used as constraints for the estimation of the unknown OD flows f_{ij}. The estimated flows have to satisfy the following equation:

$$q_a = \sum_{ij} p_{ij}^a \cdot f_{ij} \qquad (5\text{-}5)$$

where:

q_a observed flow on link *a* [veh/h]

p_{ij}^a portion of vehicles passing link *a* while travelling from origin *i* to destination *j* [-]

f_{ij} estimated flows from origin *i* to destination *j* [veh/h]

The portions p_{ij}^a express an underlying and fixed *a priori* assignment of traffic to the network, i.e. the distribution of vehicles to different alternative routes is assumed to be known and static. The portions can take real values between 0.0 and 1.0.

If there are *n* links in the network with an available count q_a, *n* different equations of the above form can be set up, resulting in a set of *n* linear equations. Given *m* possible relations *ij* in the network (internal trips from and to the same origin and destination, i.e. $i = j$, are in general excluded), all f_{ij} can be determined precisely if $m = n$. If $m < n$, a precise solution is also possible if "*n* minus *m*" equations are linearly dependent on other equations of the system. Otherwise, some equations would contradict each other, and no solution could be found.

In most cases the number *m* of unknown flows f_{ij} exceeds the number *n* of linearly independent equations. The system of linear equations is under-determined, and a unique solution does not exist. A possible way to identify an appropriate solution among the infinite number of existing solutions is to choose the most likely one. VAN ZUYLEN (1979) and VAN ZUYLEN/WILLUMSEN (1980) respectively employ the information theory to define the most likely solution. According to this theory the most likely solution contains the least information. Therefore, VAN ZUYLEN's approach to OD matrix estimation is called "information minimization" model. It would go beyond

the scope of this chapter to describe the whole theory behind the model. WANG (2008) gives a short explanation. A more detailed explanation of the model can be found in appendix B of this thesis.

Applying the information theory to the stated problem as described by VAN ZUYLEN/WILLUMSEN (1980) leads to equation 5-6 after several steps and transformations (cf. appendix B for details). It allows estimating the unknown flows f_{ij}.

$$f_{ij} = f_{ij}^0 \prod_a X_a^{p_{ij}^a / g_{ij}}, \quad g_{ij} = \sum_a p_{ij}^a \qquad (5\text{-}6)$$

where:

f_{ij}^0 flow between OD pair *ij* according to a historic matrix [veh/h]

X_a proportionality factor for each link used as constraint [-]

The estimation process does not start from scratch but depends on a historic matrix with flows f_{ij}^0. These are adjusted to new flows f_{ij} which satisfy the constraints. The proportionality factors X_a are auxiliary variables whose role will become clear in the following.

It has to be emphasized that indexes *a* under the pi and sigma signs only comprise links *a* with available counts that are used as constraints. All other links are neglected. The sum g_{ij} combines all p_{ij}^a of all links *a* concerning a specific relation *ij*. Thus, the exponent p_{ij}^a / g_{ij} can be interpreted as some sort of weighting coefficient of the information on link *a* for flow f_{ij}.

Based on equation 5-6 all current flows f_{ij} between all origins *i* and destinations *j* can be estimated. VAN ZUYLEN/WILLUMSEN (1980) proposed an iterative algorithm to find a solution $\{f_{ij}\}$ which satisfies the set of equations 5-5. The necessary steps are:

Step 1: Set iteration counter *k* to zero.

Step 2: Set all proportionality factors $X_a^{k=0}$ to 1.

Step 3: Set all flows $f_{ij}^{k=0}$ to the flows f_{ij}^0 from the historic matrix.

Step 4: Determine all currently estimated traffic flows $q_{a,est}^k$ on links *a* according to equation 5-5 using the flows f_{ij}^k of the current iteration *k*.

Step 5: Verify whether all estimated flows $q_{a,est}^k$ differ from the real flows q_a by less than a predefined percentage (e.g. ±5 percent). If so, finish the estimation. If not, continue with step 6.

Step 6: Adapt all X_a^k for the next iteration *k*+1 according to $X_a^{k+1} = X_a^k \cdot \frac{q_a}{q_{a,est}^k}$.

Step 7: Increment *k*.

Step 8: Determine f_{ij}^k according to equation 5-6 using X_a^k and go to step 4.

The algorithm finds a stable solution after several iterations *k*. The number of necessary iterations depends on the size of the network and the number of constraints and unknown flows f_{ij}.

5.3.4 Improvement of constraints for estimation

The IM model does not necessarily depend on specific links to be used as constraints. Any link with available traffic volumes can be used. The only precondition is that each flow f_{ij} to be estimated has to pass at least one link a with an available count that can be used as constraint. Otherwise no information at all is available for flow f_{ij} and it cannot be included in the estimation.

So, even if no specific links have to be used as constraints, a simple example illustrated in figure 5-6 shows that available counts on some links are more valuable for a good estimation than those on other links. In case (a) only traffic flows on the approaches of the depicted intersection are available. These links are drawn in red. An estimation of all twelve unknown flows f_{ij} between all four zones is possible but only eight constraints exist. The problem is underdetermined and the solution is not unique and thus uncertain. In case (b) traffic flows on all turning links of the intersection are available, i.e. each of the twelve flows f_{ij} can be determined exactly. In a more general network with more intersections at which all turning flows are available the final solution will not be unique as in this example, but it becomes evident that more precise constraints in the form of turning flows are more valuable for the estimation process. WANG (2008) further investigated this effect and showed for two different networks that the quality of estimation improves with higher precision of constraints.

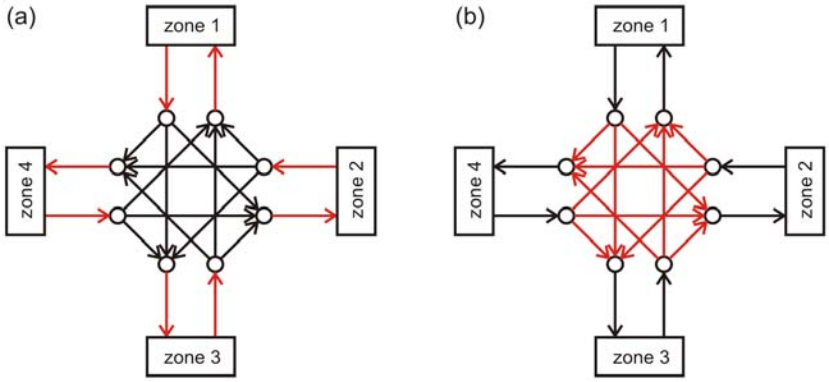


Figure 5-6: Comparison of different available constraints

Thus, directly detected turning flows at intersections are preferable for use as constraints because they contain the most precise information for the subsequent traffic demand estimation for the whole network. For the case of Germany and other countries as well, loop detectors can be expected to be located on some or all lanes approaching a signalized intersection. If lanes reserved for a specific turning movement (separate left-turn or right-turn lanes) exist, traffic volumes of these turnings can be detected directly. In many cases, however, there are only mixed lanes, and therefore only counts of combined through and turning flows are available. If some turnings are excluded from signalization (e.g. free right turns) no information on these flows is available at all. At uncontrolled intersections no information on turning flows is available either.

A simple iterative algorithm has been developed that improves the available constraints before starting the matrix estimation. The algorithm derives as many missing counts on links without counts (i.e. without detectors) as possible. This algorithm requires that all nodes of the graph either have only one input link and multiple output links or vice versa. As has been described in

paragraph 5.3.2 this property of the graph is guaranteed in this work. If a graph is used that does not fulfill this requirement from the beginning, a node with multiple input and output links as shown in figure 5-7 has to be decomposed into two nodes connected by a single link.

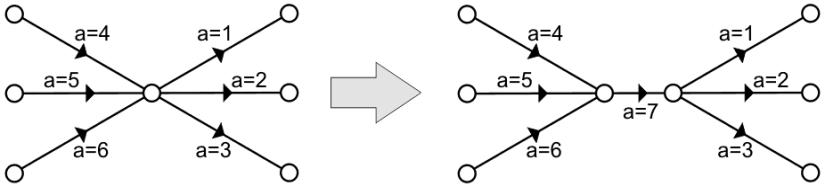


Figure 5-7: Decomposition of a multiple input – multiple output node

By going through the following steps, all derivable traffic volumes q_a on links without own counts are identified:

Step 1: All nodes having a single input link without count are stored in set N_1.

Step 2: All nodes having a single output link without count are stored in set N_2.

Step 3: For each node in N_1 the missing count on the input link is derived by summation of the counts of the output links if these are all available. In this case the node is removed from set N_1 and, if applicable, the start node of the input link from set N_2.

Step 4: For each node in N_2 the missing count on the output link is derived by summation of the counts of the input links if these are all available. In this case the node is removed from set N_2 and, if applicable, the end node of the output link from set N_1.

Step 5: All nodes having multiple input links, only one of which without count, are stored in set N_3.

Step 6: All nodes having multiple output links, only one of which without count, are stored in set N_4.

Step 7: For each node in N_3 the missing count on one of the input links is derived by subtraction of the available counts of the other input links from the count of the output link if the latter is also available. In this case, the start node of the link whose count has been derived is removed from set N_2 or N_4 if applicable.

Step 8: For each node in N_4 the missing count on one of the output links is derived by subtraction of the available counts of the other output links from the count of the input link if the latter is also available. In this case, the end node of the link whose count has been derived is removed from set N_1 if applicable.

Step 9: If no (further) count can be derived in steps 3 to 8, the algorithm stops. Otherwise it returns to step 3.

Figure 5-8 shows an example that applies the algorithm to a simple graph. The red links in the left-most image correspond to separate turning lanes or mixed lanes at a T-junction and its adjacent junctions respectively, all equipped with detectors and thus having counts. By applying the steps described above several times, all the missing traffic volumes including the desired

ones on four of the six turning links of the T-junction can be derived in the example. The link color in figure 5-8 changes from black over green when the count is derived to red.

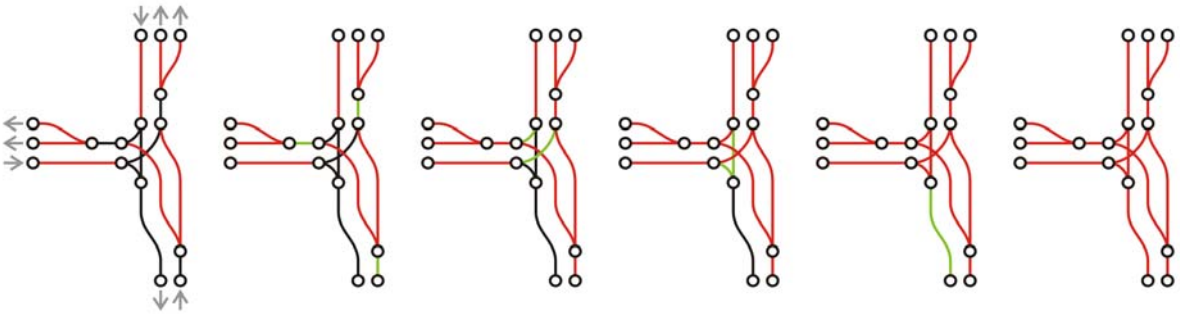


Figure 5-8: Sequence of several steps to derive missing counts

Figure 5-9 shows the links with available flow volumes (red links) after the algorithm has been applied to the graph of the List network. Comparison with figure 5-5 reveals that many additional link volumes could be derived. Especially at intersections where some of the turnings do not exist the algorithm is very successful. At some intersections at the borders of the network as well as at the two uncontrolled intersections some turning flows are still missing.

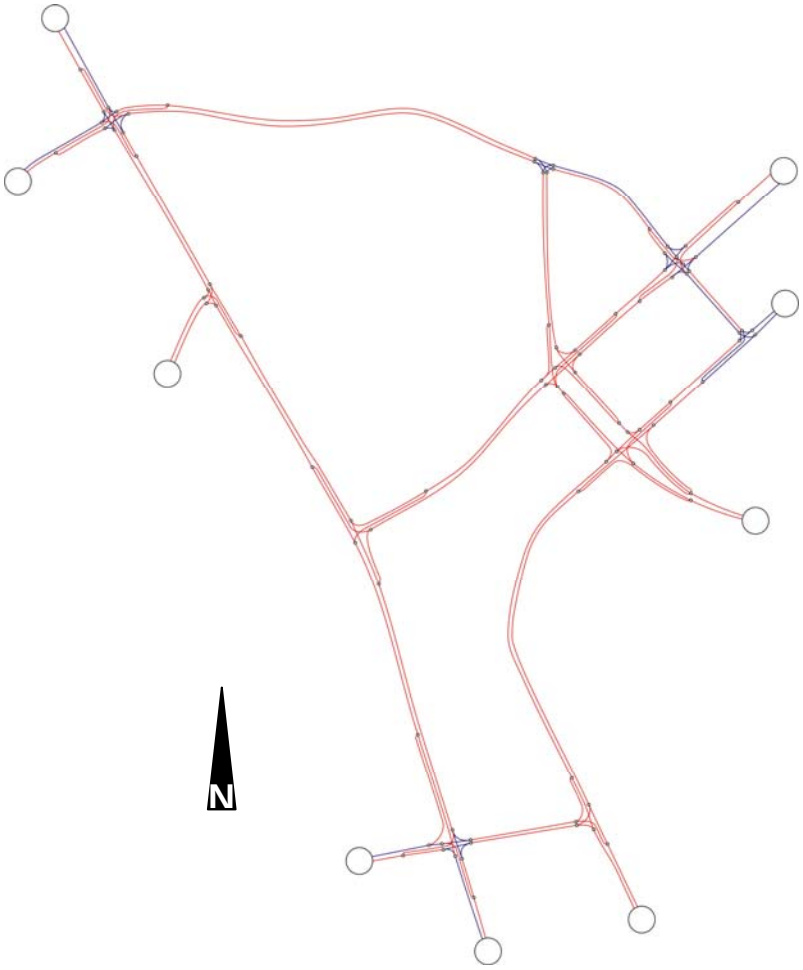


Figure 5-9: Available link flow volumes after improvement of constraints

It has to be emphasized that while improving the constraints the algorithm introduces a lot of redundant constraints that will be dealt with in paragraph 5.3.6.

5.3.5 Elimination of inconsistent constraints

Flow conservation requires that no vehicles can get lost in a network. If for example all inflows and outflows are detected (or derived) at an intersection, the sum of inflows must be equal to the sum of outflows. However, loop detectors (and other detector types as well) are known to produce counts with minor or sometimes even major errors (cf. LEHNHOFF, 2005). Furthermore, the forecasting technique used in this thesis is afflicted with some unavoidable imprecision. It is therefore very likely that the forecasted detector counts and the additionally derived counts, which are used as constraints for the OD matrix estimation, do not satisfy flow conservation. They are inconsistent to some extent.

The IM model needs traffic counts close to consistency for a good estimation of an OD matrix. Otherwise no solution can be found that satisfies all constraints in the form of equation 5-5. VAN ZUYLEN/BRANSTON (1982) developed a method to overcome inconsistencies in traffic counts. An iterative algorithm balances inconsistent traffic counts and finds a maximum likelihood estimate of corrected counts.

For each node *i* whose inflows and outflows q_a are all fully known VAN ZUYLEN/BRANSTON (1982) form the following constraint:

$$\sum_a A_{ai} \cdot q_a = 0 \qquad (5\text{-}7)$$

where:

A_{ai} coefficient relating links *a* to nodes *i* [-]

The coefficients A_{ai} relate each link *a* with count to each node *i*. If link *a* is an incoming link at node *i*, A_{ai} is set to 1. If it is an outgoing link, A_{ai} is set to -1. If link *a* is not connected to node *i* at all, A_{ai} is 0. Equation 5-7 expresses flow conservation at each node *i*. Nodes where no counts are available for one or more of the inflows and outflows are not included in these constraints.

If the network is represented as a graph where nodes correspond to intersections, no further thoughts on these nodes are necessary. However, in this thesis each intersection is decomposed into several links and nodes of a directed graph so that individual turnings can be modeled explicitly. Therefore, instead of considering single nodes *i* one has to consider the more general case of sets of nodes because it is not relevant whether all incoming and outgoing links with count are connected to the same node or not. Instead it is relevant that a set of internally connected nodes can only be entered or exited on incoming or outgoing links with counts. Figure 5-10 illustrates this concept. All red links have counts. All nodes that are connected by black links form a set of nodes. All inflows and outflows of this set are known. Therefore, instead of considering nodes *i* as VAN ZUYLEN/BRANSTON (1982) did, sets *i* of nodes are considered in this work. It is understood that a set of nodes can contain a single node only.

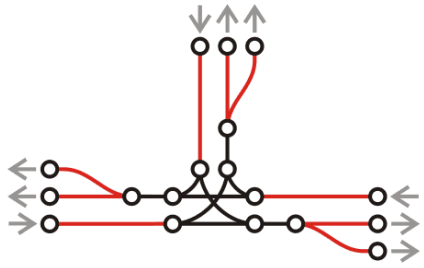


Figure 5-10: Concept of set of nodes

Thus, all sets *i* of nodes have to be identified whose inflows and outflows q_a are known, either because the relevant incoming and outgoing links *a* are equipped with a detector and thus have a forecasted count or because they contain a derived flow. An algorithm that is very similar to the one presented in paragraph 5.3.4 can be used for this purpose. It identifies all relevant sets of nodes and determines the according coefficients A_{ai}.

VAN ZUYLEN/BRANSTON (1982) assume that all counts q_a are Poisson distributed, i.e. the probability of observing $\hat{q}_a$ vehicles in one hour on one link *a* is given by equation 5-8. The probability of observing a set of flows $\{\hat{q}_a\}$ in a network is given by equation 5-9.

$$P(\hat{q}_a) = \frac{q_a^{\hat{q}_a}}{\hat{q}_a!} \cdot e^{-q_a} \qquad (5\text{-}8)$$

$$P(\{\hat{q}_a\}) = \prod_a \frac{q_a^{\hat{q}_a}}{\hat{q}_a!} \cdot e^{-q_a} \qquad (5\text{-}9)$$

where:

q_a expected number of vehicles in a time interval [veh/h]

$\hat{q}_a$ observed number of vehicles in a time interval [veh/h]

The idea of the method is to maximize equation 5-9 subject to the constraints in the form of equation 5-7 in order to find the most likely set of consistent flows $\{q_a\}$. The inconsistent forecasted or derived counts are the observed counts $\hat{q}_a$.

VAN ZUYLEN/BRANSTON (1982) use the method of Lagrange multipliers to find the maximum of equation 5-9 subject to the constraints. All intermediate steps have been reconstructed in appendix C. They lead to the following solution:

$$q_a = \hat{q}_a \Big/ \left(1 + \sum_i \lambda_i \cdot A_{ai}\right) \qquad (5\text{-}10)$$

where:

λ_i Lagrange multiplier of set *i* of nodes or constraint *i* respectively [-]

VAN ZUYLEN/BRANSTON (1982) developed an iterative algorithm to find a consistent solution $\{q_a\}$. The algorithm finds estimates $\tilde{\lambda}_i$ of the Lagrange multipliers, so that the differences δ_i between the real λ_i and these estimates are minimized. The relation between these three variables is expressed by equation 5-11.

$$\lambda_i = \tilde{\lambda}_i + \delta_i \qquad (5\text{-}11)$$

where:

$\tilde{\lambda}_i$ estimate of the Lagrange multiplier λ_i [-]

δ_i difference between real λ_i and its estimate [-]

The Taylor series of equation 5-10 about $\{\tilde{\lambda}_i\}$ is given in equation 5-12. Terms of second and higher order have been neglected. Details on the transformation can be found in appendix C.

$$q_a \simeq \hat{q}_a \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai}\right) - \hat{q}_a \cdot \sum_i \delta_i \cdot A_{ai} \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai}\right)^2 \qquad (5\text{-}12)$$

Substituting equation 5-12 in equation 5-7 for each set *i* of nodes leads to a set of *i* equations 5-13 (cf. appendix C for intermediate steps).

$$E_i = \sum_j B_{ij} \cdot \delta_j \qquad (5\text{-}13)$$

where:

$$E_i = \sum_a A_{ai} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak}\right) \qquad (5\text{-}14)$$

$$B_{ij} = \sum_a A_{ai} \cdot A_{aj} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak}\right)^2 \qquad (5\text{-}15)$$

E_i is equal to the difference of total inflow and outflow of a specific set *i* of nodes, given the current estimates $\tilde{\lambda}_i$. All E_i should therefore be preferably small. An acceptable threshold has to be defined, e.g. e = ±1 veh/h.

If a vector E, a matrix B and another vector δ are defined which comprise all E_i, B_{ij} and δ_i respectively, the set of linear equations 5-13 can be written as:

$$E = B \cdot \delta \qquad (5\text{-}16)$$

The iterative algorithm which produces consistent counts has the following steps:

Step 1: Set iteration counter *k* to zero.

Step 2: Set all $\tilde{\lambda}_i^{k=0}$ to zero.

Step 3: Determine all current E_i and B_{ij} according to equations 5-14 and 5-15. If all E_i are smaller than threshold e, go to step 7, else go to step 4.

Step 4: Determine all current δ_i^k by solving equation 5-16, i.e. $\delta = B^{-1} \cdot E$.

Step 5: Adapt all $\tilde{\lambda}_i^k$ for the next iteration k+1 according to $\tilde{\lambda}_i^{k+1} = \tilde{\lambda}_i^k + \delta_i^k$.

Step 6: Increment *k* and go to step 3.

Step 7: Determine q_a according to equation 5-10 using the last estimate $\tilde{\lambda}_i^k$ of the real λ_i.

The algorithm converges very quickly. Only two to three iterations are needed in general even with small thresholds *e* and major inconsistencies to balance.

A final remark has to be made on the prerequisite of the algorithm that all observed counts $\hat{q}_a$ must not be smaller than zero. Counts on links with detectors automatically fulfill this requirement. Due to detector malfunctions and imprecise forecasted counts, however, some additionally derived counts might be smaller than zero. Such values have to be set to zero or a small value such as 1 veh/h before starting the elimination of inconsistencies.

5.3.6 Elimination of redundant constraints

VAN ZUYLEN (1981) described a drawback of the IM model that originates from redundant information. He showed that the estimated flows between the four possible OD pairs in figure 5-11 vary depending on the link counts that are used for the estimation. Given that all links have counts and are used as constraints, a first solution is found. One would expect the result of the estimation to be the same if one of the four links is not used as constraint. The available information is not reduced in this case because the information on the missing link (no matter which one of the four is chosen) is a linear combination of the information on the remaining links. However, the result of the estimation differs depending on the omitted link information. Thus, the IM model requires that any information used as constraint is independent of any other used information. The reason for this property has been discussed in appendix B. It is a direct consequence of the derivation of the IM model. VAN ZUYLEN (1981) proposed a modification of the original IM model to deal with this effect.

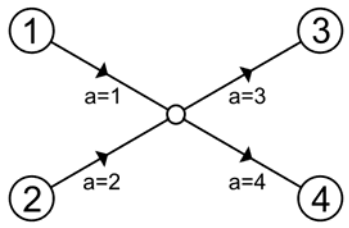


Figure 5-11: Example network used by VAN ZUYLEN (1981)

WANG (2008) showed that this problem can also be dealt with by using not only the portions p_{ij}^a of the constraining links *a* when calculating g_{ij} in equation 5-6 but those of *all* links in the network. If done so, the model always produces the same stable estimation result, whether all four links are used as constraints or one of the four links is omitted. This is because the exponent p_{ij}^a / g_{ij} better reflects the real weight of a constraining count if g_{ij} considers all links in the network. This approach turned out to perform slightly better than the modified IM model, especially when the p_{ij}^a are not known exactly but only estimated, which is the case in reality.

However, completely redundant information should still be avoided. According to WANG (2008) completely redundant information arises for example from a link that splits up into several turning links at a node. If the information on the first link is available and used as constraint, as well as the information on several or all of the turning links, the latter are completely redundant and thus this redundant information has a higher weight during the estimation process compared to other, non-redundant information. This is not justified and leads to a reduced quality of the

estimation result. However, the information on turning links is more valuable for the estimation process because it is more precise as has been shown in paragraph 5.3.4. The approaching link only contains a sort of accumulated information that is specified with more detail on the turning links. Therefore, WANG (2008) defined several rules in order to eliminate completely redundant information while maintaining the most precise information as possible. These rules consider different types and combinations of nodes with a varying number and arrangement of connected links. Figure 5-12 shows three examples of the most relevant structures considered by WANG (2008).

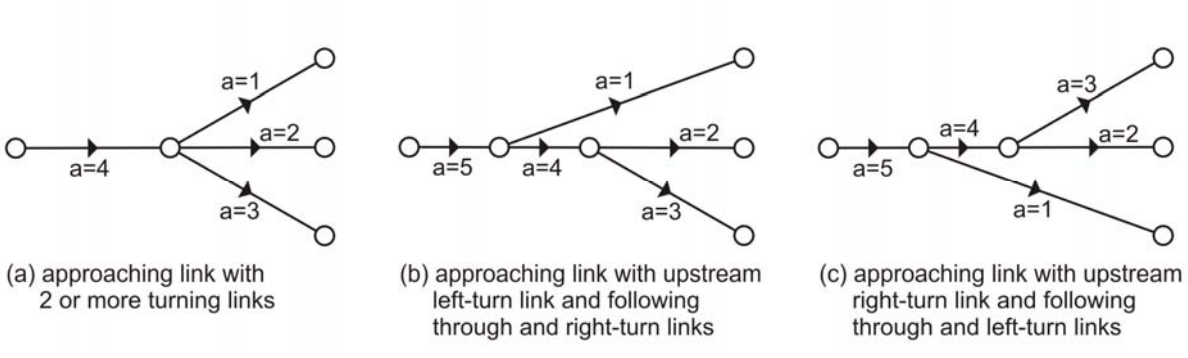


Figure 5-12: Some of the structures used for redundancy elimination rules by WANG (2008)

The information contained on each link are the portions p_{ij}^a and, if available, the count q_a. The principle of elimination of completely redundant information on approaching links is to subtract more precise information on turning links from the more general, accumulated information on the approaching link if counts on the turning links are available. In case (a) depicted in figure 5-12 the p_{ij}^a and q_a of the three turning links 1, 2, and 3 with available counts are subtracted from the p_{ij}^a and q_a on link 4. If link 4 does not have a count, only the p_{ij}^a are subtracted. If one of the turning links does not have a count, its p_{ij}^a are set to zero and are not subtracted from link 4. In other words, the most precise information which is the information on turning links is kept if the turning links have a count that can be used as constraint. In this case the information of these downstream links has to be reduced from the upstream approaching link. If a turning link has no count, it is eliminated by setting its portions p_{ij}^a to zero. The portions are not subtracted from the approaching link.

WANG (2008) defined equivalent rules for cases (b) and (c) and other link structures that may be relevant in urban networks. Some rules are also used in upstream direction. Applying the set of rules to the available information on all links of the graph before starting the IM model assures that no completely redundant information is used as constraint. However, the rules have not been generalized to cover arbitrary link structures. Therefore, an algorithm has been developed in this thesis that covers all imaginable link structures without the need for adapted rules for different structures.

Each link in the network has to be checked for containing accumulated information in form of p_{ij}^a and, if available, q_a that is also contained with more precision on other links. For this task, two sets of links are used. Set L_1 contains all links that are single input links at their end node and

set L_2 contains all links that are single output links at their start node. The graph as presented in paragraph 5.3.2 ensures that each of its links is in at least one of the two sets. Starting with set L_1, all single input links are successively checked for redundancy in downstream direction. For illustration, figure 5-13a shows an example in which link 7 is such a link. Using this link as root, a tree of links is generated by using a recursive tree spanning algorithm. The branches of the tree end at links that fulfill one or more of the following criteria:

(a) They have a detected or derived traffic volume (red links 1 and 2 in figure 5-13a).

(b) They are one of multiple input links at their end node (link 3).

(c) They end at a destination (link 6).

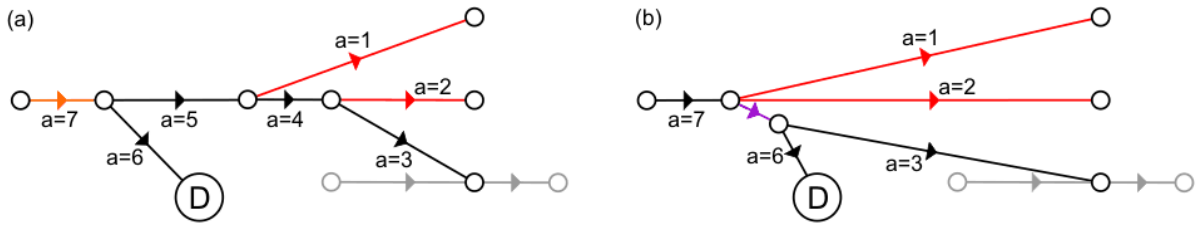


Figure 5-13: Example node structure to illustrate redundancy elimination

Thus, all traffic on any link of the tree must inevitably also have passed the root link. If a detected or derived count exists on the root link, the counts on the red links are redundant but more precise at the same time. Consequently, the latter are kept and the count on the root link has to be reduced by the counts on the red links. The same has to be done with the p_{ij}^a. If no count is available on the root link, only its p_{ij}^a are reduced by those on the red links. This is relevant in order to calculate the correct g_{ij}, which must not contain completely redundant p_{ij}^a either, even though a root link without count itself will not be used as a constraint for the IM model.

The remaining information on the root link after redundancy elimination can be interpreted as the information on a virtual link connected to the end links of the tree that have no counts. In figure 5-13b the virtual link in the example is drawn in purple and contains the accumulated information of links 3 and 6. If all branches of the spanned tree end with links with counts, the information on the root link is reduced completely and no virtual link exists.

The information on those links in the spanned tree that do not have counts (links 3 to 6 in the example) is eliminated and these links are removed from set L_1 because they do not have to be checked for redundancy again. If used as root links themselves, the links in the corresponding trees would only be a subset of the links in the current tree. If one of these links has been checked as root link before, the information will still be deleted because its p_{ij}^a are contained in the root link of the new, larger tree. Only the information on the links with counts and the reduced information on the new root link are kept. The links with counts remain in set L_1 because they might themselves be followed by branching links with more precise counts and their information might have to be reduced as well.

After all links in set L_1 have either been checked or removed, all links in set L_2 are checked or removed following the same principle, but in upstream direction. Links that are contained in both sets are checked for redundancy in both directions, which might result in two independent virtual links.

The described procedure of redundancy elimination covers all link structures considered by WANG (2008) and any other imaginable arrangement of links. The red links in figure 5-14 are used as constraining links in the test network after the algorithm for redundancy elimination has been applied to the graph. Comparison with figure 5-9 shows that the information on completely redundant links has been discarded. Comparison with figure 5-5 shows that the constraining links comprise more turning links than the originally detected links, whereas some approaching links with accumulated information are not used as constraints in favor of turning links whose counts have been derived additionally.

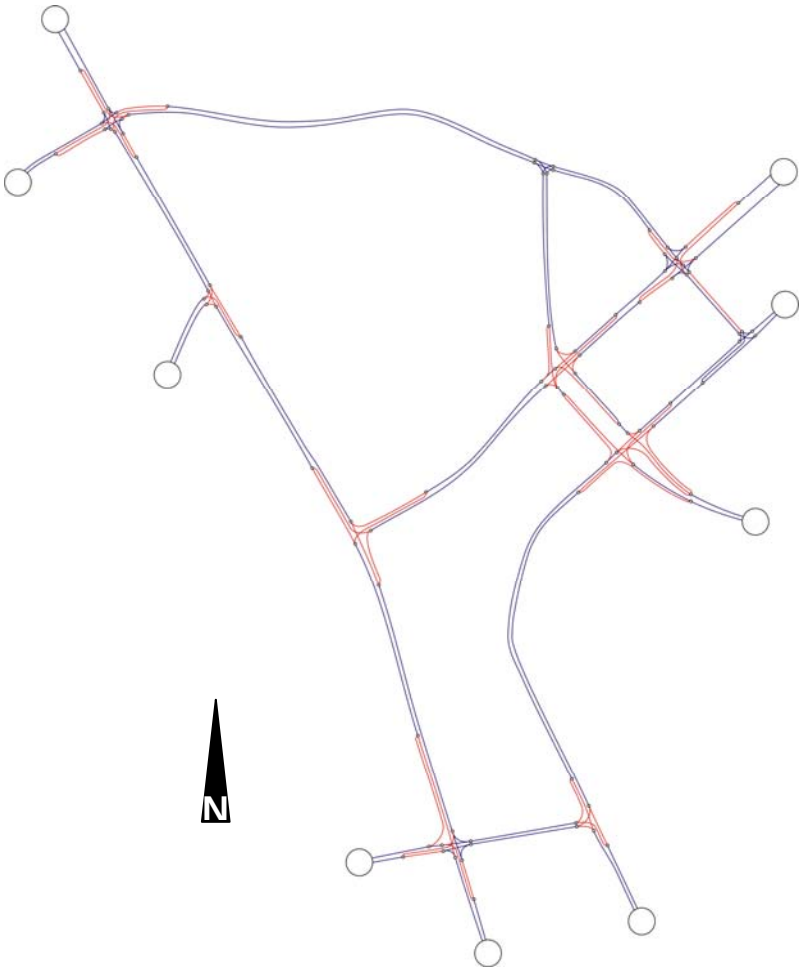


Figure 5-14: Links used as constraints after elimination of complete redundancies

At the northwestern intersection two virtual links have been generated. As in the case of derived counts it is also possible that the remaining counts on such virtual links are negative due to faulty detector measurements and/or imprecise forecasted counts. Such negative values have to be corrected to zero or even small positive values and the counts on virtual links have to be included into the elimination of inconsistencies. Therefore the algorithm presented in paragraph 5.3.5 has been programmed in such a way that it takes also virtual links into account.

5.3.7 Integration of traffic assignment techniques

5.3.7.1 Overview

As has been described in paragraph 5.3.3 the IM model requires that the portions p_{ij}^a for each link *a* and each OD pair *ij* are known, i.e. an underlying static assignment of traffic to the network which is independent of the final OD matrix is assumed. In reality however, the exact p_{ij}^a are unknown and have to be estimated. Different traffic assignment techniques are available for this purpose. Such techniques assign given OD matrices to a network, i.e. they estimate how traffic flows from origins *i* to destinations *j* disperse to the available alternative routes. They result in choice percentages p_{ij}^k of all alternative routes *k* between the OD pairs *ij*. A common assumption is that all drivers choose their routes in such a way that no time can be gained by using an alternative route, i.e. travel times on all alternative routes are equal or higher. This is the user equilibrium according to the first principle of WARDROP (1952).

WANG (2008) tried different traffic assignment techniques and combined them with the IM model in an iterative procedure. The idea is to start with a given historic matrix or a unit matrix in the first step and to assign it to the network by using a suitable assignment technique. This assignment results in choice percentages p_{ij}^k of different routes *k*. Since each link *a* belongs to one or more routes *k* the necessary portions p_{ij}^a can be easily deduced by summation of the respective p_{ij}^k of routes *k* that link *a* belongs to. A new matrix is estimated by using the IM model and the deduced p_{ij}^a. This matrix is assigned to the network again to get new p_{ij}^a which can then be used to estimate a new matrix and so on. The process of repeated traffic assignment and matrix estimation is executed until the estimated matrix converges on a stable solution.

Two of the three assignment techniques tested by WANG (2008) led to a stable OD matrix. These techniques are Successive assignment and SUE assignment. In accordance with these findings both methods have also been implemented and tested in this thesis. The assignment techniques will be discussed in the next two sections.

During the assignment process travel times t_k on different routes *k* have to be estimated repeatedly. They depend on the travel times $t_a(q_a)$ of their according links *a* which are a function of the current traffic load q_a on the links. Such a function will be presented after both assignment techniques have been described, together with an alternative approach to estimate travel times on routes *k* based on the CTM.

5.3.7.2 Successive assignment

Successive assignment is a rather simple assignment technique that is easy to implement. The total traffic demand in the form of an OD matrix is split up into *n* slices. Often equally sized slices are used, i.e. each slice contains the n^{th} part of the matrix. The slices are assigned to the network in *n* steps. During each step, the current slice of flow f_{ij} is assigned to the currently shortest route *k* given the current traffic load that has already been assigned to the network in previous steps. After *n* steps the whole matrix will have been assigned to the network.

The algorithm has the following steps:

Step 1: Determine all flows $f_{ij,n}$ of one slice by dividing the according total flows f_{ij} from the OD matrix by the chosen number n of slices.

Step 2: Determine the travel times t_k of all routes k for an empty network, i.e. free flow travel times.

Step 3: Assign each flow $f_{ij,n}$ from the current slice to the route k from i to j that currently has the shortest travel time t_k. If no further slice is left, finish the assignment; else go to the next step.

Step 4: Determine the travel times t_k on all routes k subject to the current traffic loads on all links in the network and go to step 3.

After the assignment all routes k have a specific traffic load. The choice probability p_{ij}^k of a route is the ratio of the load on route k and the total flow f_{ij} from the OD matrix.

5.3.7.3 Stochastic User Equilibrium assignment

The theory behind SUE assignment is more sophisticated. SUE factors the drivers' perception of travel times on different routes into their route choice, i.e. drivers do not necessarily choose the objectively shortest route but the one that they *perceive* to be the shortest. Thus the first principle of WARDROP (1952) is changed such that SUE is reached if no driver *believes* to get to his destination faster on a different route than the one that he has chosen.

Multinomial Logit Model

The Multinomial Logit (MNL) model (cf. SHEFFI, 1985) can be used to estimate the choice probabilities of different alternative routes under the assumption of SUE. The word "multinomial" refers to the fact that there can be more than two different discrete options to choose from (i.e. the routes in the case of traffic assignment) and that these options do not have a meaningful order. The word "logit" means that the logit formula is used to determine choice probabilities. If there are only two options k and l to choose from, the logit formula expresses the probability p_k that option k is chosen.

$$p_k = \frac{1}{1 + e^{(U_l - U_k)}} \qquad (5\text{-}17)$$

where:

U_k perceived utility of option k [-]

U_l perceived utility of option l [-]

The higher the perceived utility U_k, the more likely is the choice of option k. U_k has a deterministic component V_k that can be measured objectively and an additive random "error" ξ_k with expected value $E[\xi_k] = 0$, so that

$$U_k = V_k + \xi_k \qquad (5\text{-}18)$$

where:

V_k objective utility of option k [-]

ξ_k random "error term" of option k [-]

This random component incorporates subjective aspects of the user which influence his choice but cannot be measured directly. SHEFFI (1985) states that the ξ_k are "independently and identi-

cally distributed Gumbel variates" so that they do not influence choice probabilities in equation 5-17. The choice probability p_k of an option k out of two options k and l is thus:

$$p_k = \frac{1}{1 + e^{(V_l - V_k)}} \qquad (5\text{-}19)$$

where:

V_k deterministic utility of option k [-]

V_l deterministic utility of option l [-]

Figure 5-15 shows the choice probabilities of two options k and l according to equation 5-19. Option k will be clearly preferred over option l if the difference of V_l and V_k is negative, i.e. if V_k is greater than V_l. If V_k equals V_l both choice probabilities are 0.5.

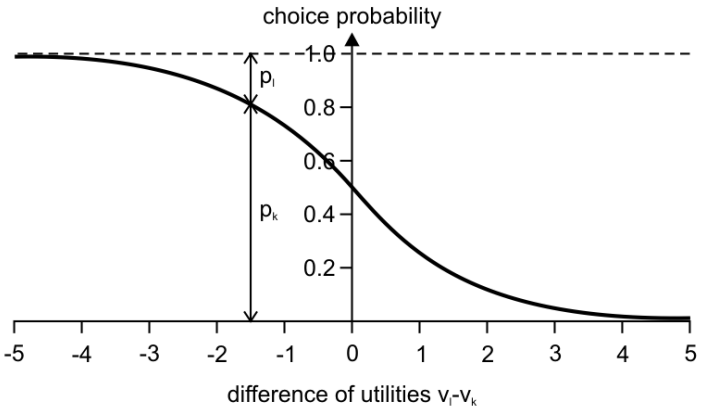


Figure 5-15: Choice probabilities of two options *k* and *l* according to the logit formula

If there are more than two options equation 5-19 can be generalized to

$$p_k = \frac{1}{1 + \sum_{\substack{l=1 \\ l \neq k}}^{K} e^{(V_l - V_k)}} \qquad (5\text{-}20)$$

where:

K number of different options [-]

Figure 5-16 shows three examples of choice probabilities of three different options k, l, and m according to equation 5-20.

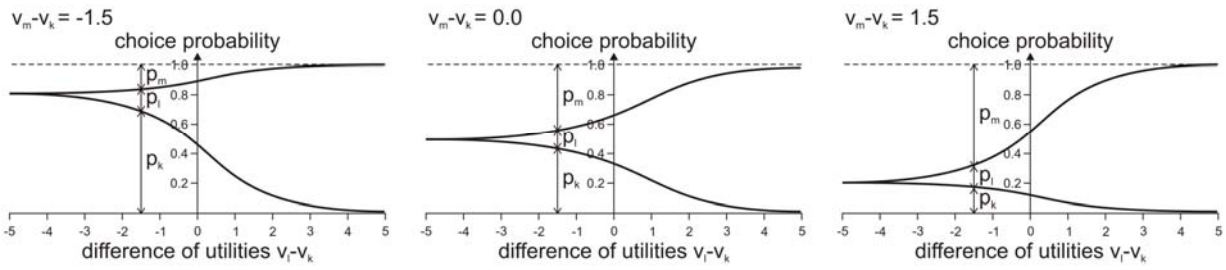


Figure 5-16: Choice probabilities of three options *k*, *l*, and *m* according to the logit formula

Equation 5-20 can be transformed as follows:

$$p_k = \frac{1}{1+\sum_{\substack{l=1 \\ l\neq k}}^{K} e^{(V_l - V_k)}} = \frac{1}{\sum_{l=1}^{K} e^{(V_l - V_k)}} = \frac{e^{V_k}}{e^{V_k} \cdot \sum_{l=1}^{K} e^{(V_l - V_k)}} = \frac{e^{V_k}}{\sum_{l=1}^{K} e^{V_l}} \qquad (5\text{-}21)$$

In the case of route choice the utility of route k is its current travel time t_k. Contrary to the previous argumentation a route has a higher utility for the drivers if its travel time is small. Therefore equation 5-21 has to be adapted slightly. The probability to choose route k with travel time t_k out of a set of available routes K is:

$$p_k = \frac{e^{-\theta \cdot t_k}}{\sum_{l=1}^{K} e^{-\theta \cdot t_l}} \qquad (5\text{-}22)$$

where:

t_k	travel time of route k [h, min, or s]
θ	unit scaling factor [1/h, 1/min, or 1/s]

The unit scaling factor θ transforms travel time into a utility without unit. Furthermore it allows free choice of the preferred unit of travel times. It can be adapted in such a way that p_k is independent of the used time unit. And finally, θ is an important parameter to calibrate the logit model. A high value increases the drivers' sensitivity to differences in travel times and a small value reduces their sensitivity.

C-Logit model

Equation 5-22 works well if all alternative routes are completely independent of each other as depicted in figure 5-17a. If travel times are the same on all three routes, choice probabilities p_1, p_2, and p_3 will be the same, too. In case (b) however, routes 2 and 3 overlap significantly. Equation 5-22 will still assign same choice probabilities to all three routes even though one would expect route 1 to be the more heavily loaded. Routes 2 and 3 share a stretch and are thus likely to have smaller choice probabilities each. These probabilities will sum up to a total choice probability with a magnitude comparable to p_1.

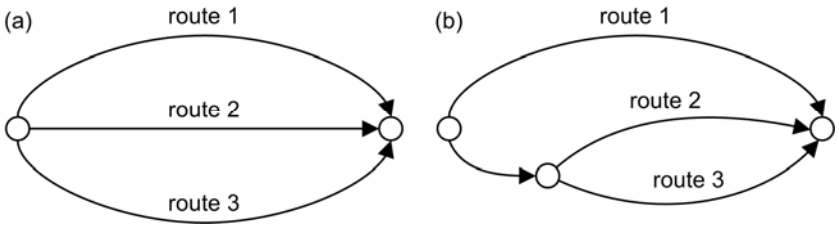


Figure 5-17: Entirely independent against overlapping routes

CASCETTA ET AL. (1996) introduced a Communality Factor (*CF*) to take the effects of overlapping routes on route choice probability into account. *CF* expresses similarity of different routes. It has to be determined for each route k out of K alternative routes and is defined as follows:

$$CF_k = \beta \cdot \ln \sum_{g=1}^{K} \left(\frac{LG_{g,k}}{\sqrt{LG_g \cdot LG_k}} \right)^{\gamma} \quad (5\text{-}23)$$

where:

LG_g	measure of links belonging to route g [h, min, s, or m]
LG_k	measure of links belonging to route k [h, min, s, or m]
$LG_{g,k}$	measure of links belonging to both routes g and k [h, min, s, or m]
γ	calibration parameter [-]
β	unit scaling factor [h, min, or s]

LG is the sum of specific measures of all links belonging to route g, k, or both. The measures can be link travel times, link distances or combinations thereof. If free flow travel times on links a are chosen (which is the case in this thesis), all CF_k have to be determined only once and are constant during each assignment process.

If two routes g and k overlap heavily the bracket term in equation 5-23 gets close to 1. If a route is completely independent of other routes, the sum in this equation is 1 and CF_k is thus zero. The more it overlaps with other routes, the higher CF_k will be.

A calibration parameter γ smaller than 1.0 attaches higher importance to even small degrees of congruence, whereas overlapping has to be more distinct with greater values of γ to have the same influence on CF_k.

The greater CF_k, the more the utility of an alternative route is reduced. To consider overlapping, equation 5-22 has to be modified to

$$p_k = \frac{e^{\theta \cdot (-t_k - CF_k)}}{\sum_{l=1}^{K} e^{\theta \cdot (-t_l - CF_l)}}. \quad (5\text{-}24)$$

Since CF_k is a dimensionless measure, the unit scaling factor β in equation 5-23 allows adaptation of CF_k to the unit of travel time t_k in equation 5-24. Moreover, β specifies the amount to which CF_k worsens the original utility of a route in terms of mere travel time t_k. If β is high, the choice probability of highly overlapping routes is reduced to a greater extent than in the case of a low β.

Method of Successive Averages

Route choice probabilities depend on travel times t_k which themselves depend on the current traffic load q_a on links a of route k. Therefore, a solution that establishes SUE has to be determined in an iterative procedure. WANG (2008) used the Method of Successive Averages (MSA) as described by SHEFFI (1985). During the first iteration $n = 0$ an initial set of traffic loads $\left\{ q_k^{n=0} \right\}$ on all routes k is determined based on free flow travel times t_k. Then all travel times have to be updated according to the new traffic loads q_a^n on all links a. Iteration counter n is incremented and a set of auxiliary traffic loads $\left\{ q'^n_k \right\}$ is derived based on the updated travel times. The solution of the previous iteration is adapted according to

$$q_k^n = \left(1 - \frac{1}{n}\right) \cdot q_k^{n-1} + \frac{1}{n} \cdot q'^n_k = q_k^{n-1} + \frac{1}{n} \cdot \left(q'^n_k - q_k^{n-1}\right) \quad (5\text{-}25)$$

where:

n	iteration counter [-]
q_k^n	traffic load on route k at iteration n [veh/h]
q'^n_k	auxiliary traffic load on route k at iteration n [veh/h]

Equation 5-25 shows that the solution in step n is the average of the solution of step n-1 and the auxiliary solution of step n. The weighting coefficient 1/n reduces the weight of the auxiliary solution with increasing number n of iterations and thus assures that MSA converges on a stable solution.

SUE Algorithm

All the previous aspects and explanations can be turned into one algorithm that assigns a given OD matrix to different routes in a network under the assumption of SUE. The steps are:

Step 1: Set iteration counter n to zero.

Step 2: Determine free flow travel times $t_{k,ij}$ of all available routes k of each OD pair ij for an empty network and the according choice probabilities p_{ij}^k of these routes according to equation 5-24 subject to these free flow travel times and the constant $CF_{k,ij}$.

Step 3: Determine traffic loads $q_{k,ij}^{n=0} = f_{ij} \cdot p_{ij}^k$ of all routes k of each OD pair ij with total flow f_{ij} and the resulting traffic loads $q_a^{n=0}$ of all links a.

Step 4: Increment iteration counter n.

Step 5: Determine the current link travel times $t_a(q_a^{n-1})$ of all links a, update all travel times $t_{k,ij}$ of routes by summation of travel times of the respective links, and determine new choice probabilities p_{ij}^k according to equation 5-24.

Step 6: Determine auxiliary traffic loads $q'^n_{k,ij} = f_{ij} \cdot p_{ij}^k$ of all routes k and update the current traffic loads $q_{k,ij}^n$ according to equation 5-25. If all changes of traffic loads on routes k are smaller than a given threshold, e.g. 1 veh/h, finish the assignment; else go to the next step.

Step 7: Derive the current traffic loads q_a^n on all links a and go to step 4.

Again, the final choice probabilities p_{ij}^k of routes k can be derived from the ratio of the final load on route k and the total flow f_{ij} from the OD matrix.

5.3.7.4 Estimation of travel times

At each step of both assignment algorithms travel times t_k on all alternative routes have to be estimated. They depend on the according link travel times $t_a(q_a)$ as a function of traffic loads on the link. WANG (2008) used the well known BPR function (proposed by the U.S. Bureau of Pub-

lic Roads – BPR) which is described in the Highway Capacity Manual – HCM (TRB, 2000). It is a common volume delay function of the following form:

$$t_a(q_a) = t_0 \cdot \left(1 + \alpha \cdot \left(\frac{q_a}{C_a}\right)^{\beta}\right) \qquad (5\text{-}26)$$

where:

q_a	traffic load on link *a* [veh/h]
t_0	free flow travel time on link *a* [h, min, or s]
C_a	capacity of link *a* [veh/h]
α, β	model parameter [-]

If there are no vehicles on link *a*, travel time t_a is equal to free flow travel time t_0. The closer the traffic load q_a gets to link capacity C_a, the higher is the travel time on this link. A reasonable estimation of link travel times depends on an adjusted choice of free flow travel time t_0, link capacity C_a and the parameters α and β.

An alternative to estimate travel times on routes that has been investigated in this thesis is to use the CTM. A method to derive travel times from the CTM has been presented in paragraph 4.4.2 and evaluated in paragraph 4.6. It has been assumed that the effect of signal timings on travel times can be taken into account in a better and more realistic way in comparison to static BPR functions since the implemented CTM model also includes signalization at each intersection.

If the CTM is used to estimate travel times during the process of traffic assignment, one run of the CTM has to be executed at each iteration step of both algorithms. At each step the CTM uses a traffic demand that corresponds to the current route and link traffic loads of the respective step. These traffic loads have to be transformed into source flows and turning portions that the CTM can work with.

According to the explanations in paragraph 4.4.2 the free flow travel time of a link or route corresponds to the number of cells multiplied by the duration of time step *T*. If a link has a traffic load of zero, its travel time stays unchanged from free flow travel time and the effects of possible traffic signals are unnoticed. A minimal traffic load on all links is needed to account for increases in travel time due to signalization.

This problem is solved by preloading all routes with a small traffic load of 0.1 veh/h before starting the assignment. In the case of Successive assignment some routes and thus links may stay unloaded during the whole assignment process. Therefore the preloaded flows can only be subtracted again after the assignment has been terminated. In the case of SUE assignment all routes will have at least a small share of traffic flows f_{ij} after the first iteration. The preloaded flows are only needed to determine free flow travel times and are subtracted immediately after this step.

5.3.8 Complete process of estimation

All the aspects described in the previous paragraphs have to be combined into one module that estimates OD matrices, route and link volumes based on a set of current detector counts of a specific time interval. Not all of the steps have to be executed repeatedly for each time interval.

Figure 5-18 shows the overall flow chart of traffic demand estimation. It will be explained in the following.

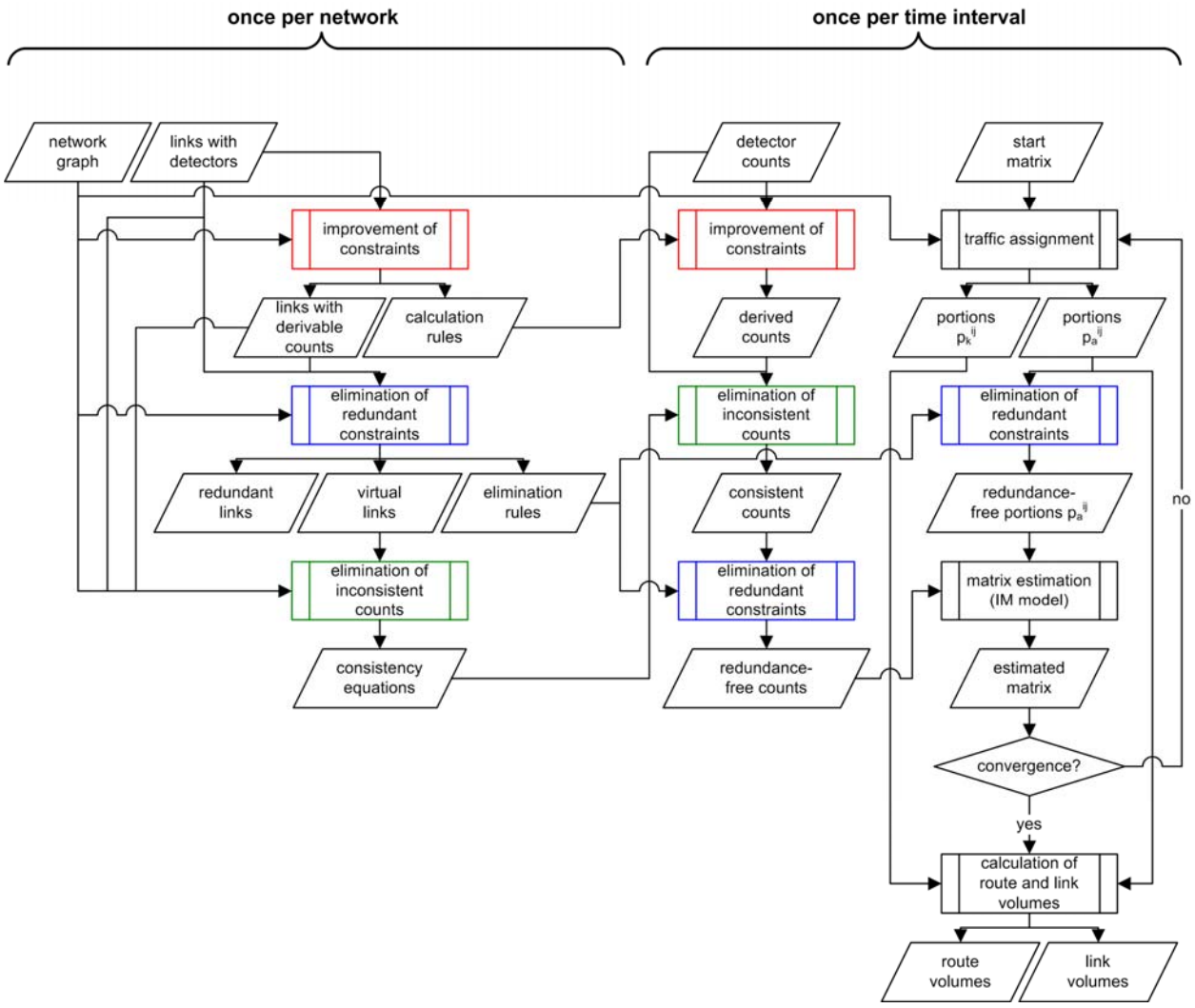


Figure 5-18: Flow chart of the complete process of traffic demand estimation

Some preparing steps have to be performed only once per network. Input data are the respective graph of the network and the knowledge of detector locations, i.e. which ones of the links are equipped with a detector. Based on this information the sub-module for improvement of constraints described in paragraph 5.3.4 can identify all links whose counts can be derived additionally. For each of these links the sub-module stores some calculation rules in the form of detected links that count positive and those that count negative when deriving the missing count.

In a next step the sub-module for elimination of redundant constraints presented in paragraph 5.3.6 uses the graph and the data sets on links with detected and derivable counts to generate redundancy elimination rules for each redundant link. The rules are stored for downstream and upstream direction respectively, if applicable. The rules are lists of links with available detected or derived counts whose information has to be subtracted from the respective redundant link. Besides these rules the sub-module also creates all necessary virtual links.

The last preparing step is to set up all consistency equations in the form of equation 5-7 in paragraph 5.3.5. This is done by the sub-module for elimination of inconsistent counts which takes detected links, links with derivable count and virtual links into account.

Once these preparing steps have been executed, OD matrices, route and links volumes can be estimated for each time interval. The detector counts of the current interval are fed to the sub-module for improvement of constraints that calculates all possible missing counts by using the previously determined calculation rules. Then the sub-module for elimination of inconsistent counts substitutes all available counts into the consistency equations and produces balanced, consistent counts. Finally, the sub-module for elimination of redundant counts uses the elimination rules to remove all redundancies.

The matrix estimation process starts with an initial matrix. This can be any given historic matrix or a unit matrix. This matrix is assigned to the network by either Successive or SUE assignment. Estimated route choice probabilities p_{ij}^{k} and the necessary portions p_{ij}^{a} of all links *a* are available as a result. The sub-module for elimination of redundant constraints transforms the latter into redundant free portions p_{ij}^{a} by means of the same elimination rules that have also been used to handle redundant counts.

Then the IM model estimates a new matrix based on the old one (i.e. the start matrix during the first iteration) and the redundancy-free constraints and p_{ij}^{a}. The new matrix is compared to the old one. If the f_{ij} of the new matrix differ by more than ±1 veh/h on average from the f_{ij} of the old matrix the new matrix is assigned to the network again, the redundant p_{ij}^{a} are eliminated and the matrix estimation is repeated. If the estimated matrix has converged sufficiently well, it is used as final solution. Based on the estimated matrix and the p_{ij}^{k} and p_{ij}^{a} of the last iteration the estimated traffic volumes on all routes and links can be calculated.

5.3.9 Evaluation

The performance of the described module for OD matrix, route and link volume estimation has been assessed in different steps that are described in the following. No forecasting has been considered in this stage of the evaluation.

5.3.9.1 Morning and afternoon peak interval

In a first step, estimated OD flows f_{ij} and route volumes q_k of the morning and afternoon peak intervals have been compared to the “real” flows f_{ij} and route volumes q_k as defined for both intervals (cf. paragraph 3.3.3). The detector counts serving as input for the estimation have not yet been taken from AIMSUN simulation runs. They have been calculated by assigning the respective OD matrices to the routes and links according to the chosen percentages of usage of the routes (cf. paragraph 3.3.3.)

For this initial test SUE assignment has been used with θ = 50/h and C-Logit parameters β = 0.05 h and γ = 1.0. These values have been chosen after several test runs of the module. Travel times on routes were estimated by means of the CTM, which used the real signal plans of the respective intervals during simulation. A unit matrix has been used as initial matrix to start with. These described settings are referred to as “standard settings” in the following.

Figure 5-19 shows the comparison of the "real" OD flows f_{ij} and the estimated f_{ij} for both morning and afternoon peak interval. Figure 5-20 compares route volumes q_k accordingly. Two cases have been considered: detectors either exist at all turning links or only at real detector locations. The latter is the more realistic case. Note that the unit of flows f_{ij} in figure 5-19 and of route volumes q_k in figure 5-20 is veh/h even though the duration of the time interval is only 15 minutes.

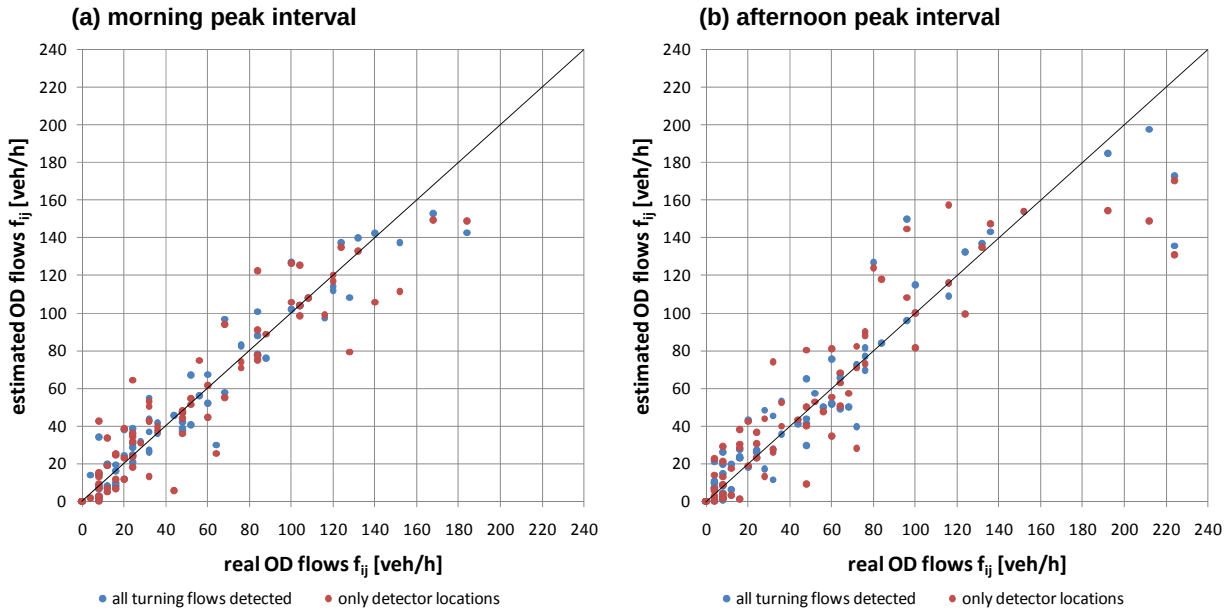


Figure 5-19: Results of OD matrix estimation for both peak intervals

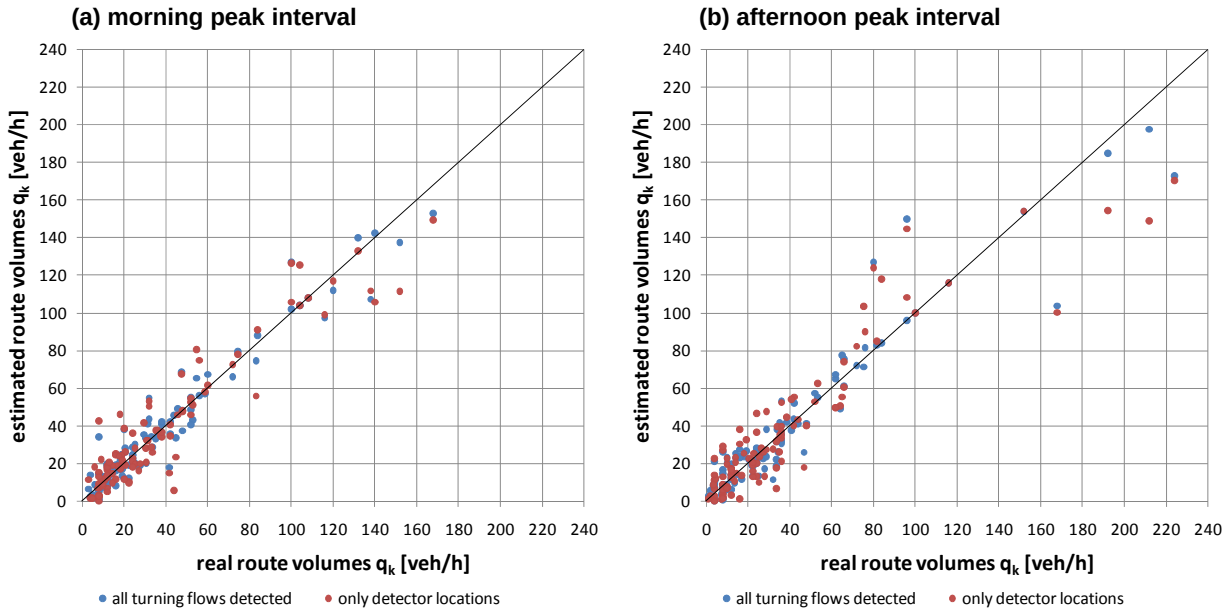


Figure 5-20: Results of route volume estimation for both peak intervals

Table 5-3 shows the achieved quality of estimation for both morning and afternoon peak interval, expressed by correlation coefficient r_{xy}, *RMSE* and *RRMSE*.

Table 5-3: Quality of OD matrix estimation for morning and afternoon peak interval

		morning peak interval		afternoon peak interval	
		all turning flows detected	only detector locations	all turning flows detected	only detector locations
OD flows	r_{xy} [-]	0.970	0.937	0.950	0.915
	RMSE [veh/h]	11.20	16.10	16.64	21.54
	RRMSE [-]	0.275	0.376	0.332	0.417
route volumes	r_{xy} [-]	0.972	0.942	0.954	0.923
	RMSE [veh/h]	8.483	12.141	12.630	16.226
	RRMSE [-]	0.281	0.381	0.339	0.424

As expected the quality of estimation decreases if counts are not available for all turning links. Even though many missing counts on other than the detected links can be derived, the information that can be used as constraints is still less precise. It can also be seen that not all f_{ij} and q_k are estimated precisely even for the case that all turning links are used as constraints. However, the good performance of the estimation is evident.

5.3.9.2 All time intervals

In a next step the evaluation of traffic demand estimation has been extended to all 56 consecutive time intervals of 15 minutes (cf. paragraph 3.3.3). This time, simulated data from 30 replications of the AIMSUN simulation has been used. For every time interval of every replication, all detector counts as well as the traffic volumes on each of the 108 routes and 146 links have been logged. Even though the same defined traffic demand has been fed to the network during each replication, detector counts and route and link volumes vary to some extent because of the random behavior of the microsimulator. This resulted in 1,680 different traffic demands and corresponding sets of detector counts (30 replications times 56 time intervals). For the estimation of travel times during SUE assignment the CTM used the signal plans that correspond to the respective intervals, i.e. morning signal plans for intervals before 1 pm and afternoon signal plans for intervals after 1 pm.

Figure 5-21 compares estimated traffic volumes on routes and links respectively to the “real” AIMSUN traffic volumes. The diagrams contain the data of all 56 time intervals of one example replication. The estimation was based on the logged detector counts of this replication. The standard settings as described before have been used.

It can be seen that the quality of estimation of traffic volumes on routes (r_{xy} = 0.889) does not guarantee a precise knowledge of the real current traffic demand on a specific route. However, to some extent it is still possible to judge from these estimates whether a specific route is currently among the most heavily loaded routes or not. This information can be used for the online optimization of coordination of traffic signals.

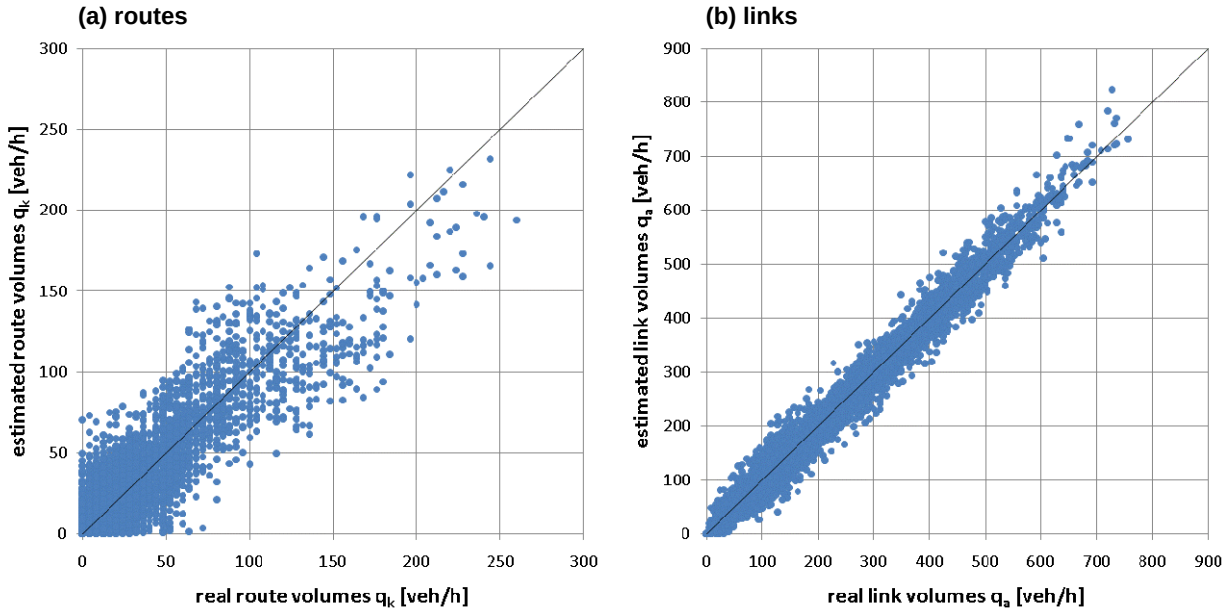


Figure 5-21: Estimation result for an example AIMSUN replication (all time intervals accumulated)

The second diagram in figure 5-21 shows that the estimation of traffic volumes on links, which result directly from the route volumes, is far more satisfying (r_{xy} = 0.991). This is because detected and derived link flows are used as constraints of the estimation, so that the finally estimated traffic demand tends to fulfill these constraints. The more redundancy-free counts on links can be used for the estimation, the better the estimated link volumes fit the real link volumes, even if the underlying OD matrix and route volumes are less accurate.

The link volumes are transformed into a traffic demand for the CTM later on in order to assess different offset combinations. Thus, a good reproduction of the current traffic situation in the traffic model based on the traffic demand estimation can be assumed.

While figure 5-21 visualizes the estimation results for only one single AIMSUN replication and does not differentiate time intervals, figure 5-22 and figure 5-23 illustrate individual and average performance of the estimation for all replications and all time intervals. The thick red lines represent the mean values of correlation coefficient r_{xy}, *RMSE*, and *RRMSE* of route and link volume estimation respectively. Mean values are given for each time interval and were calculated over all replications. The thin red lines correspond to the mean values plus/minus standard deviation. The gray curves in the background show the individual results for each replication. The axis of ordinates of corresponding diagrams in figure 5-22 and figure 5-23 have the same scale to facilitate direct comparison of the achieved quality of route and link volume estimation.

The mean correlation coefficient r_{xy} of route volume estimation has a rather stable value of 0.88 on average. Only before morning and after afternoon peak hour it is marginally lower. The mean *RMSE* is clearly proportional to the demand profile. It covers a range between 6.95 and 20.39 veh/h. The mean *RRMSE* is rather stable again with a value of 0.59 on average.

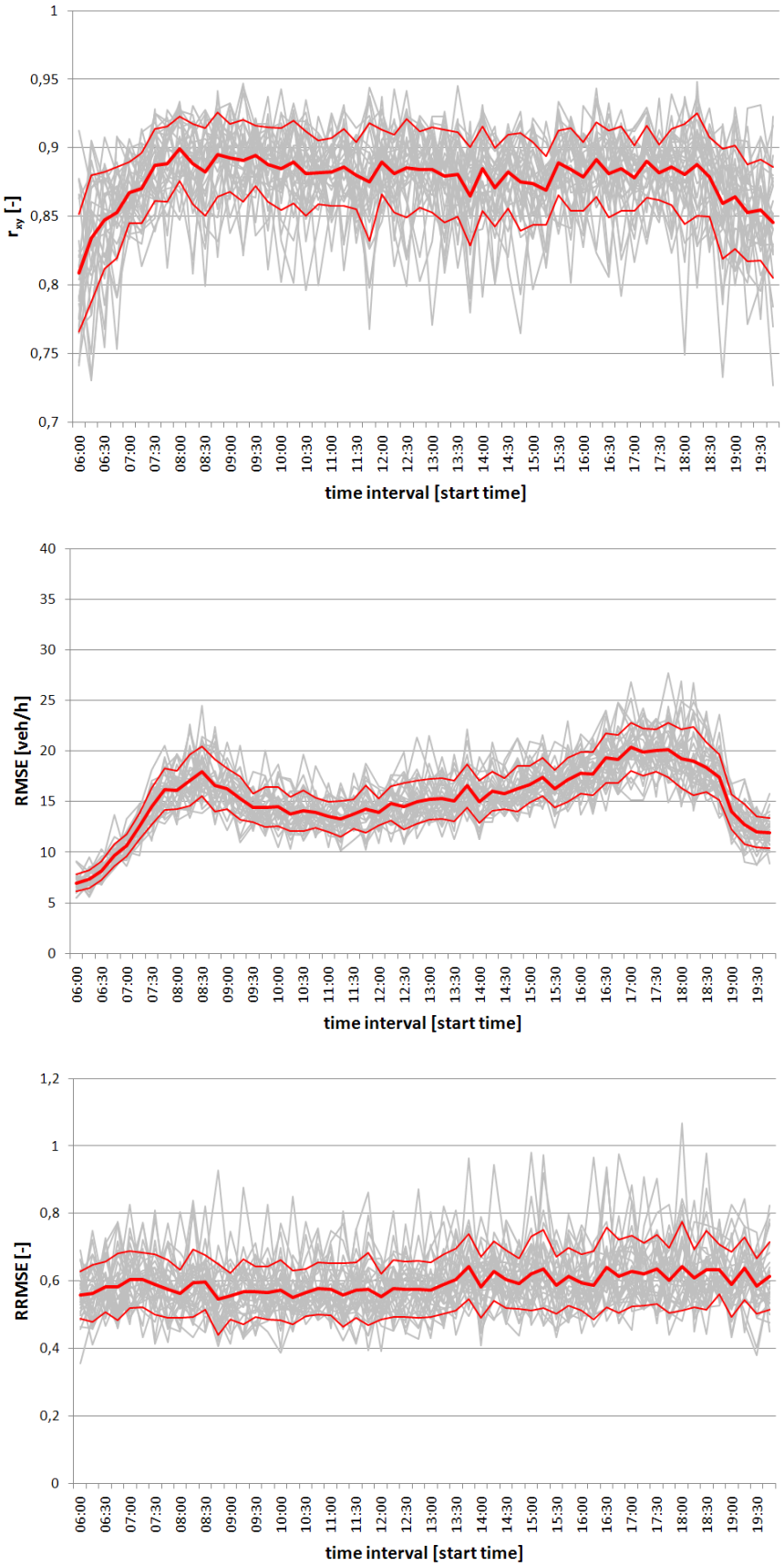


Figure 5-22: Quality of route volume estimation for all replications

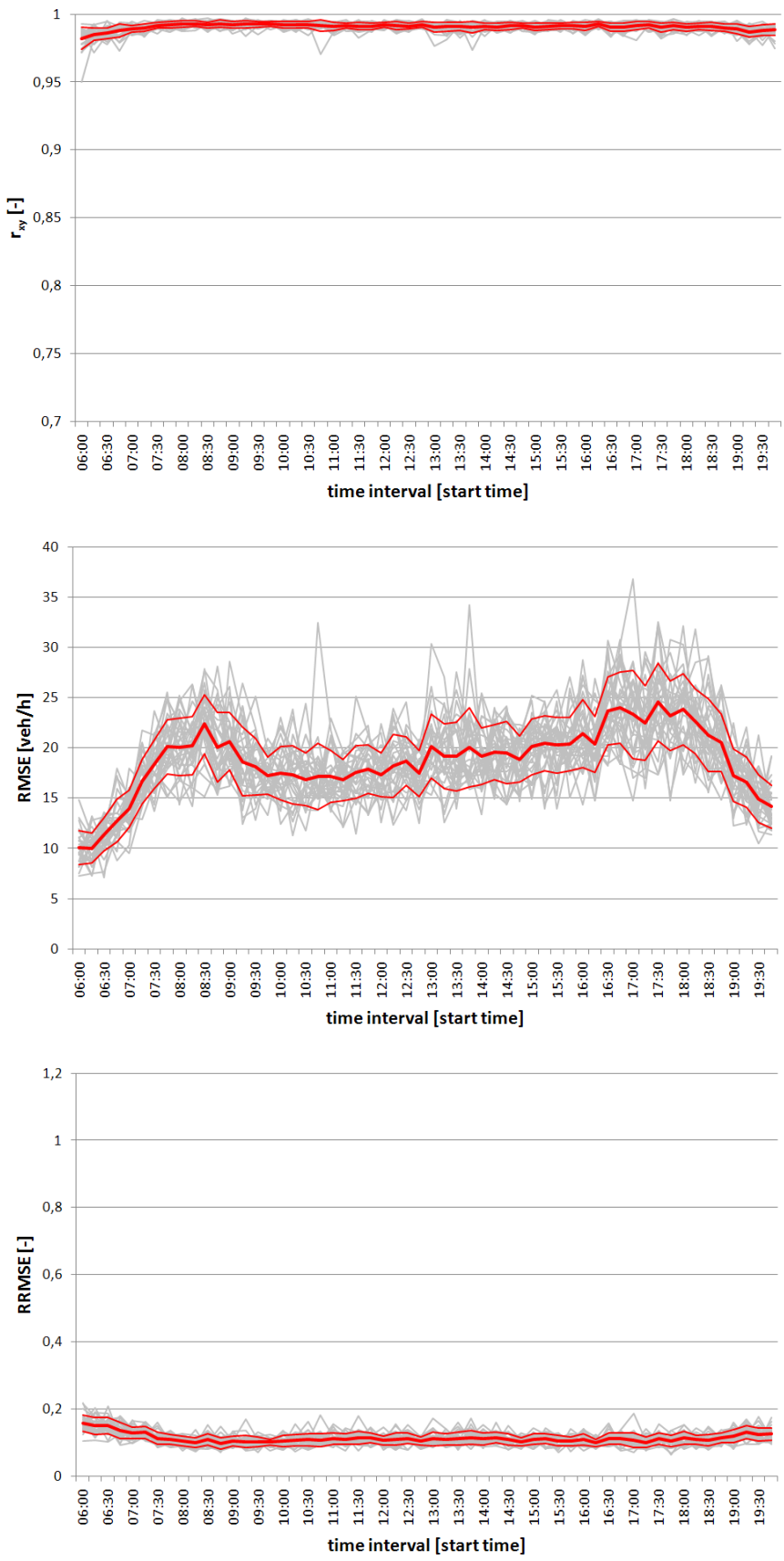


Figure 5-23: Quality of link volume estimation for all replications

The same evaluation for the estimation of link volumes reveals a much better stable mean correlation coefficient r_{xy} of 0.99 on average and a slightly worse *RMSE* in a range between 10.03 and 24.57 veh/h. However, link volumes are much higher than route volumes and thus the *RRMSE* has a good value of about 0.11 on average over time. These results for the "standard settings" are also summarized in table 5-4.

Table 5-4: Overall quality of route and link volume estimation

		route volumes	link volumes
standard settings (unit matrix, SUE, CTM travel times)	r_{xy} [-]	0.88 [1)]	0.99 [1)]
	RMSE [veh/h]	6.95 to 20.39 [2)]	10.03 to 24.57 [2)]
	RRMSE [-]	0.59 [1)]	0.11 [1)]
previous matrix as historic matrix	r_{xy} [-]	0.88 [1)]	0.99 [1)]
	RMSE [veh/h]	6.95 to 21.12 [2)]	9.89 to 23.54 [2)]
	RRMSE [-]	0.64 [1)]	0.11 [1)]
successive assignment	r_{xy} [-]	0.76 [1)]	0.97 [1)]
	RMSE [veh/h]	9.57 to 31.60 [2)]	16.12 to 46.28 [2)]
	RRMSE [-]	1.09 [1)]	0.19 [1)]
BPR-functions	r_{xy} [-]	0.82 [1)]	0.99 [1)]
	RMSE [veh/h]	7.84 to 24.40 [2)]	10.49 to 25.45 [2)]
	RRMSE [-]	0.64 [1)]	0.12 [1)]

1) average of mean value over all time intervals
2) minimum and maximum of mean value over all time intervals

Further settings of demand estimation have been analyzed. The shapes of the according diagrams are similar to figure 5-22 and figure 5-23 and are therefore not displayed. However, characteristic values of correlation coefficient r_{xy}, *RMSE*, and *RRMSE* that describe these diagrams are presented in table 5-4 in the same way as for the case of standard settings.

A first variation of the settings was to use the estimated matrix of the previous time interval as initial matrix instead of a unit matrix, except for the first interval where no previously estimated matrix was available. Table 5-4 shows that this approach makes no real difference. The iterative algorithm of repeated traffic assignment and matrix estimation is obviously able to adapt any given start matrix to a rather similar estimated matrix.

A further test was to use Successive assignment with n = 20 slices instead of SUE assignment. This approach turned out to be less accurate and is not recommended. The traffic demand used for evaluation includes no oversaturated intersections, and thus travel times on the different routes depend much more on the effects of signalization than on the traffic volumes on the links (even though the latter have also an influence that must not be neglected). Because of this, it could be observed that Successive assignment tends to assign an important share of the vehicles of an OD pair to the shortest route only, resulting in a worse estimation of the portions p_{ij}^a and thus reducing the quality of estimation.

As a last test, the traditional BPR function with $\alpha = 0.15$, $\beta = 4.0$ and a lane capacity of 1,800 veh/h has been used for travel time estimation during traffic assignment instead of the CTM travel times. Surprisingly, this approach reduced the quality of estimation only slightly, at least concerning route volumes. The performance of link volume estimation is comparable to the case of standard settings. Obviously, the estimated p_{ij}^a are still good enough and the algorithm is not highly susceptible to incorrect p_{ij}^a as long as the variations are still in a tolerable range.

All in all the evaluation suggests that the plain demand estimation without forecasting delivers satisfactory estimates of the current traffic demand in a network. However, link volume estimation clearly outperforms route volume estimation and the latter has to be used with some reservation.

5.4 Evaluation of combined forecasting and traffic volume estimation

For the task of signal plan optimization, an estimated traffic demand of the next optimization interval is needed. Therefore, the route and link volume estimation described in paragraph 5.3 has to be combined with the forecasting method described in paragraph 5.2, i.e. forecasted detector counts with less precision instead of real counts of the respective interval are used as input for the demand estimation.

Table 5-5 summarizes the results of the assessment of the combined forecasting and estimation module. The standard settings described in paragraph 5.3.9.1 were used. Thus, the values can be directly compared to the “standard setting” case in table 5-4.

It has been explained in paragraph 3.2.2 that in the framework of the developed ATCS a forecasting of traffic demand that is two time intervals ahead is needed. Table 5-5 shows results for the two cases that traffic demand is forecasted either one or two intervals ahead. Detector counts from four preceding time intervals are needed to create the current space-time-pattern. Therefore, the first interval whose demand can be estimated is either the fifth or the sixth depending on the case, and the quality assessment comprises either 52 or 51 intervals only instead of 56.

The quality criteria correlation coefficient r_{xy}, *RMSE*, and *RRMSE* have been determined for two different references. Estimated route and link volumes can be compared to the real volumes of the respective interval whose detector counts have been forecasted. The real volumes are taken from individual AIMSUN replications. This is the “real volumes” reference case in table 5-5. However, it has been mentioned in paragraph 5.2.2 that the forecasting method tends to produce average detector counts as forecasted counts. This will also affect the results of the traffic demand estimation which is directly based on these average detector counts. Therefore, in a second case the estimated route and link volumes have also been compared to the average volumes of each time interval over all 30 replications. This is the “average volumes” reference case in table 5-5.

Table 5-5: Overall quality of combined forecasting and route and link volume estimation

forecast	reference		route volumes	link volumes
1 interval ahead	real volumes	r_{xy} [-]	0.88 [1)]	0.98 [1)]
		RMSE [veh/h]	10.84 to 20.16 [2)]	22.99 to 36.97 [2)]
		RRMSE [-]	0.55 [1)]	0.16 [1)]
	average volumes	r_{xy} [-]	0.93 [1)]	1.00 [1)]
		RMSE [veh/h]	6.86 to 16.37 [2)]	8.79 to 20.56 [2)]
		RRMSE [-]	0.42 [1)]	0.08 [1)]
2 intervals ahead	real volumes	r_{xy} [-]	0.88 [1)]	0.98 [1)]
		RMSE [veh/h]	12.08 to 20.11 [2)]	22.96 to 37.86 [2)]
		RRMSE [-]	0.55 [1)]	0.16 [1)]
	average volumes	r_{xy} [-]	0.93 [1)]	1.00 [1)]
		RMSE [veh/h]	8.39 to 16.72 [2)]	8.71 to 22.84 [2)]
		RRMSE [-]	0.42 [1)]	0.08 [1)]

1) average of mean value over all time intervals
2) minimum and maximum of mean value over all time intervals

If estimated volumes are compared to the real volumes, a decrease of the quality of estimation compared to the non-forecasting case can be observed which is not surprising because of the reduced precision of forecasted detector counts. *RMSE* of both route and link volume estimation and *RRMSE* of link volume estimation are a bit higher than in table 5-4. Only *RRMSE* for route volume estimation is slightly decreased. It has to be noted that the sometimes reduced quality of forecasted traffic counts results in a few time intervals with a rather bad quality of route and link volume estimation. This is not reflected in the average values in table 5-5. No relevant difference can be observed between the two cases of forecasting one or two intervals ahead.

Comparison to average volumes reveals a much higher quality of estimation with an average *RRMSE* for link volume estimation of only 0.08 for both cases of forecasting one or two intervals ahead. This shows that the described method of forecasting and demand estimation has some potential to estimate the upcoming average traffic demand of the next one or two time intervals whereas an estimation of the real traffic demand is more difficult. The method cannot incorporate stochastic variations of traffic flow which are not contained in the reference space-time-pattern used for forecasting. However, a good estimation of the upcoming average demand might be good enough for optimization of signal settings. The impact of traffic demand estimation on the optimized signal settings will be evaluated in chapter 8.

6 Optimization of signal control settings

6.1 Overview

After forecasting and demand estimation have been completed, the ATCS prototype starts optimization of signal settings for the respective time interval. Estimated traffic volumes are available for each link in the sub-network and thus also for each turning movement at all signalized intersections. These flow volumes can be used to determine an appropriate cycle length and adapted green splits or phase durations. Paragraph 6.2 explains the process of this adjustment.

Once the new signal plans have been determined for each intersection, they have to be shifted by different offsets in order to enable a good coordination. The optimization of these offsets is dealt with in paragraph 6.3. In addition to the estimated link volumes, route volumes are used as input as well for offset optimization.

Some final remarks on the overall approach of optimization of signal control settings as implemented in this thesis will be made in paragraph 6.4.

6.2 Adjustment of cycle length and phase durations

6.2.1 Fundamentals

In this work, a classic approach to adjust cycle length and green times or, more precisely, phase durations to a given traffic demand is used. The according equations for calculation of fixed time signal plans are taken from the HBS (FGSV, 2001). The same approach is also adopted in the latest edition of RiLSA (FGSV, 2010). These equations have been transformed into a module for signal plan adjustment. It can be assumed that at least for undersaturated conditions the HBS or RiLSA equations represent a reasonable approach to reduce delays. Since these equations can be seen as a sort of instruction on how to calculate cycle length and phase durations based on given traffic volumes, the term “adjustment” is preferred over the term “optimization” in this thesis. No real optimization takes place in this module, even though the underlying objective of green split and to some extent of cycle length adjustment in the HBS and RiLSA is to minimize delays at intersections, and thus to optimize the signal plans accordingly. The necessary equations will be explained in paragraph 6.2.2.

Only cycle length and phase durations are adjusted in this work. Each phase has to be preplanned and is considered to be fixed, i.e. the signal groups that show green and red during one phase always remain the same. Phase sequences of intersections do not alter either. The phase changes, i.e. the time periods between two phases where traffic signals change from green to red and vice versa, are also preplanned and fixed. Cycle lengths have to be determined for each intersection individually. Then, the longest of these cycle lengths is chosen for the whole sub-network in order to enable offset optimization of signal plans to coordinate intersections.

To apply the HBS/RiLSA equations, traffic volumes on all lanes of an intersection that are controlled by traffic signals are needed. Estimates of these volumes are available from the preceding traffic demand estimation where each lane at an intersection is modeled as an individual link of the network graph. These estimates are available for adjustment of cycle length and phase durations.

6.2.2 Theoretical background

The cycle length t_C [s] of an intersection must be equal to the sum of all phase durations and of all phase change durations or intergreen times:

$$t_C = \sum_{i=1}^{p} t_{G,i} + \sum_{i=1}^{p} t_{IG,i} \qquad (6\text{-}1)$$

where:

p	number of phases [-]
$t_{G,i}$	duration (green time) of phase i [s]
$t_{IG,i}$	duration (intergreen time) of phase change i [s]

It has to be assured that all vehicles that arrive in one cycle can pass the intersection during the next green period, so that holds

$$\frac{q_{rel,i}}{3600\,s/h} \cdot t_C = \frac{q_{S,i}}{3600\,s/h} \cdot t_{G,i} \qquad (6\text{-}2)$$

where:

$q_{rel,i}$	relevant traffic flow of phase i [veh/h]
$q_{S,i}$	according saturation flow [veh/h]

In general, more than one flow has right of way during a specific phase i. The relevant traffic flow of phase i is the flow that is decisive for the necessary phase duration. It is the flow on a lane with right of way that leads to the highest ratio of lane traffic flow to according saturation flow. The saturation flow is the theoretical flow that could discharge from that lane during one hour of unlimited green period (cf. paragraph 2.2). The left term of equation 6-2 expresses the number of vehicles that arrive during one cycle. The right term is the number of vehicles that can discharge during one green period or phase duration $t_{G,i}$ at saturation flow rate.

Transformation of equation 6-2 gives

$$t_{G,i} = \frac{q_{rel,i}}{q_{S,i}} \cdot t_C \,. \qquad (6\text{-}3)$$

Substitution in equation 6-1 leads to the minimum cycle length $t_{C,min}$ according to the following steps:

$$t_{C,min} = \sum_{i=1}^{p} \frac{q_{rel,i}}{q_{S,i}} \cdot t_{C,min} + \sum_{i=1}^{p} t_{IG,i} \qquad (6\text{-}4)$$

$$1 - \sum_{i=1}^{p} \frac{q_{rel,i}}{q_{S,i}} = \frac{\sum_{i=1}^{p} t_{IG,i}}{t_{C,min}} \qquad (6\text{-}5)$$

$$t_{C,min} = \frac{\sum_{i=1}^{p} t_{IG,i}}{1 - \sum_{i=1}^{p} \frac{q_{rel,i}}{q_{S,i}}} \qquad (6\text{-}6)$$

Derivation of equation 6-6 assumed uniform arrival of vehicles at an intersection. The minimum cycle length is just long enough to allow each vehicle to discharge during the cycle of its arrival. WEBSTER (1958) explains that "with random traffic, however, any green time which is wasted because of the variability of the arrival times can never be recovered and the minimum cycle in consequence is associated with very high delays (theoretically, with infinite delays)." To incorporate random arrivals a degree of saturation x_i [-] is introduced that reduces the saturation flows. Typical values of x_i lie between 0.8 and 0.9. Equation 6-6 then changes to

$$t_{C,x_i} = \frac{\sum_{i=1}^{p} t_{IG,i}}{1 - \sum_{i=1}^{p} \frac{q_{rel,i}}{x_i \cdot q_{S,i}}} \qquad (6\text{-}7)$$

where:

x_i degree of saturation of relevant flow $q_{rel,i}$ [-]

Equation 6-7 has been derived analytically, but it does not incorporate any optimization of vehicle delays. Therefore, WEBSTER (1958) developed an adapted equation for an optimal cycle length that is intended to reduce vehicle delay for the relevant flows of each phase under the assumption of random arrivals. To derive this equation he used a delay function that was found by a simple simulation of traffic flow on an "Automatic Computing Engine". While WEBSTER (1958) presented all the details of the derivation of the adapted equation, only the result of his work is described here. According to SCHNABEL/LOHSE (1997) the basic idea was to find an optimal cycle length that is *k*-times the minimum cycle length. WEBSTER (1958) found *k* to be

$$k = 1.5 + \frac{5}{\sum_{i=1}^{p} t_{IG,i}} \qquad (6\text{-}8)$$

which leads to

$$t_{C,opt} = \frac{1.5 \cdot \sum_{i=1}^{p} t_{IG,i} + 5}{1 - \sum_{i=1}^{p} \frac{q_{rel,i}}{q_{S,i}}} . \qquad (6\text{-}9)$$

One can define another variable $b_{rel,i}$ to be the following ratio:

$$b_{rel,i} = \frac{q_{rel,i}}{q_{S,i}} \qquad (6\text{-}10)$$

The sum B of all $b_{rel,i}$ over all phases is

$$B = \sum_{i=1}^{p} b_{rel,i} . \qquad (6\text{-}11)$$

Equation 6-9 can then be written as

$$t_{C,opt} = \frac{1.5 \cdot \sum_{i=1}^{p} t_{IG,i} + 5}{1 - B} , \qquad (6\text{-}12)$$

and equation 6-7 as

$$t_{C,x} = \frac{\sum_{i=1}^{p} t_{IG,i}}{1 - B/x} \qquad (6\text{-}13)$$

if all x_i are the same.

Both HBS and RiLSA propose the approach based on the degree of saturation in equation 6-13 and the approach derived from delay optimization in equation 6-12. Strictly speaking, the latter should only be applied at isolated intersections where random arrival of vehicles is a legitimate assumption, and thus the first should be preferred for coordinated roads and networks. This is conform to the recommendation in the HBS.

Based on the aforementioned delay function, WEBSTER (1958) showed that the practical approach of allocating phase durations “in proportion to the corresponding ratios of flow to saturation flow” is in most cases close to the allocation that leads to the least delay. Thus, the green period or phase duration of phase i can be calculated as follows:

$$t_{G,i} = \frac{b_{rel,i}}{B} \cdot \left(t_C - \sum_{i=1}^{p} t_{IG,i} \right) \qquad (6\text{-}14)$$

This equation can be used for both alternatives of cycle length calculation. For the case of equation 6-13 it can be shown easily that equation 6-14 is equivalent to equation 6-3:

$$t_{G,i} = \frac{b_{rel,i}}{B} \cdot \left(t_C - \sum_{i=1}^{p} t_{IG,i} \right) = \frac{b_{rel,i}}{B} \cdot \left[t_C - t_C \cdot (1 - B) \right] = b_{rel,i} \cdot t_C = \frac{q_{rel,i}}{q_{S,i}} \cdot t_C \qquad (6\text{-}15)$$

Thus, using either equation 6-12 or 6-13 leads to the cycle length of an intersection, and application of equation 6-14 for each phase *i* leads to the quasi-optimal phase durations.

In the case that one or more of the final phase durations fall below minimum green time or phase duration $t_{G,min,i}$, one of the following two equations has to be used according to the previously used equation 6-12 or 6-13 to redetermine the cycle length t_C:

$$t_{C,opt} = \frac{1.5 \cdot \sum_{i=1}^{p} t_{IG,i} + \sum_{i=1}^{p-p'} t_{G,min,i} + 5}{1 - B'} \qquad (6\text{-}16)$$

$$t_{C,x} = \frac{\sum_{i=1}^{p} t_{IG,i} + \sum_{i=1}^{p-p'} t_{G,min,i}}{1 - B'/x} \qquad (6\text{-}17)$$

where:

p'	number of phases whose duration exceeds minimum phase duration [-]
$t_{G,,min,i}$	minimum duration (green time) of phase *i* [s]
B'	sum of b_i of all phases that exceed minimum green time [-]

For all phases whose duration exceeds minimum green time, the phase duration has to be calculated again according to

$$t_{G,i} = \frac{b_{rel,i}}{B'} \cdot \left(t_C - \sum_{i=1}^{p} t_{IG,i} - \sum_{i=1}^{p-p'} t_{G,min,i} \right). \qquad (6\text{-}18)$$

After application of these equations, adjusted cycle length and phase durations are available for each intersection.

6.2.3 Algorithm

The equations had to be transformed into a software module that finds a common cycle length for all intersections in the sub-network and phase durations at each intersection so that these settings are adapted to the estimated traffic volumes of the next optimization interval. A common cycle length is a prerequisite for offset optimization which will be conducted afterwards.

The user has to define all phases, minimum phase durations, the phase sequence and phase changes of each intersection in advance. This implies that the basic design of the signal plans is pre-configured. To allow for more flexibility the module has been programmed in such a way that single flows can have right of way in two consecutive phases. For instance one phase could give right of way to two through lanes of opposing directions while the following phase stops one of these flows and gives right of way to its opposing left-turn lane (cf. figure 6-1). Thus, the other through lane flow has right of way during both phases. In the context of this thesis two phases that share one or more lanes with right of way, i.e. one or more signal groups show a green signal indication during both phases, are referred to as “combined phases” which are comprised of two “sub-phases”. For determination of phase durations of such combined phases it has to be checked whether the relevant flows of the sub-phases or the shared flows of the combined phase are the decisive factor. Strictly speaking, this approach neglects the green time of the affected signal groups during the phase change between the two “sub-phases”.

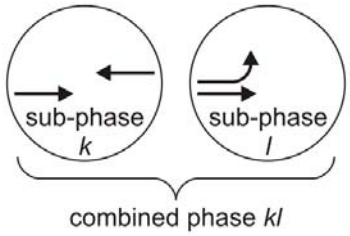


Figure 6-1: Concept of combined phase

The user also has to define saturation flow rates for each lane or link respectively that is under the regime of traffic signal control. For the case of mixed lanes that comprise through and turning traffic flows, saturation flow rates can be specified for each flow direction separately. Based on the currently estimated traffic flows the saturation flow of the mixed lane will then be calculated as recommended in the HBS (FGSV, 2001):

$$q_{S,i} = \frac{1}{\sum_{m=1}^{k} a_m / q_{S,m}} \qquad (6\text{-}19)$$

where:

k	number of traffic flow directions that share the lane [-]
a_m	ratio of traffic flow m to total flow on the lane [-]
$q_{S,m}$	saturation flow of flow direction m [veh/h]

In the following the algorithm that adapts cycle length and phase durations is presented. In a first step the adapted cycle length of each intersection has to be determined individually, i.e. the following steps are executed repeatedly for each intersection.

Step 1: Calculate all saturation flows for mixed lanes according to equation 6-19.

Step 2: Determine all relevant ratios $b_{i,rel}$ of lane traffic flow to according saturation flow (equation 6-10) for each individual phase i that is not a sub-phase of a combined phase. The relevant ratio is the largest among all ratios of lanes with right of way in phase i.

Step 3: Determine the relevant ratios $b_{k,rel}$ and $b_{l,rel}$ of the two sub-phases k and l of a potential combined phase kl. Only the flow ratios of lanes which do not have right of way in both sub-phases may be considered to identify these relevant ratios. In addition, determine the relevant ratio $b_{kl,rel}$ of the combined phase which must only include lane flows with right of way in both sub-phases.

Step 4: Determine the sum B of ratios of lane traffic flow to respective saturation flow according to equation 6-11. All $b_{i,rel}$ of individual phases can be summed up without reservation. In the case of combined phases the larger of the two values $b_{k,rel} + b_{l,rel}$ or $b_{kl,rel}$ is used, i.e. either the two sub-phases or the combined phase are decisive.

Step 5: Determine cycle length t_C either by equation 6-12 or 6-13 and round it to an integer. If B exceeds the value 1.0 (oversaturation), use the latter instead of B in both equations. If t_C is not within a predefined range from $t_{C,min}$ to $t_{C,max}$ set it to the closest of these

borders. If t_C is smaller than the sum of all minimum phase durations and intergreen times, increase it to this sum.

Step 6: Determine the phase durations $t_{G,i}$ of each individual phase *i* that is not a sub-phase of a combined phase according to equation 6-14. If the phase duration is below minimum phase duration, use the latter instead.

Step 7: Determine the phase durations $t_{G,k}$ and $t_{G,l}$ of the two sub-phases *k* and *l* of a potential combined phase *kl* by using $b_{k,rel}$ and $b_{l,rel}$ in equation 6-14. If one or both phase durations are below minimum phase duration, use the latter instead. Determine also the necessary overall duration $t_{G,kl}$ of the combined phase by using $b_{kl,rel}$ in equation 6-14. If $t_{G,k} + t_{G,l} \geq t_{G,kl}$, the durations of sub-phases are decisive. This can happen even if $b_{k,rel} + b_{l,rel} \geq b_{kl,rel}$ because of possibly relevant minimum phase durations of one or both sub-phases. If $t_{G,k} + t_{G,l} < t_{G,kl}$, the overall duration of the combined phase is decisive. In this case the durations of the sub-phases are modified to $t_{G,k} = t_{G,kl} \cdot b_{k,rel} \big/ \left(b_{k,rel} + b_{l,rel}\right)$ and $t_{G,l} = t_{G,kl} \cdot b_{l,rel} \big/ \left(b_{k,rel} + b_{l,rel}\right)$ to sum up to the necessary overall duration of the combined phase. If one of these sub-phase durations falls below minimum phase duration, the latter is used instead and the rest of $t_{G,kl}$ is assigned to the other sub-phase.

Step 8: If minimum green times became relevant in step 7, calculate *B'* with neglect of those phases with relevant minimum green times. Minimum green times of sub-phases of a combined phase are only relevant if the overall duration $t_{G,kl}$ has not been decisive in step 7. Determine the sum of all relevant minimum green times and recalculate the cycle length according to equations 6-16 or 6-17 subject to $t_{C,min}$ and $t_{C,max}$.

After these steps have been run through for each intersection, a set of cycle lengths is available. The longest among these cycle lengths is then chosen as common cycle length of the sub-network. This guarantees that the most heavily loaded intersection can still operate without oversaturation, unless the respective sum *B* has exceeded 1.0. In this case, the intersection is overloaded anyway.

The phase durations of each intersection have to be adapted to the common cycle length. To do so, the following steps have to be applied. Since the new common cycle length is now longer than the individual one at most of the intersections, all phase durations have to be adapted accordingly. Since minimum phase durations might no longer be relevant, the whole process of determination of phase durations has to be repeated entirely.

Step 9: Set the cycle length of the intersection to the common cycle length.

Step 10: Reset all previously determined phase durations.

Step 11: Determine sum *B* as in step 4.

Step 12: Determine the phase durations $t_{G,i}$ of individual phases *i*, sub-phases *k* and *l* and combined phases *kl* as in steps 6 and 7.

Step 13: If minimum phase durations became relevant in step 12, calculate *B'* and determine the sum of all relevant minimum green times as in step 8.

Step 14: Recalculate the phase durations $t_{G,i}$ of individual phases *i* whose minimum green time has not become relevant yet according to equation 6-18.

Step 15: If no minimum phase durations of the two sub-phases of a potential combined phase *kl* have been relevant yet, recalculate the phase durations $t_{G,k}$, $t_{G,l}$ and $t_{G,kl}$ as explained in step 7, but by using equation 6-18 instead of equation 6-14. If the individual durations of the sub-phases turn out to be decisive and one of them falls below minimum green time, the necessary overall duration $t_{G,kl}$ of the combined phase is no longer relevant. The sub-phases can be treated as ordinary individual phases in the following.

Step 16: If additional minimum phase durations became relevant in steps 14 and/or 15, recalculate *B'* as in step 8, determine the new sum of all relevant minimum green times and go to step 14; else finish the allocation of phase durations.

After these steps have been applied, the adapted signal plans of all intersections for the next optimization interval are ready to be used for offset optimization.

6.3 Offset optimization

6.3.1 Problem and approach

While a rather simple non-model-based approach has been used for the adjustment of cycle length and phase durations, the main focus of the new ATCS prototype is on the offset optimization, which poses a challenge especially in meshed networks. A sophisticated module with alternative optimization techniques has been developed to deal with this problem. Two of these techniques are based on the research conducted by ALMASRI (2006) for offline optimization of offsets.

Choosing offsets for arterials might be less difficult, but it is more complex in interconnected networks. In a network with *n* intersections there are *n*-1 independent offsets (one intersection generally serves as reference with a fixed offset $t_{off} = 0$). Given a common cycle length t_C for the whole sub-network, offsets of each intersection can take a value between zero and t_C minus 1, i.e. the signal plans can be shifted along the whole range of cycle seconds. From this it follows that the solution space that comprises all possible offset combinations has a size of $t_C^{(n-1)}$.

Since the solution space is highly irregular it is impossible to solve the problem of offset optimization analytically. In order to assess the potential of a specific offset combination in terms of vehicle delay, the CTM described in chapter 4 is used in this thesis. ALMASRI (2006) showed the usefulness of this model for offset optimization in urban networks. As has been said in paragraph 4.6, one run of the CTM takes 52 milliseconds on average to simulate 15 minutes of traffic flow in the test network. Due to the huge solution space the problem cannot be solved by brute force, i.e. by trying every possible solution. If the cycle length was 90 seconds for instance, the test network with 10 signalized intersections would have $3.87420489 \cdot 10^{17}$ possible solutions. Evaluating all of them with the CTM would last 638,821,202 years! As proposed by ALMASRI (2006) a heuristic GA will be used instead in this work to find a good solution. GA are designed for problems with a huge and irregular solution space that cannot be solved analytically. Even though they do not guarantee to find the global optimum, they are able to find suitable solutions.

The next paragraph will provide background information on GA. Pragraphs 6.3.3 and 6.3.4 will describe two variations of offset optimization based on a GA as proposed by ALMASRI (2006)

with some adaptations. Paragraph 6.3.4 presents a much simpler optimization algorithm that has originated from the second GA based approach during its implementation. It will be used for comparison during the final evaluation.

6.3.2 Genetic Algorithms

GA attracted major interest in traffic related applications during the last years (cf. paragraph 2.6), but they can be used for any sort of problem that cannot be solved analytically or at all in reasonable time. A comprehensive introduction to GA is given by POHLHEIM (2000). In the following, the relevant basics of GA are summarized. Where appropriate, these basics are presented in the context of offset optimization to facilitate comprehension.

Problems with a huge and highly irregular solution space are predestined to be solved by a heuristic algorithm such as a GA. A solution consists of a set of values for different variables. In this thesis these variables are offsets of different intersections, i.e. a specific offset combination is a possible solution. A solution has a fitness according to an objective or fitness function. In the case of offset optimization this fitness can be the total delay that the offset combination imposes on the traffic in the network. In this case, lower fitness values are better. However, in other cases a high fitness value might be preferable. Thus, the GA tries to find a solution that minimizes or maximizes the fitness value, depending on the case.

The basic idea of GA is to imitate the evolutionary process. In this framework a possible solution is called a chromosome. A chromosome consists of several genes, one for each variable. In the beginnings of GA these genes were in general binary representations of each variable. ALMASRI (2006) used this concept and coded each of the offsets as a binary substring of the complete chromosome. Nowadays however it is more common to use integer or real numbers for the variables or combinations thereof (cf. figure 6-2).

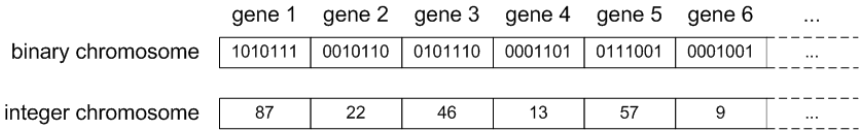


Figure 6-2: Binary and integer chromosome

Figure 6-3 shows the typical flowchart of a GA. The GA starts the search with a first set of different chromosomes or solutions, called a population. The chromosomes are individuals of the population. The initial population can be created randomly, i.e. the desired number of chromosomes is generated by choosing the values of the genes or variables of each chromosome uniformly from the range of valid values of each variable. In the context of offset optimization each gene can take any integer value between zero and the common cycle length t_c minus 1. If knowledge on some good solutions exists a priori, these solutions can be included in the initial population.

The GA will then calculate the fitness of each chromosome in the population according to the fitness function. The fitness function does not necessarily have to be a mere mathematical function. In fact, this kind of problem could be solved analytically in general with no need for a GA. On the contrary, it is more common that determination of the fitness value involves an entire process of comprehensive calculations. In this thesis for instance these calculations correspond to one entire run of the CTM for each chromosome.

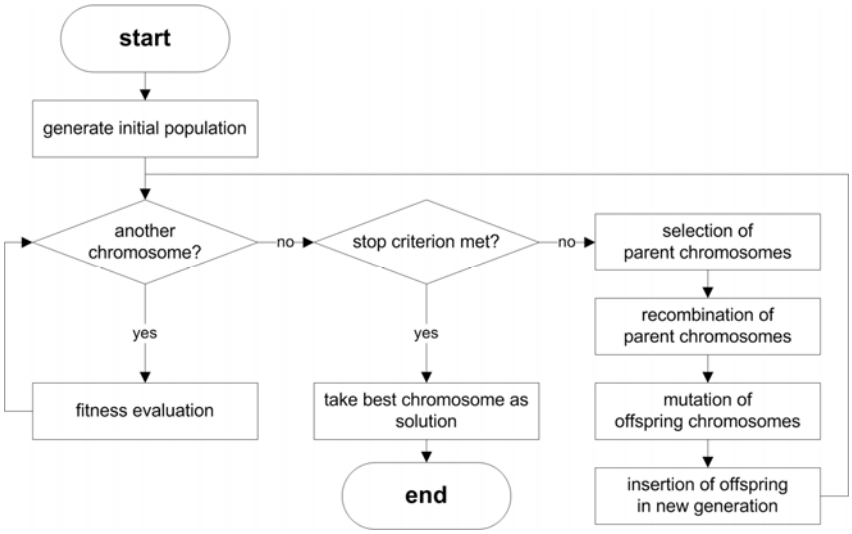


Figure 6-3: Flowchart of a Genetic Algorithm

The next step of the GA is to create a next generation of chromosomes which inherit genes from chromosomes of the initial population. The chromosomes of the initial population are thus parents to the chromosomes of the next generation which are referred to as children or offspring. The process of creating a next generation has three sub-steps.

The first sub-step is selection. The individuals from the initial population which are allowed to reproduce have to be selected. Individuals with better fitness values should be more likely to become parents. One technique to choose the parent chromosomes is roulette wheel selection. An imaginary roulette wheel is divided into several sections, one for each chromosome of the population. The share of each chromosome's section is proportional to its fitness, i.e. fitter chromosomes have a larger share of the wheel. For each selection, a uniformly distributed random number is chosen in the range from zero to the circumference of the wheel. This number falls into the section of one chromosome which is selected for mating. Selection is repeated until the necessary number of parent chromosomes has been reached. Roulette wheel selection implies that the fitness functions attributes high fitness values to good solutions and low values to bad solutions.

An alternative technique is tournament selection. A predefined number of chromosomes, in general two to three, are chosen at random and uniformly from the population. These chromosomes are participants of a tournament. The chromosome with the best fitness is chosen for mating, the others are returned into the pool of candidates. This selector only considers the rank of solutions in terms of fitness, whereas the absolute fitness is irrelevant. Tournament selection can thus be used in both cases that higher or lower fitness values are better.

A third selector is the truncation or threshold selector. A threshold in percent of the population size has to be defined that specifies the number of chromosomes that are allowed to reproduce. The chromosomes of the initial population are sorted by fitness and the best chromosomes are selected in decreasing order until the desired number of chromosomes is reached. Again, the selection only depends on the rank of solutions.

The next sub-step is recombination. A new chromosome has to be created from two parent chromosomes. If variables are integer or real numbers, discrete recombination can be used as

an option (cf. figure 6-4). The child inherits each gene from one of the two parents. Whether a gene will come from parent 1 or parent 2 is decided separately for each gene with the same probability for both parents.

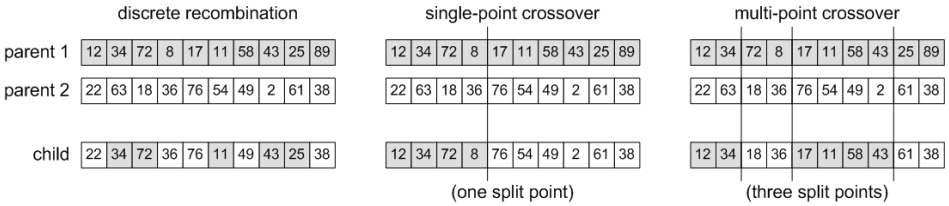


Figure 6-4: Recombination of chromosomes

Another technique is crossover recombination which was originally designed for binary chromosomes. The first bits of the chromosome come from parent 1 and the last bits from parent 2. The split point is chosen at random. In the described case of only one split point this method is called single-point crossover, but multi-point crossover with several split points exists as well. The concept can also be transferred to the case of real and integer variables. In this case, which is also shown in figure 6-4, the split points are not located between single bits but between genes. In the case of single-point crossover the child inherits the first genes up to the split point from the first parent and the remaining genes from the second parent. Multi-point crossover works in an analog way.

The third sub-step is mutation. Some of the child chromosomes are mutated slightly, i.e. the values of single genes are changed at random in order not to get stuck in local optima of the solution space. In general, mutation is only applied at a low rate.

Once the desired number of children has been generated, these new chromosomes are inserted into the population where they replace the same number of old chromosomes. This modified population represents the first generation. The whole process of fitness evaluation, selection, recombination and mutation is repeated with this new generation of chromosomes and the following generations. To make sure that the best solution of a generation does not get lost, elitism can be used. This variation guarantees that the currently best solution is always transferred to the next generation.

The GA terminates the search as soon as a stop criterion is met. This criterion might be that the best solution does not improve anymore over a certain number of generations, that a maximum number of generations has been reached or that the available computing time is exceeded. When the GA has stopped, the best chromosome of the last generation is taken as solution of the problem. Even though application of GA does not guarantee that the global optimum is found, solutions are in general rather close to it. Since the GA is capable of searching within the whole range of the solution space, the risk of getting stuck in an area of a local optimum is reduced.

6.3.3 Parallel Genetic Algorithm

The first variation for offset optimization proposed by ALMASRI (2006) is the PGA. All variables, i.e. all independent offsets of the intersections are optimized simultaneously, or in other words in parallel. ALMASRI (2006) used binary chromosomes to code the offsets. In this thesis, an integer gene has been used for each offset.

As has been said before, *n* signalized intersections result in *n*-1 independent offsets because one intersection is in general used as reference. For reasons that will be discussed in paragraph 7.4 of the next chapter and that have to do with signal plan transition, all *n* offsets are free variables in this work. Thus, a network with *n* intersections needs chromosomes with *n* integer genes.

The fitness of a chromosome, i.e. a specific offset combination, is the amount of total vehicle delay [veh·s] that it induces to the network and that is experienced by the collectivity of drivers. To estimate this delay, each chromosome requires one run of the CTM which has been programmed to internally switch traffic signals according to the previously determined common cycle length and phase durations and the currently assessed offset combination. The traffic demand that is fed to the CTM network during each run equals the estimated link volumes which have to be transformed into source inflows and turning portions at diverge cells. The implementation of the CTM is sufficiently fast to enable the GA to test hundreds of different chromosomes within only a few minutes.

Since in this first version the CTM starts each run with an empty network, a warm-up period has to be considered to allow the network to fill up with vehicles. The duration of the warm-up period has been chosen to be three times the current common cycle length t_C. All delays that arise during this warm-up period are excluded from the final calculation of total delay.

The GA framework has not been programmed entirely anew for this thesis. The open source Java Genetic Algorithms and Genetic Programming Package (JGAP) provided by MEFFERT ET AL. (2009) has been used instead. This package allows implementation of arbitrary fitness functions. A variety of parameters can be set by the user, or a set of default values can be used. The user also has a limited choice of different selection and recombination methods and is free to implement own operators.

Before integrating JGAP into the ATCS prototype and using the CTM for fitness evaluation, a simple fitness function has been tested first to make sure that the correct application of the software package has been fully understood and that it works well. The idealized artificial arterial shown in figure 6-5 with 10 signalized intersections has been chosen for this purpose, so that 10 different offsets $t_{off,i}$ had to be optimized. All adjacent intersections have an equal distance of 625 meters. Given a speed of 50 km/h the optimal relative offset $t_{rel,opt}$ between intersections is thus 45 seconds in both directions. The common cycle length has been set to 90 seconds.

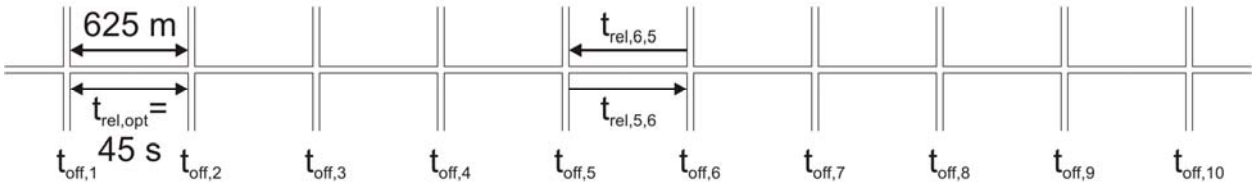


Figure 6-5: Idealized arterial for testing the GA software package

The fitness function has been programmed to calculate all relative offsets $t_{rel,i,j}$ between neighboring intersections *i* and *j* in both directions based on the absolute offsets $t_{off,i}$ and $t_{off,j}$. Then the absolute differences of these relative offsets to the optimal relative offset of

45 seconds are summed up and returned as fitness of the currently assessed offset combination. This is expressed by the following equation:

$$fitness = \sum_{i=1}^{n-1} \left(\left| t_{rel,i,i+1} - 45\ \text{s} \right| + \left| t_{rel,i+1,i} - 45\ \text{s} \right| \right) \qquad (6\text{-}20)$$

where:

n	number of intersections [-]
$t_{rel,i,j}$	relative offset between intersections i and j in direction from i to j [s]

From equation 6-20 it follows that smaller fitness values are better than high values. In the best case all relative offsets are 45 seconds and the fitness value is zero.

JGAP has been set to start with a random initial population. The population size was 50 chromosomes. The search stopped after 200 generations. Tournament selection has been chosen with a tournament size of three chromosomes. Crossover rate of the single-point crossover operator has been set to 0.35 which is a standard setting of JGAP. It means that 35 percent of the population is replaced by new offspring chromosomes at each generation. Mutation rate has been set to 12 which is also a standard setting. It means that every 12th gene is mutated on average.

Table 6-1 shows the results of 10 test runs of the GA. It displays the relative offsets between all neighboring intersections in both directions as they arise from the final solutions. The last column shows the fitness of the final solution. The last row contains the average values of the absolute differences between the final relative offsets and the optimal relative offset over all 10 test runs and for each combination of neighboring intersections.

Table 6-1: Relative offsets derived from the final solutions of the GA software package test

	relative offsets $t_{rel,i,j}$ [s]																				fitness
run	1,2	2,1	2,3	3,2	3,4	4,3	4,5	5,4	5,6	6,5	6,7	7,6	7,8	8,7	8,9	9,8	9,10	10,9	[s]
1	45	45	46	44	46	44	46	44	45	45	45	45	45	45	42	48	45	45	12
2	45	45	47	43	45	45	44	46	45	45	43	47	37	53	45	45	44	46	28
3	44	46	44	46	45	45	44	46	45	45	45	45	44	46	44	46	43	47	14
4	45	45	43	47	34	56	44	46	39	51	45	45	43	47	45	45	43	47	48
5	45	45	44	46	45	45	44	46	45	45	40	50	45	45	43	47	45	45	18
6	45	45	45	45	41	49	41	49	34	56	42	48	45	45	45	45	45	45	44
7	49	41	48	42	57	33	53	37	46	44	47	43	48	42	45	45	45	45	66
8	45	45	46	44	47	43	45	45	45	45	48	42	46	44	45	45	45	45	14
9	46	44	48	42	48	42	47	43	45	45	44	46	44	46	43	47	45	45	26
10	45	45	40	50	42	48	17	73	41	49	43	47	45	45	46	44	45	45	86
average of absolute differences | $t_{rel,i,j}$ - $t_{rel,opt}$ | [s]																					
	0.6		1.9		3.6		4.7		2.2		1.8		1.6		0.9		0.5				

Obviously the GA works fine. Most of the relative offsets have an optimal value of 45 seconds or are close to it. None of the fitness values reaches zero, but the values are in general very low. However, table 6-1 also reveals that some outliers occur from time to time which reduce the fitness (relative offsets that differ by more than 5 seconds from the optimal relative offset are

highlighted). The GA does not always find good offsets for all intersections, but all in all the final solutions are of high quality.

After this initial test, JGAP has been integrated into the software framework of the ATCS prototype. The fitness function has been programmed to use the CTM for evaluation of each chromosome or offset combination as described before. In order to find suitable parameters of the GA many optimization runs with different settings have been conducted. The List test network and the average traffic demand of the afternoon peak interval of 15 minutes have been used. The signal plans have been adjusted before each offset optimization run as described in paragraph 6.2. Since this signal plan adjustment is deterministic, the signal plans have been exactly the same for each offset optimization run. Details on the phases and phase sequences will be given in chapter 8 about the evaluation of the overall ATCS prototype.

Differences of the performance of the GA could be observed for different parameters, but in most cases they were small. The applied selection method had the greatest influence. In the end, the same parameters as for the arterial test case were chosen, i.e. random initial population, population size of 50, tournament selection with three competitors, single-point crossover operator with crossover rate of 0.35, mutation rate of 12, and elitism enabled.

The only difference was the choice of the stop criterion. The duration of 15 minutes of each optimization interval constrains the available time to finish optimization of signal settings for the next interval. Therefore, the GA has been set to stop as soon as a predefined calculation time has been exceeded. In order to have some buffer time left, a maximum calculation time of 10 minutes has been chosen in a first attempt. Several test runs revealed that the greatest improvements of the fitness occur within the first generations. No additional major improvements were obtained after a time period of 5 minutes, sometimes even earlier. The maximum computing time has therefore been set to 5 minutes (or 300 seconds).

Figure 6-6 shows the evolution of fitness of the best solution of each generation over time for six different test runs. The figure shows clearly that the most significant improvements occur at the beginning of the search which justifies the decision to stop the GA after 300 seconds. PGA was able to iterate 130 to 132 generations in this time. The figure also accentuates the stochastic behavior of the GA. The fitness values of the six final solutions are in a similar range but still differ from each other.

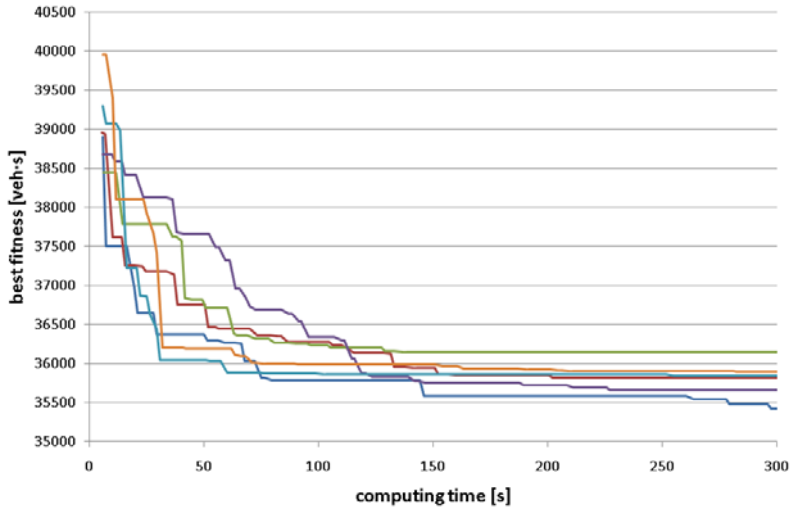


Figure 6-6: Evolution of fitness during six test runs using the PGA

6.3.4 Serial Genetic Algorithm

ALMASRI (2006) tested a second alternative which he called SGA. This approach tries to incorporate the knowledge of important routes into optimization. In a first step SGA only optimizes the offsets of the intersections along the heaviest loaded route according to the current traffic demand. The genes of the chromosomes only comprise the offsets of these intersections. All other offsets are kept constant at zero. In a next step the remaining offsets of intersections along the second heaviest loaded route are optimized by the SGA. The offsets of the intersections along the first route are kept at their previously optimized values. SGA continues the procedure with the third heaviest loaded route and so on until all offsets have been considered.

Figure 6-7 illustrates the concept of offset grouping for a small grid network with nine intersections. Part (a) shows the four heaviest loaded routes, numbered in order of decreasing traffic volumes. Part (b) shows the according groups of offsets that are optimized separately by the SGA.

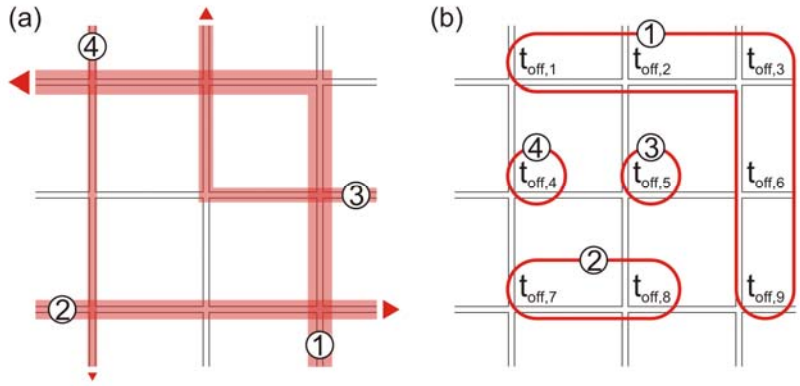


Figure 6-7: Concept of grouping offsets in the SGA

Figure 6-7 also shows that the groups may contain only one single offset. For such groups only one variable has to be optimized. Using a GA in this case does not make much sense. Therefore, in this work the SGA has been programmed to run a complete offset enumeration of all possible values between zero and t_C-1 if a group has only one member. If a group contains two or more members, the GA is used.

ALMASRI (2006) defined traffic demand in terms of source inflows and turning percentages. He had thus no direct knowledge of the heaviest loaded routes and had to derive route volumes based on the given information. In this work, however, route volumes are estimated by the demand estimation module. The estimates can be used directly to group the offsets.

The available computing time of 300 seconds has to be distributed among the different groups. This can be done by assigning shares of the total computing time to the groups in proportion to the number of offsets in each group.

As in the case of PGA, six test runs have been conducted using the SGA. The evolution of fitness of the best individual is shown in figure 6-8. It can be seen that major improvements occur directly after the SGA finishes one group and starts with the next. After these drastic improvements no major changes can be observed until the next group is considered.

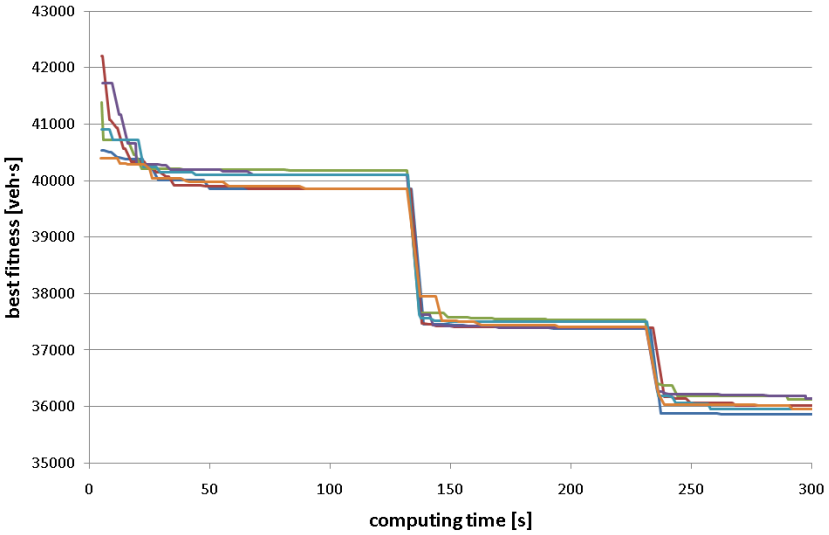


Figure 6-8: Evolution of fitness during six test runs using the SGA

ALMASRI (2006) found SGA to be faster and to find better solutions than PGA. However, a first comparison of the final results in figure 6-6 and figure 6-8 does not suggest superiority of SGA over PGA in the given case. In chapter 8 both methods are evaluated more in detail.

6.3.5 Sequential Enumeration

A third algorithm has been tested in this work which does not employ a GA. It resulted from the implementation of SGA which conducts a complete enumeration of all possible offsets for groups with only one member. The third algorithm applies this concept to all offsets and handles them one after another, which is why it has been named Sequential Enumeration (SE).

SE sets all offsets to zero in the first place, followed by a successive complete enumeration and evaluation of all possible offsets between zero and t_C-1 at every single intersection. Intersections are processed in the order of decreasing traffic loads on the routes. The algorithm starts with the first intersection along the most heavily loaded route, goes on to the second intersection and so forth until all intersections along this route have been considered. Then the second heaviest loaded route is addressed in the same way, followed by the third and so on until all offsets have been considered. In the example in figure 6-7 the processing order of intersections would be: 9, 6, 3, 2, 1, 7, 8, 5, and 4.

Before proceeding to the respective next intersection, the best offset that has been found for the current intersection is kept. After all intersections have been considered once, the algorithm returns to the first intersection and continues the search in the same order of intersections until one complete repetition over all intersection achieves no further improvement or until a maximum of four repetitions has been reached.

Figure 6-9 shows the evolution of fitness during optimization with SE for the afternoon peak interval in the List test network. The common cycle length has been set to 68 seconds by the previous adjustment module, so that 68 runs of the CTM had to be performed per intersection. In the example no further improvement is achieved after the first repetition. Since SE is a deterministic algorithm, further test runs will always produce the same result for the same given demand.

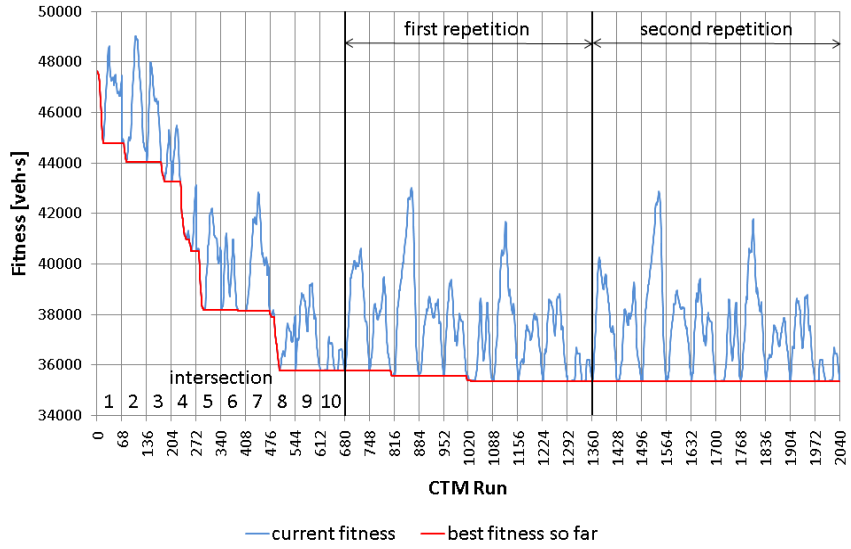


Figure 6-9: Evolution of fitness during SE

Figure 6-10 compares the evolution of the best fitness during application of SE to those of PGA and SGA. The curves for PGA and SGA are average curves of the six runs displayed in figure 6-6 and figure 6-8.

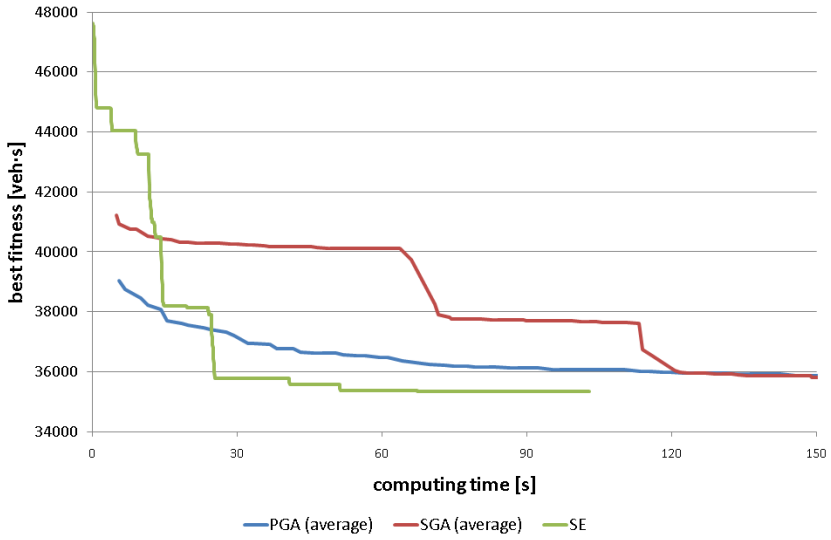


Figure 6-10: Comparison of PGA, SGA and SE

On average both GA based algorithms find and equally good solution. But it is rather surprising that SE finds a solution in the same range within shorter time. This raises the question whether complex GA are an appropriate method to optimize offsets since a much simpler approach is obviously capable of producing comparable results. Details on the further evaluation of all three algorithms will be given in chapter 8.

6.4 Additional remark

In this thesis optimization of signal settings has been divided into two sub-problems that are handled separately one after another. First, a common cycle length and adequate phase dura-

tions are determined, and then the offsets of the resulting signal plans are optimized. However, there is a close relation between cycle length and offsets. Cycle length defines the range of valid offset values and thus the solution space for offset optimization. It is conceivable that a specific cycle length creates a solution space that contains better offset combinations than a different solution space constrained by another cycle length. In order to take this effect into account, BRAUN (2008) designed a GA that optimizes different signal settings at once, including cycle length and offsets. Cycle length is represented by one of the genes of the chromosomes. Since a different cycle length changes the range of valid offsets, BRAUN (2008) could not use integer genes. Instead he coded offsets as real values between 0.0 and 1.0 to express a portion of cycle length. These portions are used to calculate integer offsets subject to the value of the cycle length gene. This implies that the same real value of an offset gene has a different meaning for different cycle lengths. The solution space is constantly changing for each chromosome with a different cycle length.

To avoid such behavior, the stepwise approach described in this chapter has been preferred in this thesis. First, a specific cycle length has to be determined and then the GA optimizes offsets with a constant solution space. In order to incorporate the effects of different cycle lengths on offset optimization, a future enhancement of the ATCS prototype could be to run the GA several times for different cycle lengths. Given the current maximum computing time of the GA of 300 seconds, three different cycle lengths could be checked in 15 minutes. However, such an approach has not yet been implemented.

BRAUN (2008) and others tried to use heuristics to optimize the whole range of signal plan settings. The author of this thesis, however, is of the opinion that at least for undersaturated conditions signal plan calculation as described in paragraph 6.2 is an adequate and reliable approach. The evaluation of the ATCS prototype in chapter 8 supports this assumption. The need for heuristics to optimize cycle length and phase durations is therefore not obvious.

7 Signal plan transition

7.1 Overview

Since the ATCS prototype repeatedly sends new signal plan settings to the controllers for each optimization interval, a signal plan transition at each intersection has to occur at the beginning of each time interval in order to switch from the previous to the new settings. Transition plays an important role not only for ATCS but also for pre-planned fixed signal plans that are chosen according to a time-of-day (TOD) schedule.

For an isolated intersection, the task of transition is rather simple. Transition can take place in just one second by jumping from the break point of the old plan to the one of the new plan, both having the same signal indication. At arterials or in meshed networks with several coordinated intersections, however, the new plans must be in sync as soon as possible in order to guarantee a well functioning green wave. Therefore, the new plans have to be shifted according to their respective offsets. Different methods for this shifting exist. During these transition periods neither the settings of the old nor the new plan are active but some intermediate state, inducing additional delays for the drivers. In order to keep these delays small, smooth transition methods are to be preferred over abrupt methods. Then again, the transition period should be preferably short, so that coordination is (re-)established quickly. A trade-off decision between smoothness and shortness has to be made.

Different transition techniques will be described in paragraph 7.2.1. Some fundamental work on the effects of transition has already been published earlier and will be presented shortly in paragraph 7.2.2. Based on these studies a special microscopic simulation study has been conducted explicitly for the case of a meshed network in order to identify a method that enables smooth and quick transition into the new coordination pattern. The study and its results will be presented in paragraph 7.3. Based on the results a transition technique has been chosen which will be used in the framework of the ATCS prototype.

Transition may also influence the overall performance of a certain offset combination. If transition is not considered when evaluating an offset combination, the same combination may perform better in the CTM than it finally does in reality. Paragraph 7.4 will show how transition can be included into the assessment of possible offset combinations.

7.2 State-of-the-art

7.2.1 Transition methods

Different transition methods exist and are used in practice with slight modifications depending on the type of controller. SHELBY ET AL. (2006), COHEN ET AL. (2007), and LEE/WILLIAMS (2009) amongst others give a good overview. In the following different transition methods are described the way they have been implemented in the experimental study. Figure 7-1 illustrates the methods by showing a transition from signal plan A to B.

Dwell: Once the coordinated phase is reached, it is held for an additional dwell period until synchronization with the new plan is accomplished. This method may result in dwell periods of up to one cycle length minus one second.

Max dwell: This method is similar to Dwell, but with a dwell period limited to a certain percentage of cycle length. If the necessary shift of signal plan exceeds this limitation, the shifting has to be performed in multiple cycles. If for example the offset has to be adjusted by 50 seconds and the limit of the dwell period is 20 seconds per cycle, the coordinated phase will be held for an additional 20 seconds in the first and second cycle after start of transition and for another 10 seconds in the third cycle.

Shortest dwell: This version of Dwell determines the shortest possible dwell period within the next three cycles and additionally holds the coordinated phase for this dwell period in the respective cycle. The shortest dwell period may happen to be the one in the first cycle, in which case the method is equal to simple Dwell.

Minimum green time: This method is described in RiLSA (FGSV, 2010). After start of transition, the current phase is terminated as soon as possible (i.e. after minimum green time is over), followed by a change to the next phase of the new and correctly shifted signal plan. It must be assured that the end of this phase according to the new plan can be reached subject to necessary intergreen times and its minimum green time. If this is not possible, the subsequent phase is chosen and switched to directly. It may happen that the next possible phase of the new plan subject to these constraints is the same as the current phase of the old plan. In this case, this method is the same as Dwell. Furthermore, this method possibly skips phases during transition.

Add in two cycles: After start of transition the next two cycles are lengthened equally according to the required adjustment of offset. This is done by lengthening every phase proportionally to its green split in the new plan.

Add in three cycles: This method is equivalent to Add in two cycles, but with three equally lengthened cycles.

Add with maximum lengthening: The cycles are not lengthened in a specified number of cycles, but constrained to a maximum lengthening expressed as a percentage of cycle length. If the signal plan has to be shifted by, for example, 60 seconds and the maximum lengthening of each cycle is 18 seconds, the transition has to be performed in four cycles since only a shift of 54 seconds would be possible in three cycles. In this example, each of the four cycles will be equally lengthened by 15 seconds. Of course, this method may be the same as Add in two or three cycles if the maximum lengthening allows it.

Subtract in two cycles: In contrast to Add, the Subtract method shortens cycles. This is possible since a positive shift of, for example, 50 seconds of a signal plan with a cycle length of 90 seconds is equal to a negative shift of 40 seconds. The two cycles are shortened equally by shortening all phases in proportion to their green splits while observing minimum green times.

Subtract in three cycles: As with Add in three cycles, this method works like Subtract in two cycles, but uses an additional cycle for shortening the phases.

Subtract with maximum shortening: Comparable to Add with maximum lengthening, the shortening of each cycle is constrained to a maximum shortening in percent of cycle length. The necessary number of shortened cycles is derived by this percentage, and all cycles are shortened equally.

Shortway: This method chooses whichever of the two methods Add with maximum lengthening or Subtract with maximum shortening establishes coordination faster.

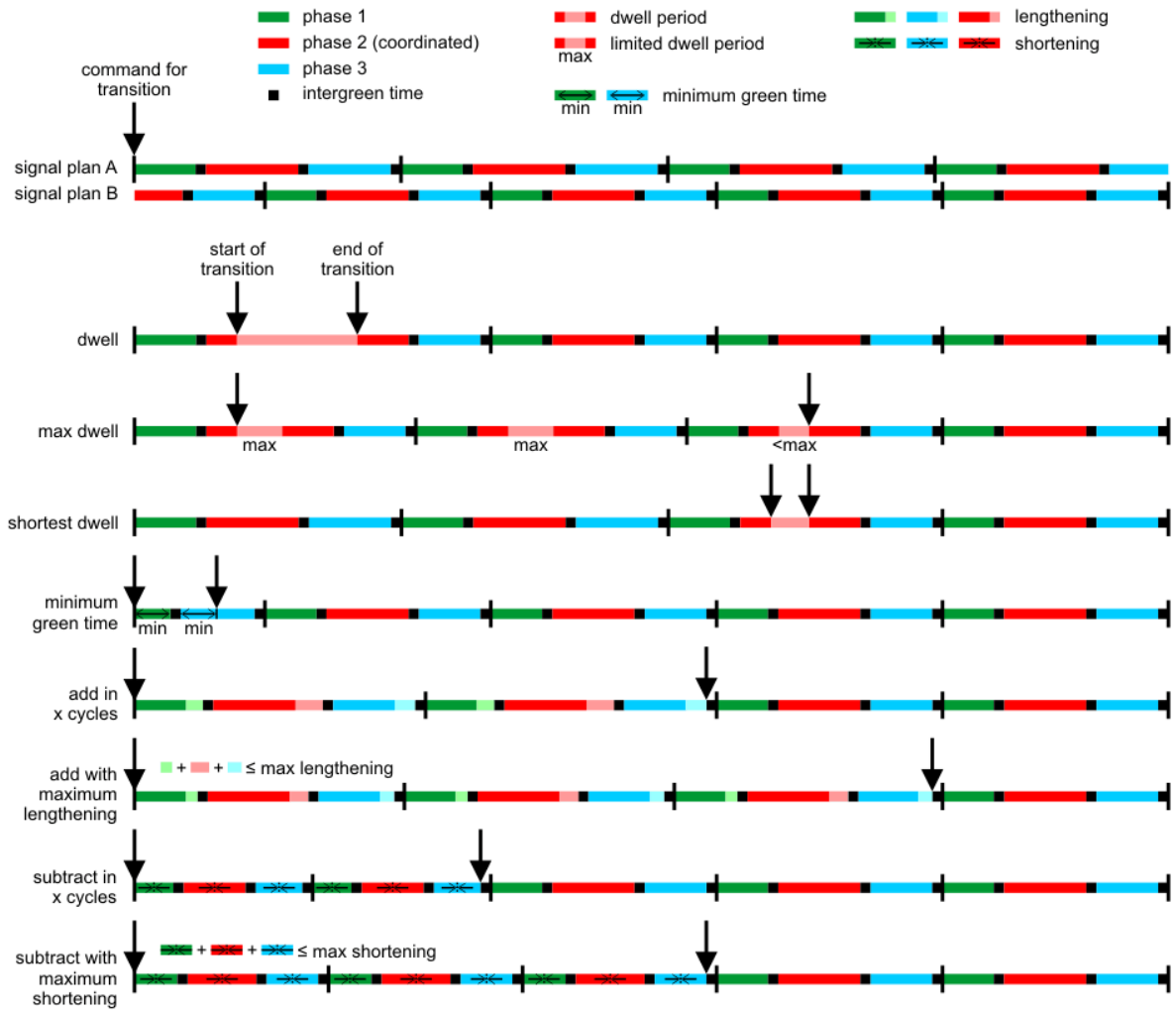


Figure 7-1: Transition methods as used in this thesis

7.2.2 Previous studies

SHELBY ET AL. (2006) give a comprehensive overview of relevant publications on transition and cite some simulation studies. The main findings of their literature review are that transition may possibly induce increased delay not only during transition itself but also some time after transition is finished, and that a smoother transition over several cycles is less disruptive than abrupt transition such as Dwell.

In order to allow for a reliable comparison of transition methods under different conditions, SHELBY ET AL. (2006) conducted an extensive simulation study using CORSIM. Eight transition methods had been tested for six saturation levels between 60 and 110 percent. They considered different offset adjustments with increments of 5 seconds. As a measure of effectiveness the average delay per vehicle in the last 30 minutes of the simulation, i.e. after start of transition, had been used. 30 replications per method were simulated. Max dwell and Add were constrained to 19 percent of cycle length and subtract to 18 percent.

Additional delay of up to 18 percent compared to the case of no offset adjustment was observed for an isolated intersection. However, also a reduction of up to 9 percent could be achieved when Add transition was used during oversaturation. In general, abrupt transition performed worst, whereas smooth transition was better. Add worked best in the highly saturated or oversaturated scenarios, whereas Subtract showed good results at low saturation rates of 80 percent or less. SHELBY ET AL. (2006) concluded that “the transition method that temporarily shifts the saturation level closer to the ‘ideal’ 90 percent target experiences less delay”.

For an arterial with four intersections, Shortway transition turned out to be the most effective method for the undersaturated and nearly saturated case. The above-mentioned beneficial effects of Add and Subtract in the oversaturated or undersaturated scenarios could not be observed in the arterial case. Again, immediate transition performed worst.

The chosen period of 30 minutes used for averaging the vehicle delay is very long and the obtained performance values for different transition methods vary little, especially in the undersaturated and nearly saturated scenarios. Therefore, COHEN ET AL. (2007) extended the study and calculated average performance measures for 18 time intervals of 90 seconds following the start of transition. 30 replications were used. This method, which they called transient profile analysis, allows for better assessment of the impact of different transition methods at a high temporal resolution.

Two arterial networks were modeled in CORSIM with 10 and 6 intersections respectively and two different transition scenarios. Transient profiles for network delay and travel times (overall and for arterial and side-street travel time separately) as well as for some intersections were generated, showing results for Dwell, Max dwell, Shortway and Subtract methods. Both Dwell methods produced major peaks of delay and have been considered by COHEN ET AL. (2006) to be inappropriate for coordinated arterials. For the second network, the additional delays did not even decrease to the same level as with smooth transition long after the transition is over. Shortway and Subtract performed comparably well and produced a much smoother transient profile with less delay and smaller travel times.

YUN ET AL. (2008) used hardware-in-the-loop simulation to evaluate the performance of transition methods as implemented on three real controllers. They modeled a corridor with four coordinated signalized intersections in VISSIM. For evaluation, only performance measures of the

last 15 minutes of the simulation (i.e. after start of transition) have been considered. Only Shortway was possible with the first controller. Transition performance was acceptable but turned out to be insensitive to the number of cycles. Both Dwell and smooth methods could be tested with the other controllers. While in one case, the smooth method was found to perform better, the other controller showed no significant differences for any of its transition methods. It can be argued again that a time interval of 15 minutes is rather long, and therefore negative or positive effects of individual transition methods might average out.

While the methods described in paragraph 7.2.1 are static transition methods, MUSSA/SELEKWA (2003) proposed a methodology to optimize transition based on a dynamic quadratic optimization. Starting and ending point of this optimization are the current and the next signal plan. The methodology tries to find an optimal sequence of different adjustments to transform the old signal plan into the new one. The number of steps of this sequence is also optimized. Since old and new plan are considered to be optimal signal plans for the respective demand, the methodology minimizes "total deviation" from these optimal signal plans during transition. The method has been tested in a CORSIM simulation with three intersections and turned out to perform better than Dwell. It has not been compared to other methods, and therefore no substantiated conclusion can be drawn. Since it behaves like a smooth transition method, the findings are in accordance with other findings mentioned above. However, the simulated transition consisted of 17 adjustment time steps, although establishing the new coordination should take place preferably fast.

Another optimization technique for traffic signal plan transition is proposed by LEE/WILLIAMS (2009). Again, settings for the different adjustment steps are optimized. The method minimizes average vehicle delay during transition, which is calculated based on the HCM delay equations. Traffic flow is assumed to increase or decrease linearly during transition, which will in general not be the case in reality. A GA is used for optimization. A transition generated with this technique for a small network with three intersections has been simulated in CORSIM. Transition consists of six adjustment steps, and again the problem of preferably fast transitions in coordinated networks has to be emphasized. The method has been compared to other abrupt and smooth transition methods. LEE/WILLIAMS (2009) claim that it produces less side street delay, whereas main street delay was more or less the same.

7.3 Experimental study

7.3.1 Setup of simulation

The effects of transition on adjacent intersections as well as in a meshed network were of special interest for this thesis. Furthermore, more detailed information on delay profiles for specific traffic streams at intersections was needed. A special experimental study has been conducted that adds to the findings of the previous studies. An artificial network with three interconnected intersections has been modeled in AIMSUN. It is shown in figure 7-2.

All three intersections have the same geometry, but with different orientation. The same applies for the signal groups S1 to S6 and the phase sequence. The small triangles in the circles indicate how to transfer the signal group scheme and phase sequence to each intersection.

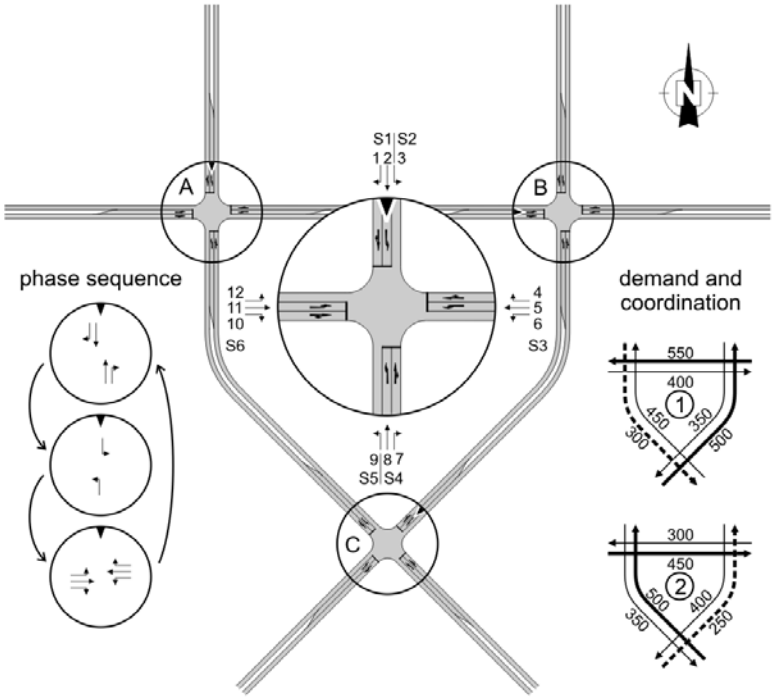


Figure 7-2: Simulation setup for transition study

Two traffic demands have been defined whose traffic volumes [veh/h] on six relations are shown. All other turning volumes have been set to 100 veh/h except for left-turning traffic with a phase of its own (streams 3 and 9), which has been set to 200 veh/h. For each demand signal plans have been constructed with a common cycle length depending on the heaviest loaded intersection. Cycle length and green splits have been calculated according to the HBS (FGSV, 2001). The optimal cycle length for demand 1 is 90 seconds, leading to degrees of saturation between 0.82 and 0.88, depending on the intersection. The optimal cycle length for demand 2 is 80 seconds, resulting in degrees of saturation between 0.78 and 0.84. Green splits vary among intersections. For coordination, the two relations with the highest traffic loads have been chosen, highlighted by bold arrows. The distance between intersection A and B is 280 meters. The distance from A to C and B to C is 320 meters. This results in optimal relative offsets of 20 seconds for A-B and of 23 seconds for A-C and B-C for a free flow speed of 50 km/h. The dashed bold lines indicate a secondary almost perfect coordination that arises unintentionally.

All transition methods described in paragraph 7.2.1 have been simulated in two scenarios:

Scenario 1: Traffic demand changes from demand 1 to demand 2. The first demand and its respective signal plans and coordination pattern are used in the first 30 minutes of simulation, the second demand and signal settings in the following 30 minutes. Thus transition starts after 30 minutes of simulation. The change of signal plans leads to necessary offset corrections of +27 seconds (which is equivalent to -53 seconds) at intersection A, +69/-11 seconds at intersection B, and +65/-15 seconds at intersection C.

Scenario 2: Traffic demand, signal plans and coordination pattern change from 2 to 1 (reversed order). This leads to offset corrections of +13/-77 seconds (intersection A), +55/-35 seconds (intersection B), and +65/-25 seconds (intersection C).

AIMSUN does not supply built-in signal plan transition methods. Its API had to be used to switch the signals directly in every simulation step. For each transition scenario the switching sequences of each signal group over the whole simulation period have been generated with a spreadsheet program and stored in text files. AIMSUN has been programmed to read these files at the beginning of each simulation run and to switch signals accordingly. This proceeding guarantees that the transition methods behave exactly as described earlier. Max dwell, Add with maximum lengthening and Subtract with maximum shortening have been constrained to 20% of the cycle length per cycle. At intersections where two competing directions are coordinated, the phase belonging to the heavier loaded coordinated relation is held during Dwell transition methods.

For evaluation of the transition methods, transient profile analysis as proposed by COHEN ET AL. (2007) has been used with intervals of 90 seconds. The performance measures of each time interval are average values of 30 replications. The first five intervals comprising the warm-up period are not used for evaluation. This means that transition starts at the beginning of the 16th interval in the diagrams presented in the next paragraph.

7.3.2 Results

Figure 7-3 to figure 7-6 show the results for both scenarios. The charts show average delay times of selected streams at different intersections and average travel times on selected relations. During the first 15 intervals, all curves are the same. Differences appear when transition starts in interval 16. Depending on the transition method, its impact is more or less prominent. The number of intervals that are influenced by transition varies. However, it can also be seen clearly that after a certain number of intervals performance settles down at a stable level and the curves become rather similar again, reflecting the new coordination pattern. A reduction in average travel time can be observed for those relations that change from an uncoordinated to a coordinated state. For relations with an increase in travel time the situation is reversed. All streams belonging to a relation that is coordinated after transition experience a reduction in average delay time, while those with an increase in delay time are uncoordinated after transition. Relation North-Southwest is uncoordinated both before and after transition. Therefore, stream B5 belonging to this relation has the same delay time before and after transition with some disturbances in between.

Relation Southwest-North is always coordinated, which leads to rather unchanged delay times of streams B11 and C8 before and after transition.

For the sake of clarity, the transition methods have been grouped in different figures. Figure 7-3 and figure 7-4 illustrate the performance of the rather abrupt transition methods, i.e. Dwell, Max dwell, Shortest dwell and Minimum green time. In figure 7-4 for scenario 2, the curves for Dwell and Shortest dwell are congruent because the shortest dwell time in three cycles is in the first cycle at all intersections, and thus Dwell and Shortest dwell are the same in this case.

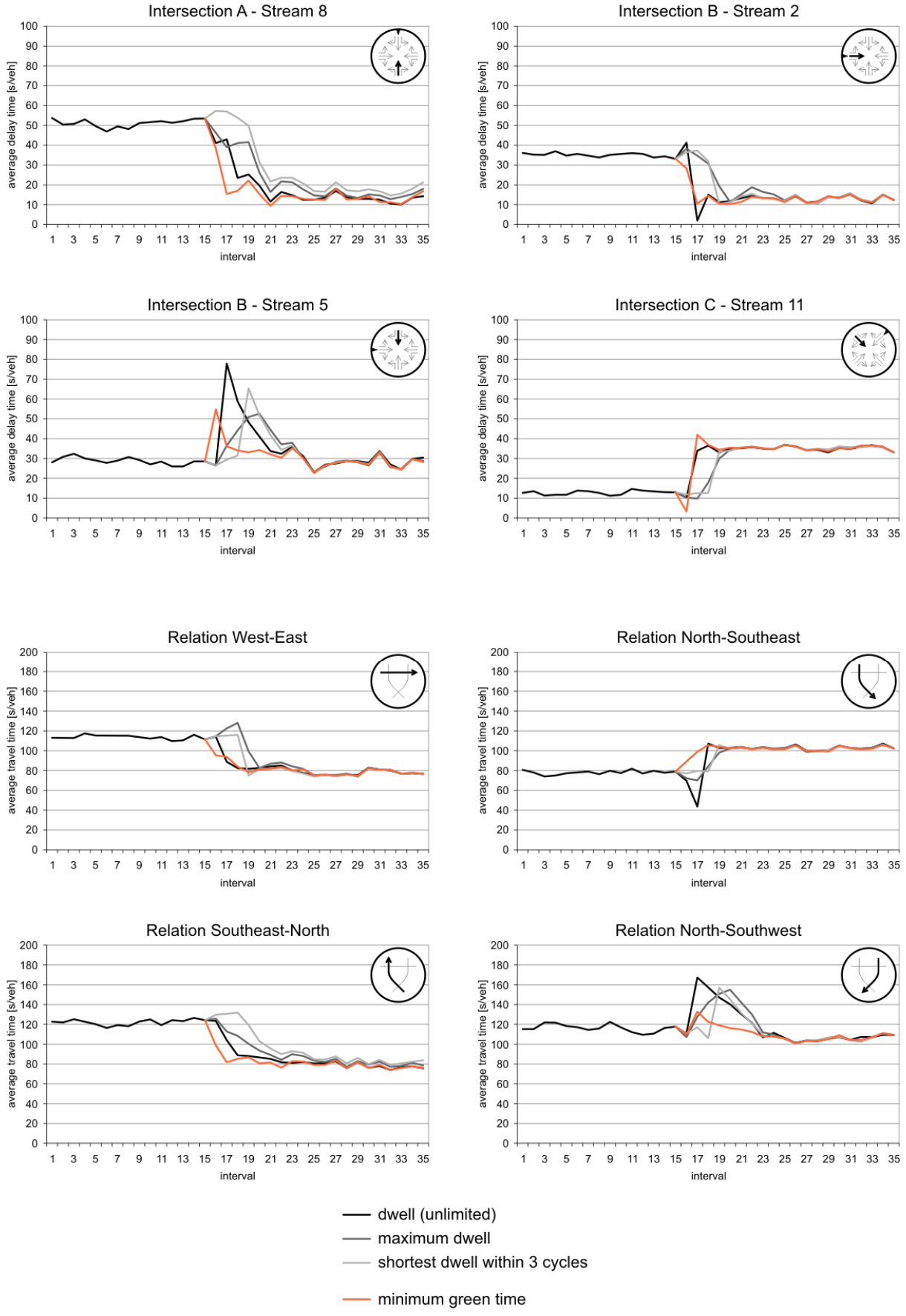


Figure 7-3: Results for Dwell and Minimum green time transitions in scenario 1

Figure 7-4: Results for Dwell and Minimum green time transitions in scenario 2

Figure 7-5: Results for Add and Subtract transitions in scenario 1

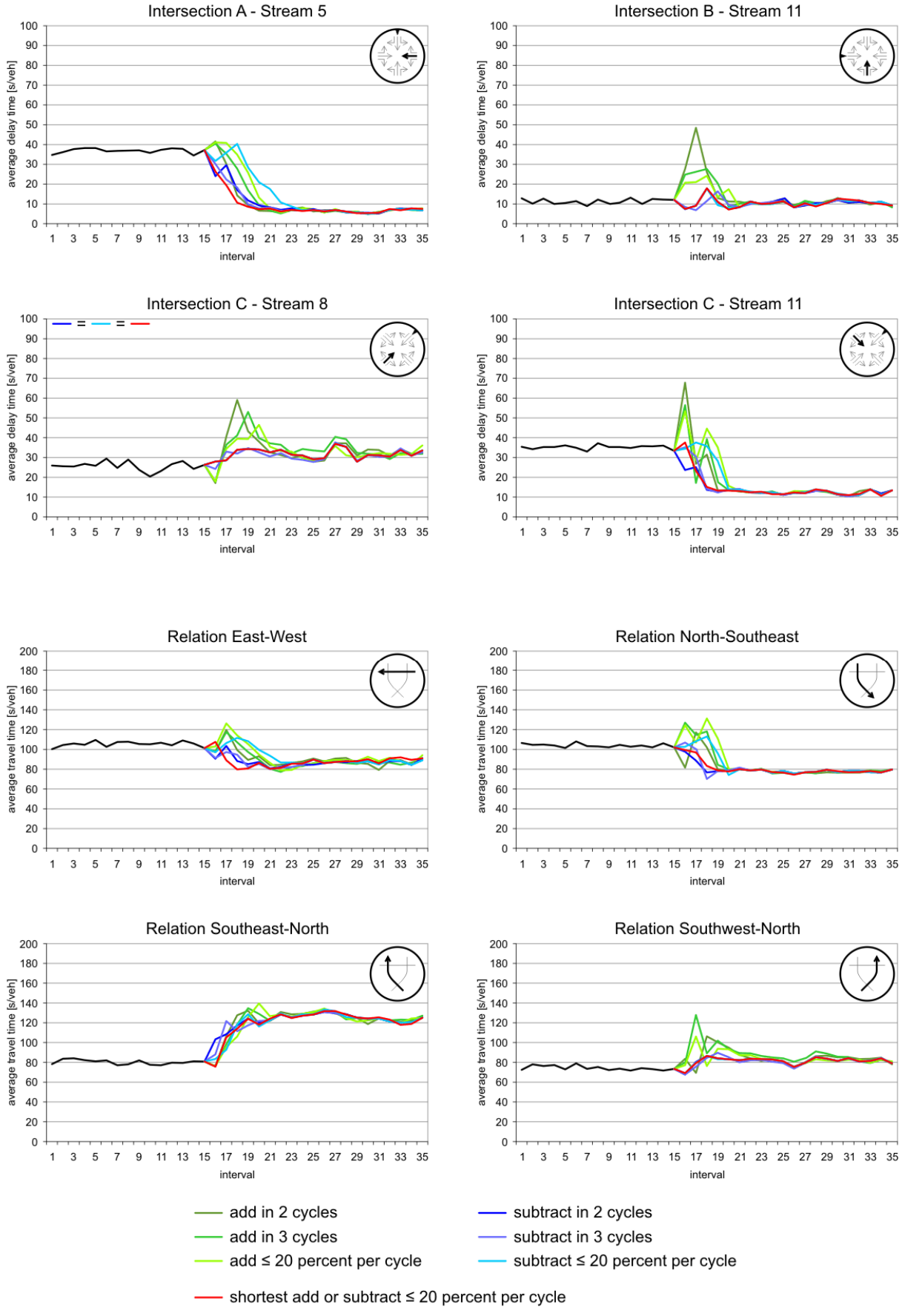


Figure 7-6: Results for Add and Subtract transitions in scenario 2

As can be seen the effect of Dwell can be very severe. High peaks occur at those streams experiencing long red times, and it may take several intervals until the disturbance is over (e.g. scenario 1: B5). In scenario 2 streams B11 and C11 have no green at all in interval 17, which is illustrated by a dashed line connecting the values of intervals 16 and 18. Other streams, however, benefit from a long green time and have a sharp drop of delay time (e.g. scenario 1: B2 in interval 17) or show a smooth decrease in waiting time without any peaks. A drop in waiting time may also be followed by a peak (e.g. scenario 2: A5). This peak is produced because dwell time at intersection A is much shorter than at intersection B, causing some vehicles to arrive at A during the red phase. The green wave on relation East-West is not established immediately.

The effects observed at intersection level may affect travel times on different relations. For example, the peak of stream B5 in scenario 1 directly influences travel time on relation North-Southwest, and the peak of stream 11 in scenario 2 is reflected in the peak of relation North-Southeast. For other relations however, the adjustment of travel time is rather smooth, even though Dwell favors immediate transition. This can happen, for instance, if Dwell periods are rather short at all concerned intersections of one relation or if the effects at both intersections even out. All in all, major disturbances at some streams have to be expected and this method is therefore not recommended for transition in coordinated networks.

If Max dwell is used, some peaks can be reduced or damped over several intervals (e.g. scenario 1: B5; scenario 2: C11). In other cases that have already performed well when simple Dwell was used, transition to the new coordination pattern is just slowed down by Max dwell without further benefit for these streams (e.g. scenario 1: A8, C11). Both effects can also be observed for travel times on different relations. Since even with Max dwell severe peaks cannot be avoided completely, this method should also be avoided.

Shortest dwell within three cycles may lead to smaller and delayed peaks (scenario 1: B5), but may as well just delay establishing the new coordination without further improvement (scenario 1: A8, B2, C11, relations West-East and Southeast-North). It cannot be recommended either.

The last of the abrupt methods is Minimum green time transition. In the cases depicted in figure 7-3 and figure 7-4 this method shows no real peaks, except for stream B5 in scenario 1 that gets only minimum green time in interval 16. Travel times on all relations quickly adapt to the new coordination pattern. However, high peaks may arise for those streams whose phase is possibly skipped. Nevertheless, among the immediate transition methods the minimum green time method seems to perform best.

Figure 7-5 and figure 7-6 show the performance of the smooth transition methods. Add in two cycles enables a smooth adjustment to the new coordination pattern for most streams, but also produces some peaks for a few streams (scenario 1: B2; scenario 2: B11, C8, C11). At the respective intersections, each cycle is lengthened by approximately 30 seconds, and therefore strongly extended green and red times arise. It seems that for these streams the negative effect of prolonged red times outweighs the positive effect of prolonged green times. Furthermore, B2 in scenario 1 and C11 in scenario 2 are coordinated streams after transition and in both cases the lengthening per cycle at the preceding intersection A is much shorter, resulting in platoons arriving during red phase and thus increasing delay time. A good coordination is only established after transition is over. Some travel time peaks can also be observed when looking at the

different relations. Generally speaking, Add in two cycles may also induce peaks if prolongation per cycle is too big.

In some cases Add in three cycles has no different effect than Add in two cycles. However, for some streams the previous peaks are reduced (scenario 2: C8, C11) or flattened over some more intervals (scenario 1: B2; scenario 2: B11). It can also be observed that this method has a tendency to shift the effects of Add in two cycles to following intervals (scenario 1: B5, C11; scenario 2: C8).

Add with maximum lengthening leads to a lengthening over 1 cycle at intersection A and over 5 cycles at intersections B and C in scenario 1 and again over 1 cycle at intersection A and over 4 cycles at intersection B and C in scenario 2. It can be observed that a quick adjustment of a coordinated stream at an intersection has no benefit if the preceding intersection reacts sluggishly (scenario 1: A8; scenario 2: A5) or vice versa (scenario 1: B2). An improper arrival pattern induced by a highly differing number of cycles for offset adjustment at adjacent intersections may also result in peaks (scenario 2: C11). All in all, adjustment of travel times on different relations appears to be sluggish, and not all peaks can be avoided.

Subtract in two cycles performs quite well for most streams depicted in figure 7-5 and figure 7-6. At intersections B and C shortening of cycles is not very distinct in both scenarios, and transition is fast and smooth. Stream A8 in scenario 1 suffers from a minor peak in delay time, that could arise from the fact that green time is too short and some vehicles have to wait for the next cycle to pass the intersection. However, this effect cannot be observed in scenario 2, where shortening of the cycles at intersection A is even a little bit stronger. Except for the mentioned peak, travel times on all depicted relations adapt smoothly and quickly to the new coordination pattern.

Subtract in three cycles is comparable to subtract in two cycles. The peak in stream A8 is reduced and transition is either equal or even a little bit smoother for the other streams. However, not much is to be gained by spreading the offset adjustment at intersections B and C over three instead of two cycles.

When it comes to Subtract with maximum shortening, similar effects as with Add with maximum lengthening (i.e. a certain sluggishness) can be observed. Offsets are adapted over four cycles at intersection A and only 1 cycle at intersection B and C in scenario 1 and over five cycles at intersection A and two cycles at intersections B and C in scenario 2. As can be seen, adjustment of delay times is slowed down again (scenario 1: B2; scenario 2: A5, C11). The same applies for travel time adjustments (scenario 1: relations West-East and North-Southeast; scenario 2: relation East-West and North-Southeast).

Obviously, using a smooth method like Add or Subtract has some potential in reducing peaks. However, if lengthening or shortening per cycle is too pronounced, peaks still occur, especially for the case of lengthening. If shortening or lengthening per cycle is limited, transition is predominantly smooth, but also tends to be sluggish. Therefore, Shortway tries to benefit from the positive effect of a limited lengthening or shortening while keeping transition short in order to establish the new coordination pattern quickly. Figure 7-7 shows the necessary length of transition against the necessary adjustment of offset, both expressed as a percentage of cycle length. The graph shows that offset adjustments of up to 40 percent of the cycle length are handled by lengthening while larger adjustments (that correspond to adjustments of up to -60 percent of

cycle length) are carried out by shortening. This guarantees that in any case, offset adjustment can be handled in no more than 259 percent of cycle length.

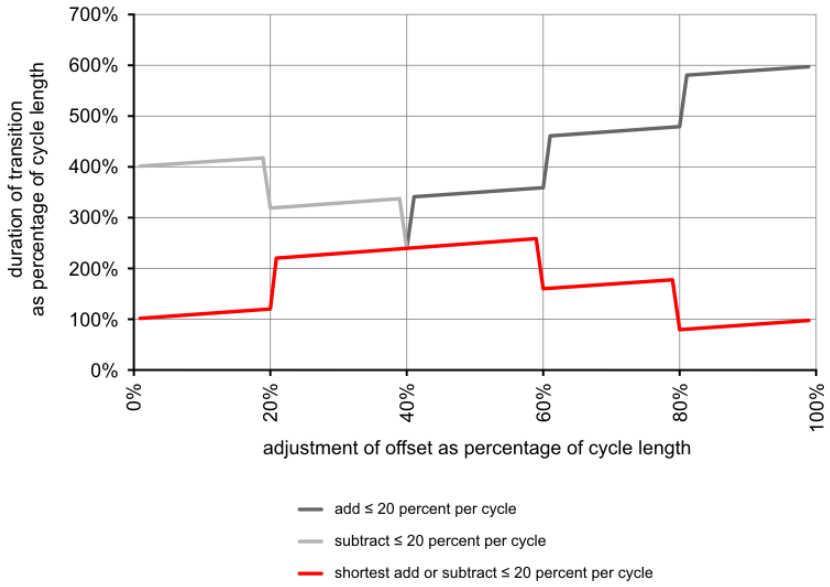


Figure 7-7: Necessary transition length of Add, Subtract and Shortway against offset adjustment

Figure 7-5 and figure 7-6 show clearly that Shortway does not produce any peaks at all in both scenarios and that adaptation of delay times and travel times is smooth and fast at the same time. Therefore, this transition method is recommended for coordinated meshed networks. This result strongly corroborates previous findings by SHELBY ET AL. (2006) and COHEN ET AL. (2007) for the arterial case. In addition, the study delivered further insights into system performance during transition.

7.4 Implementation into the ATCS prototype

According to the findings described in the previous paragraph, Shortway transition has been implemented into the framework of the ATCS prototype. Before the new signal settings are sent to the controllers, the necessary switching sequences from the old plans of the previous time interval are calculated according to the description of Shortway transition in paragraph 7.2.1. These transition sequences are also sent to the controllers.

In addition, Shortway transition can already be considered optionally during optimization. For this purpose the CTM has been programmed in such a way that it can also simulate the necessary transition from the old signal plans to the new ones, shifted according to the currently assessed chromosome or offset combination. Thus, the fitness or total delay of an offset combination also contains possible additional delays due to transition effects. Two possible solutions may perform equally well if transition is not considered, but if one of these solutions suffers from major transition disturbances which are taken into account during fitness evaluation, the GA is less likely to choose this solution over the other for recombination.

The consideration of transition during offset optimization implies that for each chromosome to be evaluated an individual switching sequence has to be generated. The amount of time to

calculate these sequences is very small compared to the computation time of one run of the CTM. The number of generations that can be iterated in 300 seconds is thus not reduced significantly. Dynamic transition methods as those proposed by MUSSA/SELEKWA (2003) and LEE/WILLIAMS (2009) however cannot be used in the framework of the ATCS prototype. These transition optimization techniques are time consuming and would have to be executed for each evaluated potential solution. LEE/WILLIAMS (2009) even use a GA for optimization of transition. If used in the framework of the ATCS prototype, this would result in a nested GA.

As explained in paragraph 6.3.3 each CTM run includes a warm-up period of three times the current cycle length. This can only be done if transition is not considered during optimization. The effects of transition are modeled at the beginning of each run of the CTM. Delays that are experienced in this time cannot simply be truncated. Another solution has to be found instead. If transition is taken into account, a first run of the CTM is executed before starting the optimization. During this run the signal plans of the previous time interval and the respective traffic demand are used. At the end of the run, all numbers n_i of vehicles in each cell i are stored. These numbers can then be used at the beginning of each following run of the CTM during optimization to preload the network. By doing so, no warm-up period is needed and the effects of transition can be included into the total delay.

In paragraph 6.3.3 it has also been mentioned that in this thesis all offsets of the n intersections in a network are considered as free variables even though in general one of them could be kept at zero to serve as a reference for the other offsets. This change has been made because of transition. If transition effects are taken into account during offset optimization, all offsets should be free to take any valid value. A specific offset combination might have a detrimental effect during transition. The same combination, only shifted by some seconds (i.e. the same value is added to all offsets which does not change the coordination pattern at all), might have a completely different and maybe less disruptive effect during transition. But if one offset is always fixed at zero, this second combination is not part of the solution space. Therefore, in this thesis the GA is allowed to modify the offsets of all intersections without exception to overcome this effect. No reference offset at one intersection is used.

8 Evaluation

8.1 Overview

This chapter describes the evaluation of the new ATCS prototype. A comprehensive simulation study has been conducted for this purpose. The ATCS prototype has been tested with different settings and for different degrees of precision of the input demand used for the optimization. All considered cases will be described in the following paragraphs.

Paragraph 8.2 covers the results for the test network in the List district of Hanover. The simulation setup has been described in paragraph 3.3. In addition, a second test network in the Südstadt district of Hanover has been used to evaluate the ATCS prototype under different circumstances. This network and the according results will be presented in paragraph 8.3. A summary of the basic findings is given in paragraph 8.4.

While the following paragraphs contain some selected diagrams that compare different test cases, appendix D comprises a much wider range of diagrams that compare travel times and number of stops not only for the overall sub-network, but also for all different routes. These diagrams will be referred to in the following.

8.2 Hanover List network

8.2.1 Modifications of the original simulation setup

8.2.1.1 Traffic demand

The first network that has been used for evaluation of the ATCS prototype is the List district network described in paragraph 3.3. The prototype optimized the common cycle length, phase durations and offsets of all ten signalized intersections and pedestrian crossings. All 56 time intervals of 15 minutes from 6 am to 8 pm have been simulated. The prototype optimized signal timings for all of these intervals consecutively.

First test runs of the simulation using the original fixed time signal plans and the given demand described in paragraph 3.3 revealed that none of the intersections came close to saturation even during peak hours. When the ATCS prototype was used, optimized cycle lengths were rather short even during peak hours. They covered only a small range of different values. In order to enable the prototype to make use of a wider range of cycle lengths, the entire daily

demand has been increased successively by up to 25 percent, i.e. the shape of the traffic demand profile displayed in figure 3-4 remains unchanged, but the amount of traffic during each interval is increased. This new demand leads to intersections close to saturation during both peak hours if the original fixed time signal plans are used. As desired, the ATCS prototype uses a wider range of cycle lengths after this adaptation.

Oversaturated time periods have still been avoided. The prototype in its present form is not able to handle oversaturated traffic conditions. While optimization is based on delay minimization, throughput maximization might be more appropriate during oversaturated conditions. Furthermore, demand estimation would be corrupted during oversaturation. Instead of the real demand, the loop detectors would only measure lane capacity during oversaturated periods. This would lead to improper constraints for the demand estimation.

8.2.1.2 Signalization

For evaluation of the new ATCS prototype, an adequate and incontestable reference case is needed. It would be best to compare the prototype to an existing ATCS (cf. paragraph 2.5.2). However, the software cores of existing ATCS are not publicly available in general and often difficult to calibrate. This task has to be performed by the vendors themselves. Therefore, fixed time signal plans are often used as reference cases in comparative studies. When doing so, the fact that fixed time control cannot react to a changing traffic demand must be considered when discussing the comparative results.

In this work, a fixed time control has been used as reference. However, the original fixed time signal plans could not be used because on the one hand they might be outdated, and on the other hand they have been optimized for a traffic demand that differs more or less severely from the one used in the simulation. As has been said, the traffic demand used in the simulation is semi-fictitious and may not represent the real traffic demand in all details. Furthermore, it has been increased by 25 percent. Therefore, two fixed time signal plans for both morning and afternoon peak hour have been generated using the established TRANSYT 7F (T7F) tool (HALE, 2005). They have been optimized for traffic demand during peak hours as used in the simulation. The original phases and phase sequences as used in the real network have been applied with only slight adaptations at some intersections (cf. appendix E for the final phases). Intersections in the network have either two or three phases. T7F was allowed to optimize the common cycle length, green splits and offsets of all controllers including the two pedestrian lights. T7F also offers a GA option which has been used for optimization. Total delay has been used as performance index to assess possible solutions.

Both resulting signal plans have a cycle length of 90 seconds. As in the real network, transition from morning to afternoon signal plans takes place at 1 pm, i.e. the first plan is used from 6 am to 1 pm and the second plan from 1 pm to 8 pm.

Figure D-1 in appendix D compares the performance of the original signal plans and the T7F signal plans during simulation. (The introducing paragraphs of appendix D explain how to read the diagrams). All values are averages over 30 AIMSUN simulation runs. Overall travel time [s/veh·km] and overall number of stops [-/veh·km] in the whole sub-network are reduced by the T7F signal plans at all time intervals. Improvements on single routes, however, are more distinct on the less heavily loaded routes. On some routes travel times and number of stops are even a little bit increased. The peaks that can be observed at the 1-pm-interval are caused by a rather

abrupt transition from morning to afternoon signal plans. AIMSUN has no inbuilt transition techniques and switches signal plans immediately. The switching of signal plans is also visible in the two route performance diagrams as a faint horizontal line.

8.2.2 Optimization for exact demand

8.2.2.1 Proceeding

In the first test case the exact traffic demand of each time interval has been used as input data for optimization. The idea behind this was to assess the capabilities of the two optimization modules described in paragraphs 6.2 and 6.3 without any impairment caused by imprecision of the demand estimation.

In a first step, 30 simulation runs using the original fixed time signal plans have been executed in AIMSUN. During each run the traffic volumes on all routes have been logged at 15-minute-intervals. Based on this data, average route volumes over all 30 simulation runs have been calculated for each interval. These volumes have been stored in a file.

Then the two optimization modules of the ATCS prototype were executed repeatedly for each of the 56 optimization intervals. During each execution the average route volumes of the respective time interval were read from the file. They were used to calculate the according link volumes by superposition. Based on these volumes the optimization modules determined the signal settings and stored the resulting switching sequences of all signal groups in another file.

Optimizing signal settings for all time intervals one after another took about 4:40 hours which is 5 minutes per interval. This amount of time is mainly consumed by offset optimization. Adjustment of cycle length and phase durations only takes a split second. The optimization process had to be done only once. The phase sequences used for optimization were the same as those of the reference T7F signal plans (cf. appendix D).

Finally, 30 simulation runs have been carried out that realized the stored, optimized switching sequences, i.e. during all replications the same switching sequences optimized for the average demand over all simulation runs have been used. AIMSUN has been programmed via the API to read the relevant switching sequences at the beginning of each time interval and to execute them accordingly. Smooth Shortway transition as presented in paragraph 7.2.1 is used to switch to the new signal settings at the beginning of each time interval.

8.2.2.2 Findings

Before discussing the overall performance, the cycle lengths adjusted by the ATCS prototype have to be examined. Cycle length has been calculated in two different ways. The saturation based approach has been used with a degree of saturation of $x = 0.85$ as recommended in the HBS (FGSV, 2001). This is the usual approach for coordinated arterials or networks. For comparison, the Webster formula has been used as well even though it is intended for isolated intersections only.

Figure 8-1 shows a comparison of the adjusted cycle lengths. They clearly follow the demand profile. Differences between the saturation based and Webster cycle lengths vary between 4 and 20 seconds. The Webster formula produces much higher cycle lengths.

The saturation based cycle length has a constant value of 49 seconds during the first six and the last four intervals. This corresponds to the minimum cycle length dictated by the intersection with the largest sum of minimum phase durations and intergreen times. As traffic increases, the cycle length is extended to up to 79 and 80 seconds respectively at peak intervals. This is still 10 seconds below the cycle length of the T7F fixed time signal plans serving as reference. The Webster cycle lengths reach 94 and 95 seconds respectively which is a little bit higher than the cycle lengths optimized by T7F.

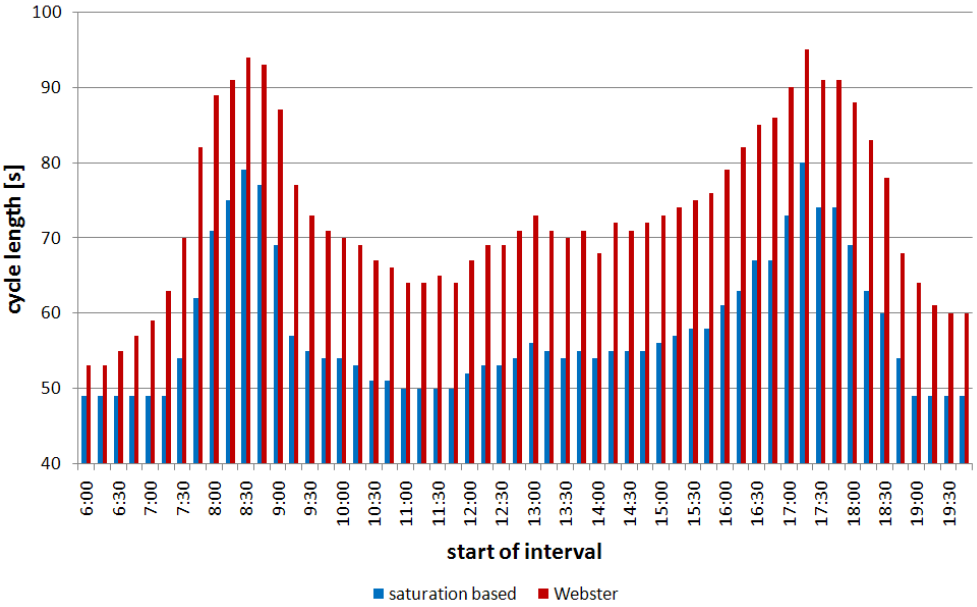


Figure 8-1: Cycle lengths adjusted for exact demand (List network)

Based on the exact demand, the ATCS prototype optimized signalization four times with different settings. The first variant only adapted cycle length and phase durations. No offsets have been optimized, i.e. all offsets were kept at zero which led to no intended coordination at all. The next three variants included additional offset optimization either with PGA, SGA or SE. Consideration of additional delays during signal plan transition has been included in the optimization. Figure 8-2 compares the temporal changes of overall travel time and overall number of stops in the sub-network. The diagrams show the overall results for the T7F reference case and the four ATCS variants. Again, all values are averaged over 30 AIMSUN runs. Even though the ATCS prototype has been designed to optimize total delay, overall travel times are displayed in all following diagrams. These travel times are directly influenced by the total delay and therefore reflect the achieved changes in total delay appropriately.

For the case of saturation based cycle lengths the ATCS prototype mainly produces smaller overall travel times than the T7F reference case even with offset optimization turned off. For off-peak intervals the prototype has the clear advantage of adapted cycle lengths which lead to much smaller travel times during off-peak intervals even without an underlying coordination. Only during morning peak hour the coordinated T7F signal plans perform better. However, a look on the overall number of stops reveals that the ATCS without offset optimization leads to a higher overall number of stops. Thus, vehicles have to stop more often but wait less due to

shorter cycle lengths. Figure D-2 shows that most routes are affected by significantly higher numbers of stops.

If offset optimization is turned on, overall travel times can be further reduced by 5 to 16 seconds depending on the interval. The comparative results in paragraph 6.3.5 surprisingly suggested that PGA, SGA and SE perform equally well. This assumption is confirmed by the simulation results. A slight advantage of PGA during some intervals might be assumed, but in general all three curves are virtually the same. The overall number of stops has decreased compared to the case without offset optimization and is now at the level of the T7F reference case. Obviously, coordination in terms of number of stops is already good in the reference case and the ATCS prototype is not able to find a better coordination pattern. The reductions in travel time must thus be mainly attributed to the shorter cycle lengths. However, figure D-3 compares the two cases of offset optimization turned off and on, which reveals that the concept of offset optimization used in this work performs well. In the coordinated case, travel times and number of stops are improved on the majority of routes, even though larger improvements on the heaviest loaded routes would have been desirable.

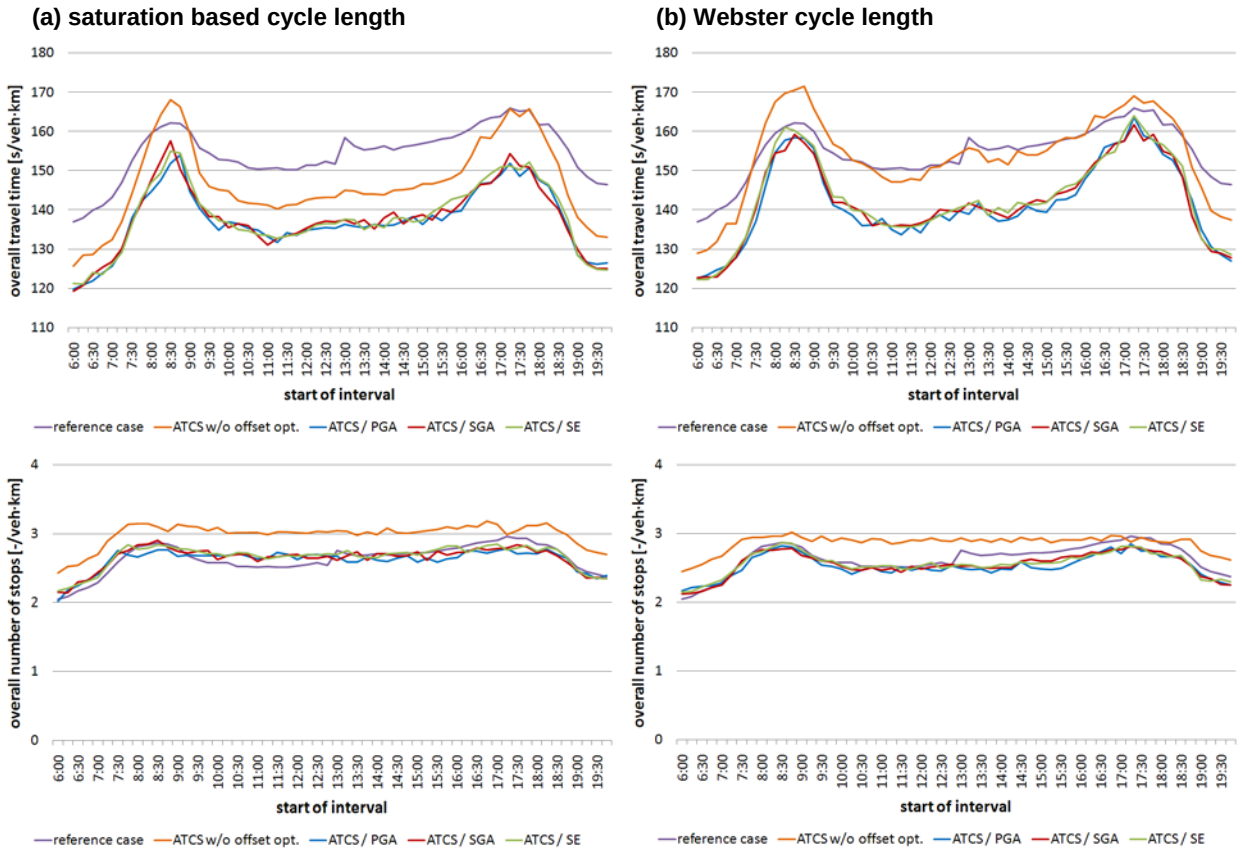


Figure 8-2: Overall performance of the ATCS prototype (exact demand, List network)

Figures D-4 to D-6 highlight the comparable performance of all three offset optimization techniques. The patterns of route travel times and number of stops are very similar. This implies that all three algorithms favor the same routes and obviously lead to similar coordination patterns. Almost all route travel times are improved even though improvements on the heavier loaded routes are less intense. Obviously, the good coordination pattern of the reference case leaves not much room for further improvement. But since volumes on the less heavily loaded routes

add up to an important share of the total traffic demand in the network, major improvements on these routes also contribute largely to an improvement of the overall performance. Concerning the number of stops, differences to the reference case are also more distinct on less loaded routes, where some more stops occur on some routes while other routes benefit from the ATCS prototype. Figures D-7 to D-9, too, show similarity of all three algorithms by comparing them directly with each other.

Figures D-10 and D-11 underline the importance of an adequate reference case. If compared to the performance of the original signal plans, the ATCS prototype produces even larger improvements of travel time and numbers of stops.

The right part of figure 8-2 shows the results for the case of Webster cycle lengths. They will not be discussed in the same detail. It can be seen that without offset optimization there are only few intervals where the T7F reference case performs worse. Switching on the offset optimization leads to travel time improvements between 5 and 17 seconds so that the overall travel times are better than in the reference case and the number of stops are comparable to the reference case. Again, PGA, SGA and SE perform equally well. More details are shown in figures D-12 to D-21.

Figures D-22 and D-23 compare the saturation based approach and the Webster based approach directly. Overall travel times are better if the saturation based approach is used, whereas the overall number of stops is better in the case of Webster based cycle lengths. Thus, the general assumption that less stops also lead to smaller delays is not necessarily true. Stops can be more numerous with shorter cycle lengths but the total waiting time spent in queues can be smaller. Figure D-23 reveals that the differences of travel times on different routes are not very distinct. However, the shorter saturation based cycle lengths lead to more stops. It might be that the longer Webster based cycle lengths enable better possible coordination patterns. But since the ATCS prototype is designed to reduce total delay only and not number of stops, the results are in accordance with this objective.

Paragraph 7.4 has shown how the effects of transition can be incorporated into optimization of offsets. Figure 8-3 compares consideration of transition to the case of its neglect during offset optimization with PGA. The blue curves are the same as the PGA curves in figure 8-2 where transition has been considered during optimization.

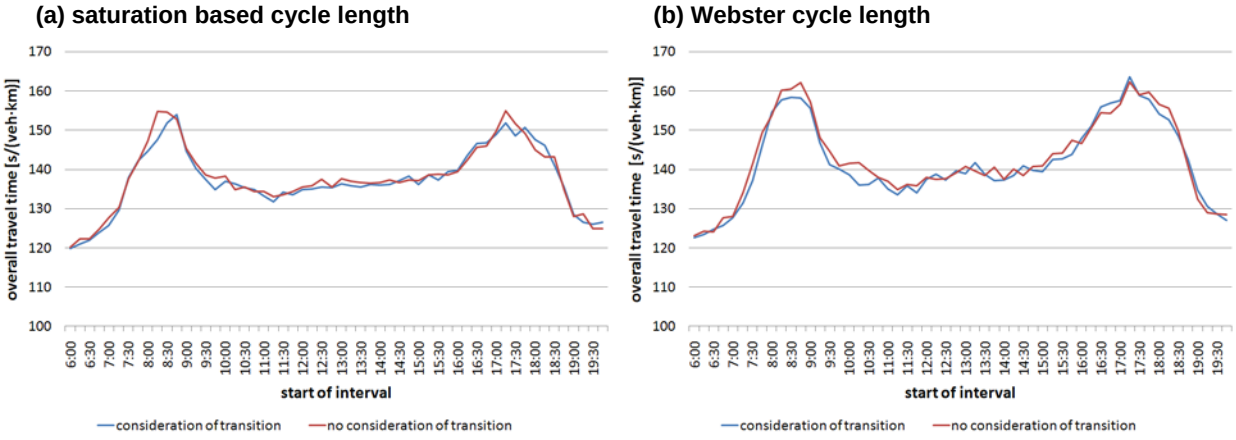


Figure 8-3: Overall performance with and without consideration of transition

A slight positive effect of consideration of transition on overall travel times can be observed, but it is not very distinct. Figures D-24 and D-25 show that the same applies to the overall number of stops. Variations of travel times and number of stops on routes seem to be more or less coincidental. However, one has to bear in mind the findings in paragraph 7.3.2 which have shown that Shortway transition produces no visible peaks in travel time at all. It is a very smooth and quick transition technique. Even if the effects of Shortway transition are not considered explicitly during optimization, this transition technique is used anyway to switch from the old to the new signal timings. Thus, traffic flow can still benefit from its smoothness and does not suffer from major travel time peaks in this case either.

Another analysis of the two cases is shown in figure 8-4. For all six combinations of saturation and Webster based cycle length on the one hand and PGA, SGA and SE on the other hand, two complete optimization runs over all 56 time intervals have been performed, one with consideration of transition, the other one without. Each optimization comprises 55 transition periods. Given the ten intersections in the network, 550 decisions per optimization had to be made on whether to shorten or to lengthen the transition cycles. In a few cases, the necessary adaptation of offsets is zero and no shortening or lengthening is needed. Figure 8-4 compares the portions of these three transition types. In all six cases more shortenings of cycles are used if transition is considered during optimization. As has been seen earlier, shorter cycles often lead to shorter travel times and thus it is consistent to use more shortened cycles if delays are to be reduced. On average the portion of shortened cycles increases from 53.2 to 63.0 percent whereas the portion of lengthened cycles reduces from 44.7 to 33.8 percent.

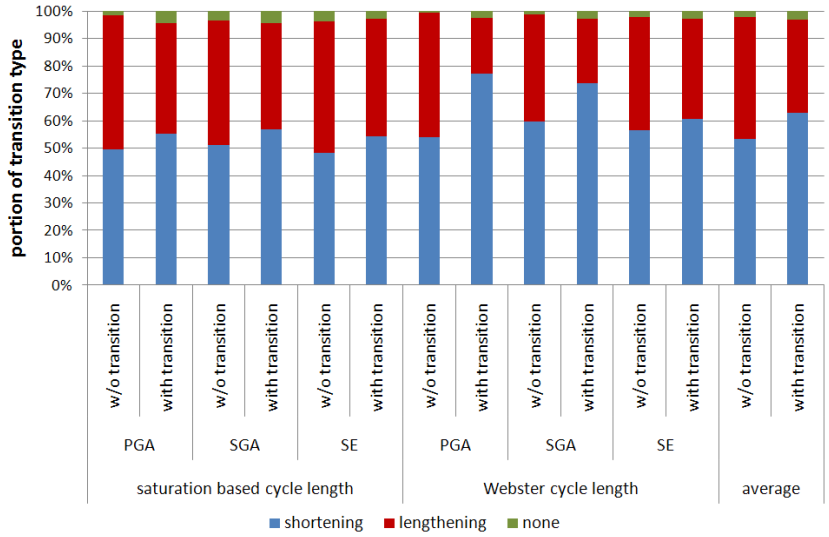


Figure 8-4: Portions of transition type with and without consideration of transition

8.2.3 Optimization for estimated demand based on average detector counts

8.2.3.1 Proceeding

Until now the optimization modules have only used the exact average demand of each time interval as input data. In a next step, the forecasting and demand estimation modules have been used in addition. As has been described earlier, forecasting and demand estimation rely on detector data as input data. In this first attempt, average detector data of each time interval

has been used that has been produced in the same way as the average route volumes described in paragraph 8.2.2, i.e. from logged detector counts of 30 AIMSUN runs simulating the original fixed time control.

Again, the modules of the ATCS prototype had to be executed 56 times. For each interval, the average detector counts of the four preceding time intervals have been read. They served as space-time-pattern to forecast the detector counts of the current optimization interval. Since the forecasting algorithm needs data from four previous time intervals and moreover looks two intervals ahead, the first optimization interval whose counts could be forecasted is the sixth time interval. For intervals 1 to 5 the average detector counts of these intervals have been used directly as constraints for the route and link volume estimation without any forecasting. The reference pattern needed for forecasting has also been derived from 30 simulation runs using the real fixed time plans. This implies that in this test case the quality of forecasting must be considered to be perfect.

Based on the forecasted counts the demand estimation and subsequent optimization of signal timings could be executed. As for the first test case, switching sequences for all intervals have been stored in a file and were read by AIMSUN at the beginning of each interval during each of the 30 simulation runs. Again, the same switching sequence has been used for all 30 runs. Transition effects have been considered during optimization.

8.2.3.2 Findings

During demand estimation, either BPR functions or the CTM can be used to estimate travel times needed for traffic assignment. Figure 8-5 compares the resulting saturation based cycle lengths of both cases. If BPR functions are used, cycle lengths are slightly higher than with CTM travel times. Due to imprecision of demand estimation the cycle lengths in both cases tend to be a little bit higher than in the case that the exact demand is used as input data for optimization, especially during peak hours.

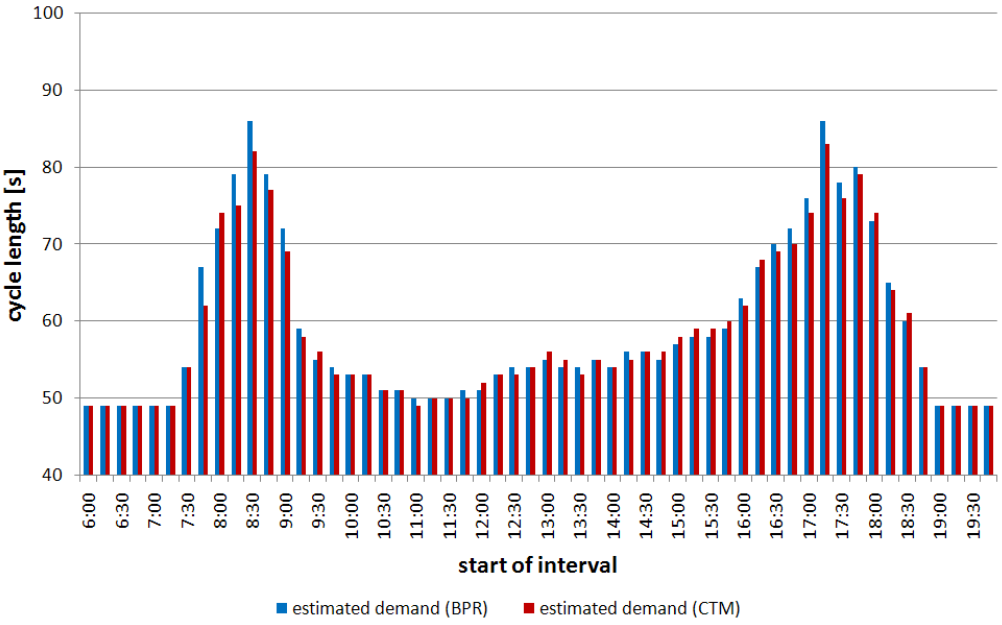


Figure 8-5: Cycle lengths adjusted for estimated demand (List network)

Figure 8-6 compares overall travel times of the three cases of optimization for exact demand and estimated demand with BPR and CTM travel times. PGA has been used for offset optimization. The blue curve corresponds to the curve of the same color in figure 8-2. Travel times in all three cases are comparable. During peak hours, optimization for the exact demand performs slightly better due to the lower cycle lengths, but all in all the differences are negligible, i.e. the estimated demand based on average detector counts is sufficiently good to be used as input data for the optimization modules.

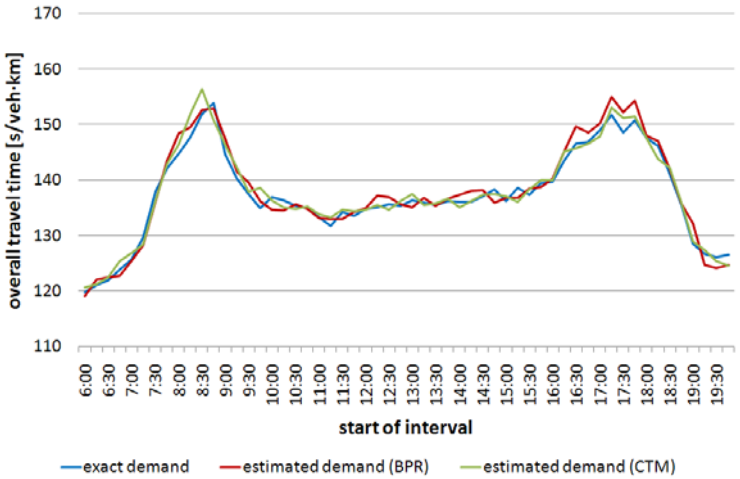


Figure 8-6: Overall performance with optimization for exact and estimated demand

Figure D-26 compares the T7F reference case to the ATCS prototype using BPR travel times for demand estimation. The two patterns of route travel time and numbers of stops are very similar to the case of exact demand. Figure D-27 compares these two cases directly and also reveals their similarity. Finally, figure D-28 shows again that using BPR or CTM travel times does not lead to major differences. Both approaches are appropriate with no apparent superiority of the one over the other.

Figures D-29 to D-31 examine the case of Webster based cycle length. All observations mentioned for saturation based cycle lengths also hold for this case.

Another question of interest was whether forecasting of detector counts contributes to the overall performance of the ATCS prototype. Therefore, in addition to the aforementioned proceeding where forecasted detector counts have been used as constraints for the demand estimation, a second optimization run has been performed that used the average detector counts of the last complete time interval as constraints for the demand estimation. This implicates that signal settings of an optimization interval are optimized for an estimated demand that is based on detector counts of the interval which lies two intervals in the past. For the first two intervals, the detector data of these intervals was used since no data from preceding intervals had been available.

Figure 8-7 shows the overall travel times for the two cases with and without forecasting. BPR functions have been used for traffic assignment. The blue curve is the same as the red curve in figure 8-6. Travel times increase more steeply before the morning peak hour because of cycle lengths which are too low for the rising demand. The decrease of travel times after this peak hour is delayed. Figure D-32 shows that during morning peak hour the number of stops is in-

creased significantly. During afternoon-peak hour the delayed adjustment of cycle lengths seems to be beneficial at first, but finally the cycle lengths are too short and lead to higher travel times in the end which are reduced with some delay only. Figures D-32 and D-33 show some more details of the effects without forecasting. Even though traffic does not crash, the diagrams show the retarding effect that arises if no forecasted detector counts are used for demand estimation. It is therefore recommended to include forecasting into the demand estimation, at least if time intervals are rather long as in this thesis.

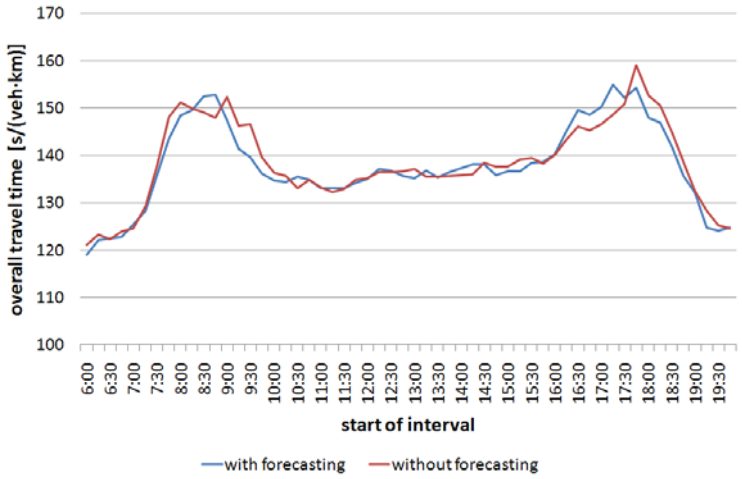


Figure 8-7: Overall performance with and without forecasting

8.2.4 Real online optimization

8.2.4.1 Proceeding

This final test of the ATCS prototype comprises its online application. The prototype is used in the same way as it would operate in a real network. No average data is used, but each simulation run in AIMSUN is optimized individually.

In this test case, AIMSUN and the ATCS prototype exchange data directly during each simulation run. At the beginning of each time interval, AIMSUN gets the optimized switching sequences for each signal group from the ATCS prototype. At the end of each time interval, AIMSUN provides the ATCS prototype with recent detector counts. The ATCS prototype stores these counts internally in order to generate space-time-patterns covering the last four intervals. For forecasting, the same reference space-time-pattern as in the previous paragraph is used. Again, the problem remains that no forecasting is possible for the first five intervals. For these five intervals the same average detector counts as in paragraph 8.2.3 are used instead of forecasted counts.

The simulation in AIMSUN and the ATCS prototype are executed in parallel. As has been said before, execution of all four modules of the prototype takes about 5 minutes for one optimization interval. Only a few seconds thereof are consumed by forecasting and demand estimation. AIMSUN simulates a 15-minute-interval in about one second. This implies that AIMSUN spends most of the time waiting while the ATCS prototype is working. The duration of a simulation run is thus dictated by the ATCS and amounts to the previously mentioned 4:40 hours. Therefore, only

10 simulation runs have been carried out in this case, and only saturation based cycle lengths and PGA have been considered.

8.2.4.2 Findings

Since all simulation runs are optimized individually, the cycle length of single intervals varies among these runs, depending on the achieved quality of demand estimation. Figure 8-8 shows box plots including outliers (circles) and arithmetic means (black diamonds) of these variations. While variations are small during off-peak intervals, they are partly very severe during peak hours. No average detector counts are used as input data in this test case but those taken from single simulation runs with higher variations. This makes it harder for the forecasting module to identify the correct reference sub-pattern, and therefore the quality of the forecasted detector counts is reduced. This effect is more distinct, when traffic demand changes more rapidly. The reduced quality directly influences the performance of the subsequent demand estimation and consequently of the cycle length adjustment as well.

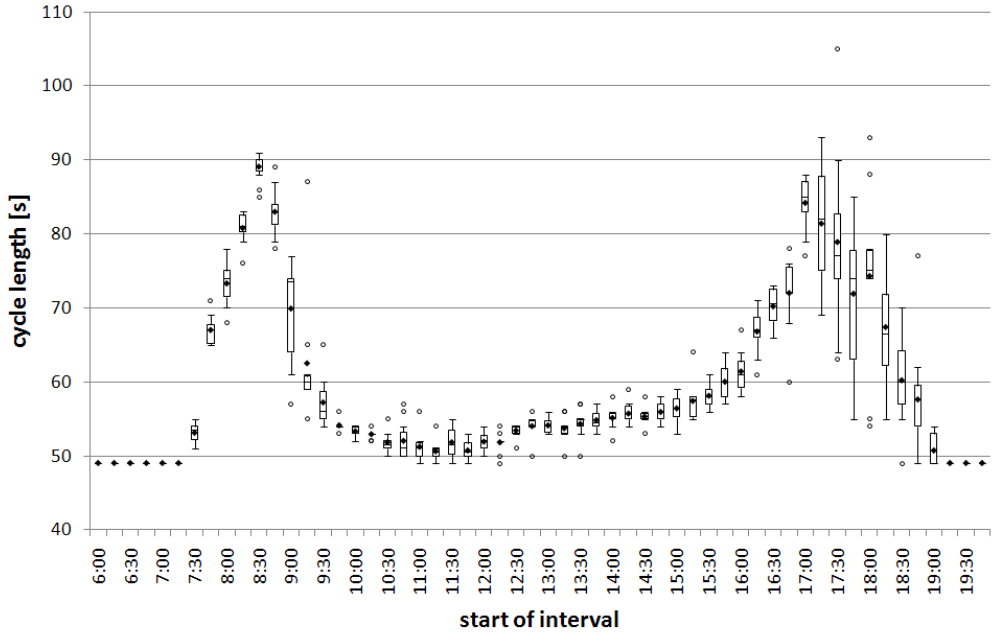


Figure 8-8: Variation of cycle lengths during online optimization (List network)

On the average, however, subsuming all 10 simulation runs, the overall performance of the ATCS prototype is still good. Figure 8-9 compares the online case to the two cases of optimization for exact and estimated demand based on average detector counts. During peak hours the online application of the prototype performs slightly worse on average. Figures D-34 to D-36, however, show that all in all the average travel times and number of stops are comparable to the previous cases. But it has to be highlighted that these are only average results over all 10 AIMSUN runs. Given the sometimes badly adjusted cycle lengths, a temporarily reduced quality of traffic flow during some of the simulation runs must be assumed.

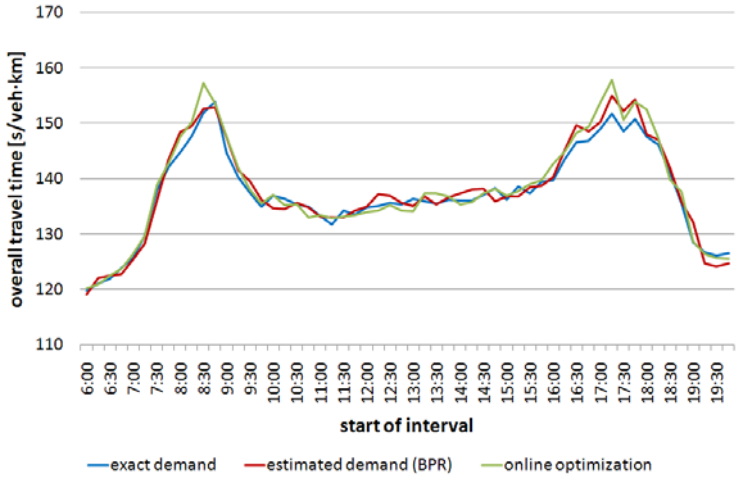


Figure 8-9: Comparison of optimization for exact and estimated demand and online optimization (List network)

8.3 Hanover Südstadt network

The ATCS prototype performed well in the List district network, but this does not allow for generalization of the findings. Therefore a second test network has been set up to assess the performance of the ATCS prototype under different circumstances. This time only saturation based cycle lengths have been used and transition effects have always been considered during optimization.

8.3.1 Simulation Setup

8.3.1.1 Specifics of the network

The second test network is in the Südstadt district of Hanover, south of the city center. It is shown in figure 8-10.

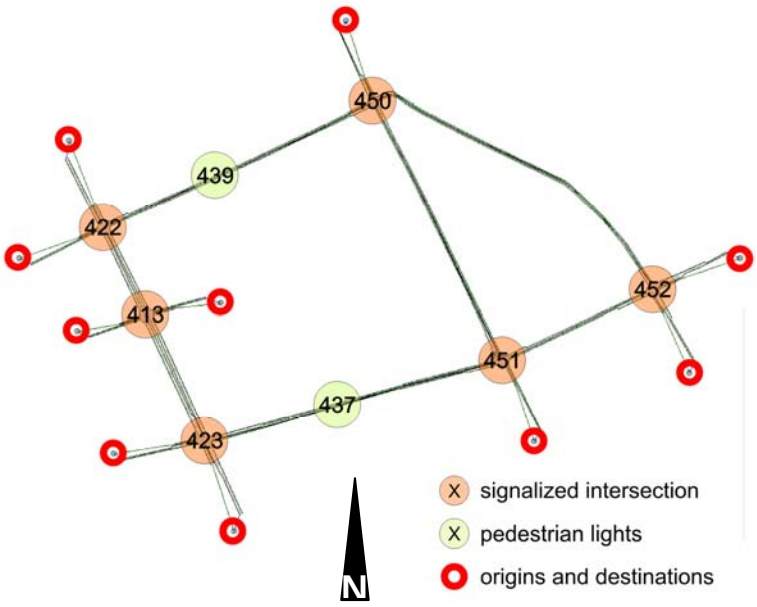


Figure 8-10: Test network in Hanover Südstadt district

The network has six signalized intersections, two pedestrian lights and a total of 52 detectors. Traffic signal 413 is only a pedestrian light in reality, but it has been upgraded to an entirely signalized intersection in the simulation.

The road passing intersections 423, 413, and 422 has two lanes per direction and is important for commuters. All other streets have one lane per direction. Most streets have a speed limit of 50 km/h. The eastern and western side streets at intersection 413 and the southern approach of intersection 451 have speed limits of 30 km/h. The tramway operates underground in the network. Some bus routes cross the network, but they have not been included into the simulation.

8.3.1.2 Traffic Demand

Traffic counts were available for the network at the author's research institute. They had been collected in other projects. However, data for all intersections was only available for the morning peak hour. Counts for the afternoon peak hour cover intersection 423 only. Based on the counts of the morning peak hour, an OD matrix has been derived that matches these counts subject to manually chosen route choice probabilities. 10 origins and destinations were used which are connected by 89 reasonable routes. For the afternoon peak interval, a second OD matrix has been defined based on some assumptions about variations of commuter flows. A daily demand has been set up next as has been done for the List network. It covers the time period between 6 am and 8 pm. The 56 single OD matrices for each interval are again linear combinations of the two peak interval matrices. The same shares of the two matrices over time as in the List network have been used. Finally, the resulting overall traffic demand has been increased by 10 percent afterwards for the same reasons that lead to an increase of 25 percent in the List district network.

During morning peak hour, the main flow is on the South-North relation on the aforementioned two-lane-road. In the afternoon the main flow travels in the opposite direction. However, the road is highly loaded in both directions at all times. Other important flows are on the East-West and West-East relation passing intersections 423, 451, and 452.

8.3.1.3 Signalization

The real fixed time signal plans were available for this network, too. Again, the original morning and afternoon signal plans have a common cycle length of 90 seconds, and signal plan transition occurs at 1 pm. Based on the real phases and phase sequences (cf. appendix E), two T7F signal plans have been optimized. Their cycle lengths are shorter than in the first test network. The morning signal plan has a cycle length of 75 seconds, and the afternoon signal plan of only 70 seconds. Figure D-37 shows that the T7F signal plans perform better than the original signal plans, given the demand fed to the simulation.

8.3.2 Optimization for exact demand

Figure 8-11 compares the overall performance of the ATCS prototype to the reference case with T7F optimized signal plans. It can be seen clearly that mere cycle length and phase duration adjustment without offset optimization performs worse than the reference signal plans at all time intervals. No benefit is drawn from adjusted cycle lengths since the cycle lengths of the T7F signal plans are already rather short. Figure D-38 shows that travel times and number of stops along the heaviest loaded routes are worsened by the ATCS prototype.

If offset optimization is turned on additionally, the ATCS prototype can achieve some improvements. During the morning peak hour, however, the reference case is still better. In this second network the heaviest loaded roads are very distinct and do not change much over time. The reference signal plans are obviously well adapted to these flows and a continuous re-optimization of signal plans and especially offsets does not lead to major improvements. Due to its characteristics the second network obviously does not benefit very much from a constant adaptation of signal settings.

Figures D-39 to D-40 show that all three offset optimization techniques perform almost equally well and produce similar patterns of route travel times and numbers of stops. On the two heaviest loaded routes, which correspond to the opposite directions of the two-lane-street, no improvements and even some minor degradation of travel times and numbers of stops are produced by the ATCS prototype. Since both directions compete against each other and have a high weight in the fitness function due to the high traffic loads (the reader is reminded that total delay is given in veh·s), none of the two directions benefits from a well functioning green wave.

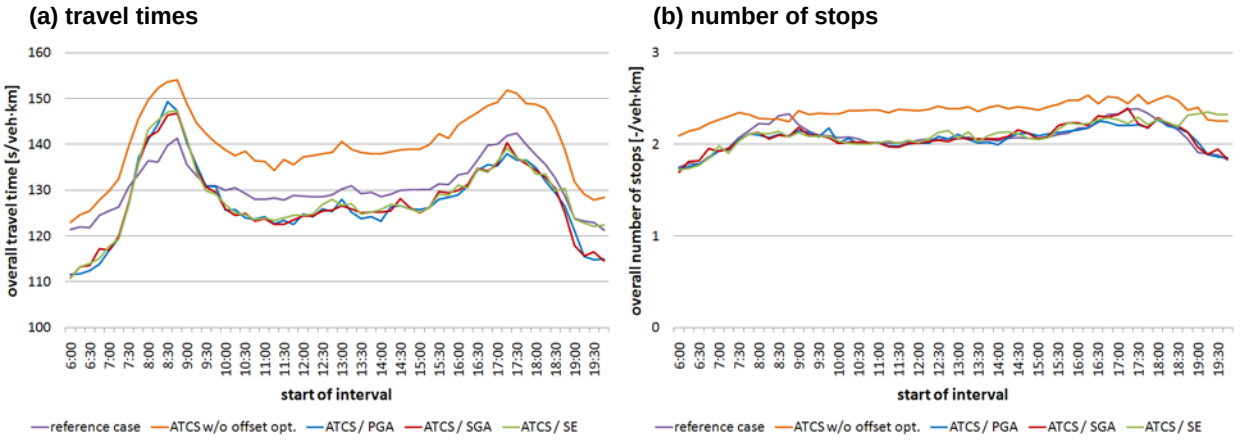


Figure 8-11: Overall performance of the ATCS prototype (exact demand, Südstadt network)

Figure D-42 confirms again the suitability of the concept of offset optimization used in this thesis even though the prototype does not perform as well as in the first network. Overall travel times and number of stops are lower if offset optimization is turned on. Figures D-43 to D-45 further confirm that all three offset optimization techniques lead to comparable results. However, it can also be seen clearly that SE has some deficiencies during the last intervals. Travel times are not reduced to the same level as with PGA and SGA and the number of stops stays as high as during the afternoon peak hour. Even though this special case has been analyzed carefully, the cause for this behavior of SE could not be revealed. For some reason SE does just not manage to find a good coordination pattern during the last few intervals.

Figures D-46 and D-47 show again the importance of a suitable reference case. If compared to the original signal plans, the ATCS prototype performs much better, at least if offset optimization is turned on.

8.3.3 Optimization for estimated demand based on average detector counts

Only BPR travel times have been used in the second test network to optimize signal settings based on an estimated demand. In accordance with paragraph 8.2.3 figure 8-12 shows that cycle lengths are slightly higher if the optimization is based on an estimated instead of the exact

traffic demand. Figure 8-14 and figures D-49 and D-50, however, show that the overall performance of the ATCS prototype is not much corrupted if an estimated demand based on average detector counts is used. This confirms the findings in paragraph 8.2.3.

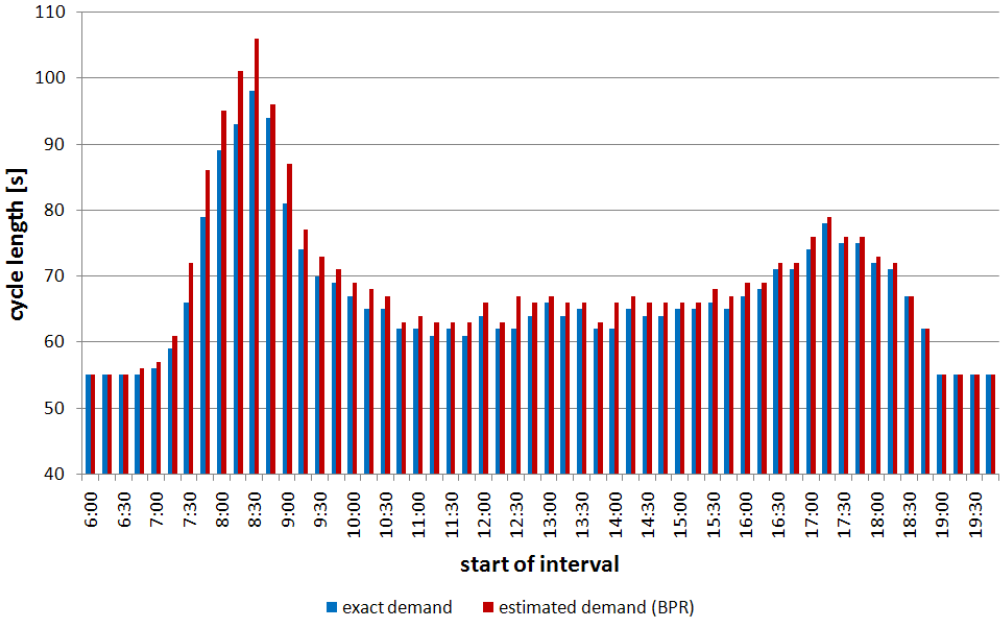


Figure 8-12: Cycle lengths adjusted for exact and estimated demand (Südstadt network)

8.3.4 Real online optimization

Finally, online application of the ATCS prototype has also been tested for the second test network. Again, variations of the cycle lengths over all 10 simulation runs could be observed. They are shown in figure 8-13. They depend on the time interval but are in general less severe than in the first test network.

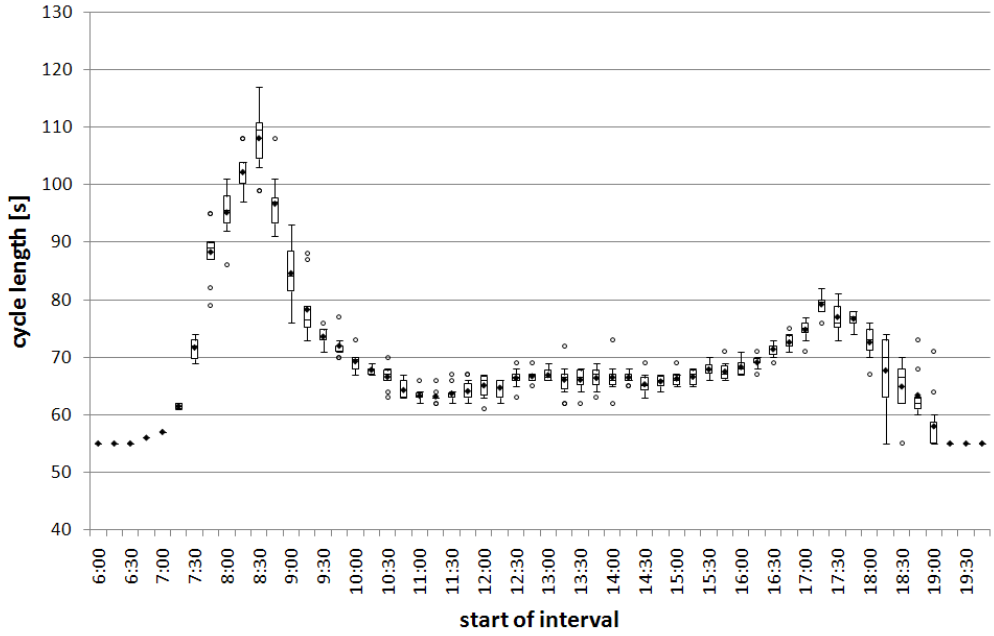


Figure 8-13: Variation of cycle lengths during online optimization (Südstadt network)

Figure 8-14 and figures D-50 to D-52 show that on average the online application is almost as good as the previous cases with exact demand and estimated demand based on average detector counts. Again, the variations of cycle lengths suggest that major differences among single simulation runs have to be assumed.

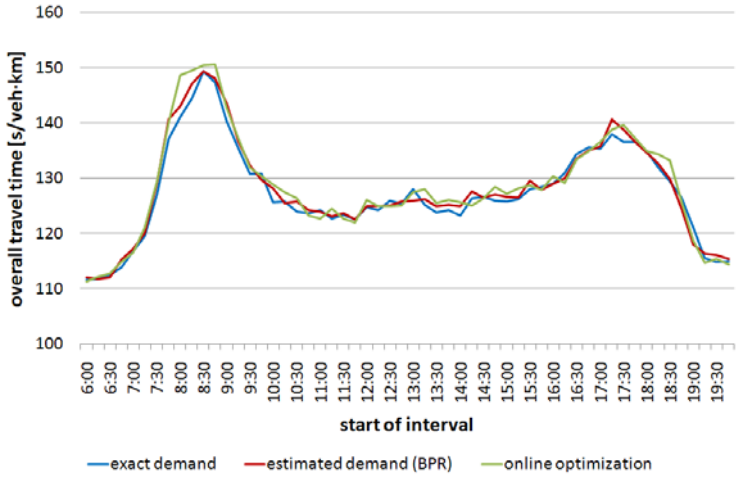


Figure 8-14: Comparison of optimization for exact and estimated demand and online optimization (Südstadt network)

All in all it has become evident that the ATCS prototype still works fine in the second network, however, its general applicability (or applicability of any ATCS at all) in this network is questionable. The reference signal plans are already well adapted to the major traffic flows which are very constant over the day. A further continuous adaptation of signal settings does not have a major positive impact. During the morning peak hour the ATCS prototype even worsens the overall travel times and number of stops.

8.4 Summarized findings

The general findings of the study for evaluation of the performance of the ATCS can be summarized as follows:

- An appropriate adjustment of the cycle length contributes significantly to a reduction of delay, especially during off-peak hours. This underlines the general benefit that can be gained from adapting the cycle length to the current average traffic demand.
- A mere adaptation of cycle length and phase durations without offset optimization may reduce delays depending on the reference case, but the number of stops must be expected to increase due to a missing coordination pattern.
- In comparison to the case of mere adjustment of cycle lengths and phase durations without offset optimization the model-based offset optimization is able to further reduce delays and even the number of stops. This shows the general usefulness of the model-based approach that is used in this thesis. It confirms the overall result of the work by ALMASRI (2006).
- No major differences between PGA, SGA, and even SE for offset optimization could be observed, which contradicts some of the findings of ALMASRI (2006). This observation leads to the assumption that a good coordination of signalized intersections which is close to the

system optimum can already be identified by means of a rather simple approach for offset optimization. The application of GA might not be necessary, at least in networks of the size used in this thesis.

- The number of stops could not be further reduced compared to the fixed time signal plans optimized with T7F. The latter obviously already establish a good coordination pattern.
- Webster cycle lengths are in general longer than saturation based cycle lengths. This leads to higher delays but also to less stops, i.e. vehicles have to stop less often but they wait longer at each intersection if longer cycle lengths are applied.
- Attainable improvements of delays and number of stops are less distinct on the heavier loaded routes. A reason for this might be that offset optimization which searches for a system optimum that favors all routes does not always and not necessarily produce perfect progressive signal systems for the heaviest loaded routes. The PI or fitness used for optimization in this thesis is total delay in veh·s. To some extent this indicator takes the number of vehicles on each route into account. However, many routes with few vehicles may outweigh a few heavily loaded routes, and thus no perfect coordination will be created for the latter. This observation confirms the findings of MÜCK (2008a, 2008b) who stated that the GA-based approach for offset optimization implemented in MOTION does not produce real “green waves” along the whole stretch of arterials.
- Shortway transition performs well and helps to avoid major disturbances when the new signal plans are established at the beginning of each time interval. The positive effect of Shortway transition is noticeable even if the effects of transition are not explicitly considered during offset optimization. Problems during transition which originate from randomly changing coordination patterns as assumed by BRAUN (2008) and stated by MÜCK (2008a, 2008b) have not been observed in this thesis.
- The performance of the ATCS prototype is only reduced slighty if an estimated instead of the exact traffic demand is used for optimization of signal timings. Cycle lengths increase only slightly. Whether travel times are estimated by BPR functions or by the CTM during traffic demand estimation makes no significant difference. Both ways are appropriate.
- If no forecasting of detector counts is used, a retarding effect can be observed which does not lead to total crashes of traffic flow but which may reduce the quality of traffic flow at least during and shortly after peak hours. If intervals are rather long as in this thesis, forecasting should not be neglected. No conclusion can be drawn for ATCS that can use shorter optimization intervals such as BALANCE or MOTION.
- The average performance of the ATCS if used as a real online application is comparable to the cases which use an exact or average estimated demand. However, different AIMSUN runs resulted in major variations of the cycle length at some time intervals, especially during and after peak hours when traffic demand changes more heavily. This observation suggests that the forecasting technique is not always able to find good estimates of the detector counts of an interval, which corrupts the demand estimation and in consequence also the optimization of signal timings. A worsened quality of traffic flow during single AIMSUN runs is therefore very likely.

- Results from the second test network underline the influence of specific network characteristics on the usefulness of the application of ATCS. An ATCS which performs well in one network might not be able to produce similar improvements of traffic flow in another network. This is obviously true for networks where the heaviest loaded routes do not change over time. This leaves not much room to further improve an already optimized set of offsets. Furthermore, the cycle lengths of the two T7F signal plans of the second network were rather short, which makes it harder for the ATCS to further reduce delay by adaptation of the cycle length.

9 Summary and outlook

9.1 Summary

In this thesis a new ATCS prototype has been developed and evaluated that is directly based on two previous research projects. The first of these two projects (ALMASRI, 2006) dealt with model-based offset optimization for signalized intersections in an urban network under the regime of traffic signal control. The project used GA in combination with the CTM. The second project (WANG, 2008) dealt with the estimation of OD matrices and the influence of redundant estimation constraints, i.e. redundant traffic counts, on the quality of the estimation result. Several rules had been developed in this project to eliminate redundant constraints. Furthermore, the process of OD matrix estimation had been incorporated into an iterative procedure of traffic assignment and matrix estimation until convergence of the estimated matrix on a stable solution is achieved.

This thesis started with an overview of the state-of-the-art of science and technology of traffic signal control. Some fundamentals on fixed time and vehicle actuated control have been provided. A concise presentation of different existing ATCS and their performance, which had been investigated in different field tests, has been given. Finally, examples of the application of GA and/or the CTM in the context of urban traffic signal control have been shortly discussed. Since no publications on the application of GA in combination with the CTM in the framework of an online control strategy had been available so far, this thesis addressed such an online application to close this gap.

Before discussing all modules of the new ATCS prototype in detail, an overview of the conceptual design of the prototype has been given. It applies a centralized concept, i.e. signal timings of all signalized intersections are optimized on a central computer. The optimization interval of the strategy is 15 minutes. At the end of each interval, the signal controllers at each intersection send current counts from their detectors to the ATCS central computer where they are used as constraints for the estimation of the current traffic demand. Based on this estimated demand, the ATCS adjusts a common cycle length, phase durations and offsets of all intersections. After termination of this optimization procedure, the new signal timings are sent back to the controllers at the beginning of the next time interval. There, the new signal timings are implemented immediately.

One of the basic elements of the ATCS is the traffic flow model which is used to estimate the impact of different offset combinations in terms of total delay. The CTM has been used for this purpose. The original model equations have been presented thoroughly, followed by some extensions developed explicitly in this thesis to enable better modeling of urban intersections of arbitrary geometry. Modeling of permitted left turns has also been included into the CTM. Besides the model-based estimation of delays, an additional approach has been developed and tested to estimate travel times on different routes in the network. The model has been carefully validated. It could be shown that the produced results are in good accordance with the results of the microsimulator AIMSUN, and that the CTM can be used to deliver reliable, even though not perfect estimates of delay and travel times in an urban network.

Since the ATCS prototype needs an estimate of the upcoming traffic demand of the next optimization interval in order to be able to perform an optimization of signal timings, the first two modules of the ATCS prototype cover traffic demand estimation. The first module performs a forecasting of detector counts. An approach by FÖRSTER (2008) has been implemented and tested for this purpose. The approach relies on space-time-patterns of detector counts of the four previous time intervals. The current pattern, which represents the traffic demand that has been observed during the last hour, is searched for in several reference patterns (in this thesis, only one such reference pattern has been used). The best matching sub-pattern within the reference pattern is identified and the values of the following time interval in the reference pattern can then be used to calculate estimates of the forecasted detector counts. In this thesis it could be shown that from time to time systematic over- or underestimation of forecasted counts occurs at some time intervals. The reason for these systematic errors has been investigated and the method has been slightly adapted. The effect could be reduced, but not eliminated entirely, i.e. some imprecision of the forecasting module remained.

The second module for demand estimation uses the forecasted detector counts as constraints for the subsequent estimation of OD flows, route and link volumes. The fundamental research by WANG (2008) has been adapted for this purpose. The module is based on the IM model by VAN ZUYLEN/WILLUMSEN (1980) for OD matrix estimation. The model uses traffic counts on several links in a network as constraints for the matrix estimation. The final estimated matrix has to reproduce the counts when assigned to the network. The more detailed the counts are, i.e. the more turning flows are counted directly, the better is the estimation. Therefore, an algorithm has been developed in this thesis which tries to reproduce a maximum of counts on these turnings based on a given set of counts. Since the IM model is sensitive to redundant counts which do not add new information to the estimation process, WANG (2008) developed several rules to eliminate such redundancies before the estimation process is started. In this thesis these rules have been transformed into a general algorithm that is applicable to any urban network. Furthermore, an additional algorithm by VAN ZUYLEN/BRANSTON (1982) has been implemented in order to eliminate inconsistent counts before starting the estimation. WANG (2008) tested different traffic assignment techniques which have been incorporated into an iterative procedure of repeated traffic assignment and matrix estimation. Two of these techniques have been tested in this thesis, too. For the estimation of travel times on different routes, which are needed during the process of traffic assignment, the CTM has been used in this thesis in addition to the traditional BPR functions. Furthermore, the approach by WANG (2008) has been adapted in such a way that not only OD flows, but also traffic volumes on different routes and all links of the network could be estimated.

An evaluation of the modules for traffic demand estimation has been conducted based on artificial data derived from a microsimulation study with AIMSUN. It revealed that link volumes can be estimated rather well, whereas the estimation of route volumes suffers from some imprecision. However, the latter is still good enough to identify those routes which are among the currently most heavily loaded routes.

The two next modules of the ATCS which have been developed in this thesis perform an adjustment of cycle length and phase durations and an optimization of offsets. The first is done by simply implementing the classic formulas for the calculation of fixed time signal plans as described in the HBS (FGSV, 2001) or RiLSA (FGSV, 2010). The module has been designed to calculate a common cycle length either based on the approach by WEBSTER (1958) or based on the desired degree of saturation. The common cycle length is dictated by the heaviest loaded intersection in the network.

The main focus of the adjustment of signal settings is on the model-based offset optimization. The CTM is used to evaluate the effects of different offset combinations in terms of total delay. For each offset combination to be tested, a single run of the CTM has to be performed. The two optimization approaches developed by ALMASRI (2006) have been implanted. The PGA optimizes the offsets of all intersections simultaneously, whereas the SGA optimizes offsets consecutively in groups of intersections in the order of the heaviest loaded routes. While ALMASRI (2006) used static inflow volumes and turning percentages to derive the volumes on different routes, the results of the demand estimation modules developed in this thesis have been used directly. SE has been developed as a third optimization algorithm in this thesis. It is a deterministic optimization technique that optimizes offsets of all intersections separately one after another by complete enumeration of all possible offsets at each intersection. The order of intersections is based on estimated route volumes, too. After offsets of all intersections have been optimized, the process is repeated until no further improvement can be achieved. A first comparison of all three optimization techniques revealed that they perform equally well.

The last major object of research of this thesis was signal plan transition. At the beginning of each time interval, the new signal timings have to be implemented at each intersection. This requires application of an appropriate technique of signal plan transition that does not induce major disturbances of traffic flow. The state-of-the-art of signal plan transition has been discussed, followed by a microsimulation study that simulated a simple meshed network with three intersections. It could be shown that the rather smooth Shortway transition technique, which shortens or lengthens up to three cycles in order to implement the new signal plan, performs best and does not produce major disturbances. This smooth transition technique has been implemented in the framework of the ATCS in order to handle signal plan transition at each intersection at the beginning of each time interval. It has also been incorporated into the model-based offset optimization in order to consider the effects of signal plan transition directly during the optimization process.

Finally, a comprehensive microsimulation study has been performed to evaluate the performance of the ATCS prototype. The prototype has been applied to two networks in the city of Hanover, Germany. The results reveal that the prototype of the newly developed ATCS has some potential to improve travel times in a sub-network compared to an optimized fixed time signal control. However, the degree of this improvement depends on the network. The overall improvement of travel times is clearly visible in the first network. Both cycle length adjustment

and offset optimization contribute to this improvement and should be used as complementary modules. When offset optimization is switched off, the overall travel time is much higher which shows that the concept of combining GA with the CTM works well to optimize offsets. But it could also be shown that the three optimization techniques PGA, SGA, and SE perform equally well. This is rather surprising, given the simplicity of the last deterministic approach. This finding suggests that good offsets can obviously be produced without the need for heuristic optimization techniques, at least for small networks as used in this thesis.

Major improvements of travel times could be achieved more easily on the less heavily loaded routes. Since the optimization tries to find a system optimum, the ATCS prototype in its present form cannot be expected to produce perfect “green waves” along the main arterials. This was one of the main criticisms on GA expressed by MÜCK (2008a, 2008b). His observations have therefore been confirmed for this point. However, the described negative effects of signal plan transition did not arise in this study due to the implemented smooth transition technique.

The results of the second network are less promising. The structure of the network, the traffic demand pattern and the already very good optimized reference case with fixed time signal plans did not leave much room for further improvements. This finding suggests that the benefit of the application of an ATCS highly depends on the given situation in a specific network. Major improvements cannot always be expected and the installation of an ATCS might therefore not always be necessary or advisable.

The prototypical implementation of the ATCS proves that the CTM in combination with GA is real-time capable which had been doubted before (MÜCK, 2008a, 2008b). Optimizing signal settings for one time interval takes 5 minutes in the case study, whereas one time interval has a duration of 15 minutes, leaving an additional buffer of 10 minutes. The method can therefore also be used to optimize larger networks, but only to some extent.

The performance of the ATCS stands or falls on the quality of the demand estimation. If the estimated demand is of poor quality, even the best optimization algorithm will produce inferior results. The demand estimation used in this thesis obviously works well if average detector counts are used as input data. When single simulation runs are optimized, which corresponds much more to a real world application of the ATCS prototype, the variations of cycle length reveal that some partly major imprecisions of forecasted detector counts remain. Moreover, it has to be highlighted that the method can only forecast detector counts according to the available reference space-time-patterns. Therefore, reference patterns for various days and special situations must be used and updated regularly.

9.2 Outlook

It could be shown that the adjustment of cycle length contributes strongly to a reduction of travel times and thus delay. However, the reference cases used for comparison in this thesis consisted of only two signal plans optimized for morning and afternoon peak hour respectively. This must be regarded with caution. On the one hand, especially the cycle length of 90 seconds in the first network seems to be a bit generous, since 80 seconds as chosen by the ATCS prototype obviously do not provoke a breakdown during peak hour either. And on the other hand, major travel time improvements during off-peak hours had to be expected because the reference signal plans have not been optimized for this demand. In this regard, creating reference

cases containing at least one or even more additional signal plans for the off-peak hours should be considered for future tests. Even though the advantage of adapted cycle length has been shown, it cannot be concluded from this study whether this has to be done necessarily by use of an ATCS. Providing an according number of adapted fixed time signal plans which are activated based on the currently measured traffic demand might also be an appropriate alternative. It should therefore be investigated in the future, which degree of improvement can be already achieved by such a rather simple approach instead of a much more sophisticated ATCS.

The study also revealed that the performance of the ATCS prototype and the attainable improvements depend on the network where the ATCS is applied. It might be useful to investigate the specific constraints of a network which promote usage of an ATCS in order to be able to differentiate between reasonable and less reasonable applications of ATCS.

The performance of the simple SE algorithm was rather surprising. This observation raises the question whether heuristic optimization techniques such as GA are really necessary to identify good offsets. On the one hand, other optimization techniques should be considered as well. On the other hand, it might be useful to investigate the problem of offset optimization more generally and to analyze the solution space for different simpler networks analytically in order to answer the question why the simple SE approach performs so well. Such an investigation might help to develop simple but efficient offset optimization techniques in the future.

It could be observed that the ATCS prototype does not necessarily produce perfect “green waves” along the major arterials. The identified solution might be optimal for the whole system, but it does not correspond to what the drivers and the authorities in charge of signal control expect. Further investigations should therefore try to find a way to incorporate “green wave” implementation along the major arterials into the ATCS. Simple weighting of routes might not be appropriate since the ATCS prototype already implements such a concept to some extent by optimizing total delay in veh·s. This can be seen as a weighting strategy based on traffic volumes. This weighting can be problematic if two opposing directions of an arterial have equal traffic volumes. In this case no perfect “green wave” will be implemented for either direction, unless the distances between the intersections happen to enable “green waves” in both directions. Therefore, other techniques should be searched for which enable the implementation of perfect “green waves” in the framework of the ATCS prototype.

Forecasting of detector counts has been shown to be a difficult task. Even though the approach used in this thesis could be improved, it still produces forecasted counts of inferior quality at some intervals, which directly worsens the quality of the optimized signal timings. Therefore the question should be addressed whether the forecasting technique can be further improved or replaced by a better alternative.

Even though the ATCS prototype uses a traffic model that can explicitly model oversaturated traffic conditions, the demand estimation in its present form cannot cope with oversaturation. If detector counts from oversaturated links are used as constraints, they do not reflect the real demand but only the capacity of the link. This reduces the quality of the overall demand estimation for the whole network. Future research should therefore be conducted to find solutions to enable the ATCS prototype to also control oversaturated networks effectively.

References

Abu-Lebdeh, G., Benekohal, R.F. (1997): Development of Traffic Control and Queue Management Procedures for Oversaturated Arterials. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1603, Transportation Research Board of the National Academies, Washington, D.C., 1997, pp. 119-127

Almasri, E. (2006): A New Offset Optimization Method for Signalized Road Networks. Doctoral thesis at the Institute of Transport, Road Engineering and Planning, Leibniz Universität Hannover, Hannover, Germany

Andrews, C.M., Elahi, S.M., Clark, J.E. (1997): Evaluation of New Jersey Route 18 OPAC/MIST Traffic-Control System. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1603, Transportation Research Board of the National Academies, Washington, D.C., 1997, pp. 150-155

Barnhart, C., Laporte, G. (Eds.) (2007): Handbook on OR & MS, Vol. 14, Chapter 11, Elsevier B.V.

Bell, M.G.H., Shield C.M., Busch, F., Kruse, G. (1997): A stochastic user equilibrium path flow estimator. In: Transportation Research Part C, Vol. 5 (3-4), pp. 197-210

Bielefeldt, C., Busch, K. (1994): MOTION – A New On-line Traffic Signal Network Control System. Proceedings of the Seventh International Conference on Road Traffic Monitoring and Control, London, UK, April 26-28

Bielefeldt, C., Condie, H., Kosmatopoulos, E., Richards, A., McDonald, M., Mück, J., Hanitzsch, A. (2004): SMART NETS Final Report, Deliverable 26 of the Project IST-2000-28090 Signal Management in Real Time for Urban Traffic Networks

Böttger, R. (1972): Optimale Koordinierung von Signalanlagen in einem Straßennetz (Planungs- und Steuerprogramm VERO). In: Straßenverkehrstechnik 2/1972, pp. 52-58

Bowen, G.T., Bretherton, R.D. (1996): Latest Developments in SCOOT – Version 3.1. Proceedings of the 8th International Conference on Road Traffic Monitoring and Control. London, UK, April 23-25

Braun, R. (2008): Ein echtzeitfähiger Evolutionärer Algorithmus zur netzweiten Optimierung der Lichtsignalsteuerung. Doctoral thesis at the Chair of Traffic Engineering and Control, Technische Universität München, München, Germany.

Braun, R., Weichenmeier, F. (2005a): Automatische Offline-Optimierung der lichtsignaltechnischen Koordinierung des mIV im städtischen Netz unter Verwendung genetischer Algorithmen. Proceedings of Heureka ’05, Karlsruhe, Germany, March 2-3

Braun, R., Weichenmeier, F. (2005b): Automatic Offline-Optimization of Coordinated Traffic Signal Control in Urban Networks using Genetic Algorithms. Proceedings of the 12th World Congress on ITS, San Francisco, USA, November 6-10

Braun, R., Kemper, C., Weichenmeier, F. (2008a): TRAVOLUTION – Adaptive Urban Traffic Signal Control with an Evolutionary Algorithm. Proceedings of the 4th International Symposium Networks for Mobility. Stuttgart, Germany, September 25-26

Braun, R., Kemper, C., Weichenmeier, F., Wegmann, J. (2008b): Improving Urban Traffic in Ingolstadt. Proceedings of the 7th European Congress on ITS, Geneva, Switzerland

Braun, R., Busch, F., Kemper, C., Hildebrandt, R., Weichenmeier, F., Menig, C., Paulus, I., Preßlein-Lehle, R. (2009): TRAVOLUTION – Netzweite Optimierung der Lichtsignalsteuerung und LSA-Fahrzeug-Kommunikation. In: Straßenverkehrstechnik 6/2009, pp. 365-374

Bretherton, R.D., Wood, K., Bowen, G.T. (1998): SCOOT Version 4. In: Traffic Engineering and Control, 39 (7-8) 1998, pp. 425- 427

Brillouin, L. (1956): Science and Information Theory. Academic Press. New York

Brilon, W., Wietholt, T., Pott, A., Zelke, U. (2009): Adaptive koordinierte Signalsteuerung in Münster. In: Straßenverkehrstechnik 09/2009, pp. 565-573

Busch, F., Kruse, G. (1993): MOTION – Ein neues Verfahren für die städtische Lichtsignalsteuerung und seine Erprobung im Rahmen des EG-Programms ATT. Proceedings of Heureka ’93, Karlsruhe, Germany

Busch, F., Kruse, G. (no date): MOTION – Ein neues Verfahren für die städtische Lichtsignalsteuerung. Special reprint, SIEMENS AG, Munich, Germany

Cascetta, A., Nuzzolo F.R., Vitetta, A. (1996): A modified logit route choice model overcoming path-overlapping problems. Proceedings of the 13th International Symposium on Transportation and Traffic Theory. Lyon, France

Ceylan, H., Bell, M.G.H. (2004): Traffic Signal Timing Optimisation based on Genetic Algorithm Approach, including Drivers’ Routing. In: Transportation Research B 38, pp. 329-342

Ceylan, H., Bell, M.G.H. (2005): Genetic Algorithm Solution for the Stochastic Equilibrium Transportation Networks under Congestion. In: Transportation Research B 39, pp. 169-185

Chow, A.H.F., Gomes, G., Kurzhanskiy, A.A., Varaiya, P. (2010): AURORA RNM – A Macroscopic Simulation Tool for Arterial Traffic Modeling and Control. Proceedings of the 89th TRB Annual Meeting, Washington, D.C., USA, January 10-14

City of Hamburg, Ingenieurbüro Vössing GmbH (2005): Zwischenstandsbericht zur Vorher-Nachher-Untersuchung, Hamburg (unpublished)

Cohen, D., Head, L., Shelby, S.G. (2007): Performance Analysis of Coordinated Traffic Signals During Transition. In: Transportation Research Record: Journal of the Transportation Research Board, No. 2035, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 19-31

Daganzo, C. (1994): The Cell Transmission Model: A dynamic Representation of Highway Traffic consistent with the Hydrodynamic Theory. In: Transportation Research B 28 (4), pp. 269-287.

Daganzo, C. (1995): The Cell Transmission Model, Part II: Network Traffic. In: Transportation Research B 29 (2), pp. 79-93.

Diakaki, C., Papageorgiou, M., Aboudolas, K. (2002): A Multivariable Regulator Approach to Traffic-responsive Network-wide Signal Control. In: Control Engineering Practice 10, pp. 183-195

Diakaki, C., Dinopoulou, V., Aboudolas, K., Papageorgiou, M., Ben-Shabat, E., Seider, E., Leibov, A. (2003): Extensions and New Applications of the Traffic-Responsive Urban Control Strategy. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1856, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 202-211

Dinopoulou, V., Diakaki, C., Papageorgiou, M. (2006): Applications of the urban traffic control strategy TUC. In: European Journal of Operational Research 175, pp.1652-1665

Donati, F., Mauro, V., Roncolini, G., Vallauri. I. (1984): A Hierarchical-Decentralized Traffic Light Control System – The First Realization: “Progetto Torino”. Proceedings of the 9th IFAC World Congress, Budapest, Hungary

Feldman, O., Maher, M. (2002a): Optimisation of Traffic Signals Using a Cell Transmission Model. Proceedings of the 9th Meeting of the Euro Working Group on Transportation, Bari, Italy, June 10-12

Feldman, O., Maher, M. (2002b): The Application of the Cell Transmission Model to the Optimisation of Signals on Signalised Roundabouts. Proceedings of the European Transport Conference (ETC), Cambridge, UK, September 9-11

FGSV, Forschungsgesellschaft für Straßen- und Verkehrswesen (2001): Handbuch für die Bemessung von Straßenverkehrsanlagen – HBS (Manual for the dimensioning of road infrastructure). FGSV-Verlag, Cologne, Germany

FGSV, Forschungsgesellschaft für Straßen- und Verkehrswesen (2010): Richtlinien für Lichtsignalanlagen – RiLSA (Guidelines for Traffic Signals), FGSV-Verlag, Cologne, Germany

Flötteröd, G., Nagel, K. (2005): Some practical extensions to the Cell Transmission Model. Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems. Vienna, Austria, September 13-16

Förster, G. (2008): Kurzfristprognose auf Basis von Raum-Zeit-Mustern. Proceedings of Heureka ’08. Stuttgart, Germany, 5 & 6 March

Foy, M.D., Benekohal, R.F., Goldberg, D.E. (1992): Signal Timing Determination Using Genetic Algorithms. In: Transportation Research Record: Journal of the Transportation Research Board (1365), Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 108-115

Friedrich, B. (1997): Ein verkehrsadaptives Verfahren zur Steuerung von Lichtsignalanlagen. Doctoral thesis at the Chair of Traffic Engineering and Planning, Technische Universität München, München, Germany

Friedrich, B. (2000a): Steuerung von Lichtsignalanlagen: BALANCE – ein neuer Ansatz. In: Straßenverkehrstechnik 7/2000, pp. 321-328

Friedrich, B. (2000b): Models for Adaptive Urban Traffic Control. In: Proceedings of the 8th Meeting of the Euro Working Group on Transportation, Rome, Italy, September 11-14

Friedrich, B., Almasri, E. (2005): Modellbasierte Optimierung der Versatzzeiten mit dem Cell Transmission Model. Proceedings of Heureka ‘05, Karlsruhe, Germany

Friedrich, B., Wang Y. (2008): Optimierung der Matrixschätzung durch Elimination redundanter Informationen. Proceedings of Heureka ‘08, Stuttgart, Germany

Friedrich, B., Wang Y. (2006): Optimizing O-D Estimation with Respect to Redundant Information and Route Choice. Proceedings of the 11th IFAC Symposium on Control in Transportations Systems, Delft, Netherlands

Ganser, M (2003): Neue Verfahren - neue Praxis. In: Steierwald, M., Martens, S. (Eds.): Steuerung kommunaler Verkehrsnetze - Was können und was kosten die neuen Verfahren? Was setzen Sie voraus? Proceedings Akadamie für Technikfolgenabschätzung in Baden-Württemberg

Garben, M., Heck, H.M., Hotop, R., Keller, H., Meißner, J.D., Sahling, B.M., Stottmeister, V. (1988): SIGMA: Ein Optimierungsverfahren zur koordinierten Lichtsignalsteuerung. In: Straßenverkehrstechnik 4/1988, pp. 135-140

Gartner, N.H. (1982): Demand-responsive Decentralized Urban Traffic Control – Part I: Single-Intersection Policies. Report No. DOT/RSPA/DPB-50/81/24 for the U.S. Department of Transportation, Research and Special Programs Administration, Washington, D.C., USA

Gartner, N.H. (1983): OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In: Transportation Research Record: Journal of the Transportation Research Board, No. 906, Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 75-81

Gartner, N.H., Kaltenbach, M.H., Miyamoto, M.M. (1983): Demand-responsive Decentralized Urban Traffic Control – Part II: Network Extensions. Report No. DOT/OST/P-34/85/009 for the U.S. Department of Transportation, Research and Special Programs Administration, Washington, D.C., USA

Gartner, N.H., Farhad, J.P., Andrews, C.M. (2001): Implementation of the OPAC Adaptive Control Strategy in a Traffic Signal Network. Proceedings of the IEEE Intelligent Transportation Systems Conference, Oakland, CA, USA, August 25-29

GEVAS (2005): TRENDS Version 4.2/5.0 Benutzerhandbuch. GEVAS software Systementwicklung und Verkehrsinformatik GmbH, München, Germany

GEVAS (2010): BALANCE Produktinformation, http://www.gevas.eu/index.php?id=149, date of query: 2010-04-21

Goldman, S. (1953): Information Theory. Prentice Hall. New York

Greenough, J.C., Kelman, W.L. (1998): Metro Toronto Scoot: Traffic Adaptive Control Operation. In: ITE Journal, vol. 68 (5)

Hadi, M.A., Wallace, C.E. (1993): Hybrid Genetic Algorithm to Optimize Signal Phasing and Timing. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1421, Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 104-112

Hale, D.K. (2005): Traffic Network Study Tool – TRANSYT 7F, United States Version (Manual). McTrans Center, University of Florida, Gainesville, Florida, USA

Hellinga, B. R. (1994): Estimating Dynamic Origin-Destination Demands from Link and Probe Counts. PhD thesis, Queen's Universitsy, Kingston, Ontario, Canada

Herz, R., Schlichter, H.G., Siegener, W. (1976): Angewandte Statistik für Verkehrs- und Regionalplaner. Werner-Verlag, Düsseldorf, Germany

Hounsell, N.B., McDonald, M. (2001): Urban network traffic control. In: Proceedings of the Institution of Mechanical Engineers Part I: Journal of Systems and Control Engineering, 215, (4), pp. 325-334

Hunt, P.B., Robertson, D.I., Bretherton, R.D., Winton, R.I. (1981): SCOOT – A Traffic Responsive Method of Coordinating Signals. TRRL Laboratory Report 1014, Transport and Road Research Laboratory, Crowthorne, Berkshire, UK

Keller, H., Ploss, G. (1987): Real Time Identification of O-D-Network Flows from Counts for Urban Traffic Control. Proceedings of the 10th International Symposium on Transportation and Traffic Theory, Elsevier, New York, Amsterdam, London, pp. 267-289

Kimber, R.M., Hollis, E.M. (1979): Traffic Queues and Delays at Road Junctions. TRRL Laboratory Report 909, Transport and Road Research Laboratory, Crowthorne, Berkshire, UK

Koch, O. (2006): Innovative Verkehrssteuerung im Rahmen eines Pilotprojektes im Stadteil Barmbek. VSVI Information

Kosmatopoulos, E., Papageorgiou, M., Bielefeldt, C., Dinopoulou, V., Morris, R., Mück, J., Richards, A., Weichenmeier, F. (2006): International comparative field evaluation of a traffic-responsive signal control strategy in three cities. In: Transportation Research Part A 40, pp. 399-413

Kruse, G., Busch, F. (2002): MOTION for SITRAFFIC – Optimierung der Lichtsignalsteuerung im Einsatz. Proceedings of Heureka '02, Karlsruhe, Germany

Lämmer, S. (2007): Reglerentwurf zur dezentralen Online-Steuerung von Lichtsignalanlagen in Straßennetzwerken. Doctoral thesis at the Faculty of Transportation and Traffic Sciences "Friedrich List", Technische Universität Dresden. Dresden, Germany.

Lämmer, S., Krimmling, J., Hoppe, A. (2009): Selbst-Steuerung von Lichtsignalanlagen in Straßennetzwerken – Regelungstechnischer Ansatz und Simulation. In: Straßenverkehrstechnik 11/2009, pp. 714-721

Lee, C., Machemehl, R.B. (1998): Genetic Algorithm, Local and Iterative Searches for Combining Traffic Assignment and Signal Control. Proceedings of the International Conference on Traffic and Transportation Studies, Beijing, China, July 27-29, pp. 488-497

Lee, J., Williams, B.M. (2009): Development and Testing of a Constrained Optimization Model for Traffic Signal Plan Transition. Proceedings of the 88th Annual Meeting of the Transportation Research Board, Transportation Research Board of the National Academies, Washington, D.C., 2009

Lehnhoff, N. (2005): Überprüfung und Verbesserung der Qualität von automatisch erhobenen Daten an Lichtsignalanlagen. Doctoral thesis at the Institute of Transport, Road Engineering and Planning, Universität Hannover, Hannover, Germany.

Li, Z., Chang, G.L. (2010): Modelling Arterial Signal Optimization with Enhanced Cell Transmission Formulations. Proceedings of the 89th Annual Meeting of the Transportation Research Board, Transportation Research Board of the National Academies, Washington, D.C, January 10-14

Lighthill, M.J., Whitham, J.B. (1955): On Kinematic Waves. I. Flow Movement in Long Rivers. II. A Theory of Traffic Flow on Long Crowded Roads. Proceedings of Royal Society (A229), pp. 281-345.

Lin, W.H., Wang, C. (2004): An Enhanced 0-1 Mixed-Integer LP Formulation for Traffic Signal Control. In: IEEE Transactions on Intelligent Transportation Systems Vol. 5 (4), pp. 238-245

Lo, H.K. (1999): A novel traffic signal control formulation. In: Transportation Research A 33, pp. 433-448

Lo, H.K. (2001): A Cell-Based Traffic Control Formulation: Strategies and Benefits of Dynamic Timing Plans. In: Transportation Science 35 (2), pp. 148-164

Lo, H.K., Chang, E., Chan Y.C. (2001): Dynamic network traffic control. In: Transportation Research A 35, pp. 721-744

Lo. H.K., Chow, A.H.F. (2004): Control Strategies for Oversaturated Traffic. In: Journal of Transportation Engineering 130 (4), pp. 466-478

Lowrie, P.R. (1982): The Sydney Co-ordinated Adaptive Traffic System – Principles, Methodology, Algorithms. In: Proceedings of the International Conference on Road Traffic Signalling, London, UK

Mauro, V., Di Taranto, C. D. (1990): UTOPIA. Proceedings of the 6th IFAC Symposium on Control, Computers, and Communication in Transportation, Paris, France

Mazzamatti, M., Netto, D.V.V.F., Vilanova, L., Ming, S.H. (1998): Benefits gained by responsive and traffic adaptive systems in Sao Paulo. Proceedings of the 9th International Conference on Road Transportation Information and Control, pp. 109-113

McDonald, M., Hounsell, N.B. (1991): Road traffic control: TRANSYT and SCOOT. In: Papageorgiou, M. (Ed.), Concise Encyclopedia of Traffic and Transportation Systems, Pergamon, pp. 400-408

Meffert, K. et al. (2009): JGAP - Java Genetic Algorithms and Genetic Programming Package. URL: http://jgap.sf.net.

Mertz, B. (2001): Ein mikroskopisches Verfahren zur verkehrsadaptiven Knotenpunktsteuerung mit Vorrang des öffentlichen Verkehrs. Doctoral thesis at the Chair of Traffic Engineering and Planning, Technische Universität München, München, Germany.

MIZAR Automazione (1993): General Information on UTOPIA Concepts and SPOT Control Unit. Turin, Italy

Mück, J. (2008a): Neue Schätz- und Optimierungsverfahren für Adaptive Netzsteuerungen. In: Straßenverkehrstechnik 12/2008, pp. 761-773

Mück, J. (2008b): Schätz- und Optimierungsverfahren in der Adaptiven Netzsteuerung SITRAFFIC Motion MX. Proceedings of Heureka '08, Stuttgart, Germany

Mussa, R., Selekwa, M. (2003): Proposed Methodology of Optimizing Transitioning Between Time-of-Day Timing Plans. In: Journal of Transportation Engineering, No. 129 (4), American Society of Civil Engineers, Reston, VA, pp. 392-397

Papageorgiou, M., Diakaki, C., Dinopoulou, V, Kotsialos, A., Wang, Y. (2003): Review of Road Traffic Control Strategies. In: Proceedings of the IEEE, 91 (12), pp. 2043-2067

Papageorgiou, M., Ben-Akiva, M., Bottom, J., Bovy, P.H.L., Hoogendoorn, S.P., Hounsell, N.B., Kotsialos, A., McDonald, M. (2007): ITS and Traffic Management. In: Barnhart, C., Laporte, G. (Eds.): Handbook in OR & MS, Vol. 14, Elsevier

Park, B., Messer C. J., Urbanik, T. (1999): Traffic Signal Optimization Program for Oversaturated Conditions – Genetic Algorithm Approach. In: Transportation Research Record: Journal of the Transportation Research Board (1683), Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 133-142

Park, B., Messer C. J., Urbanik, T. (2000): Enhanced Genetic Algorithm for Signal-Timing Optimization of Oversaturated Intersections. In: Transportation Research Record: Journal of the Transportation Research Board (1727), Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 32-41

Peck, C., Gorton, P.T.W., Liren, D. (1990): The application of SCOOT in developing countries. Proceedings of the 3rd International Conference on Road Traffic Control, pp. 104-109

PEEK (2010): UTOPIA Network Management and Control, http://www.peektraffic.nl/page/484, date of query: 2010-04-22

Peters, J.M., McCoy, J., Bertini, R. (2007): Evaluating An Adaptive Signal Control System in Gresham. (www.its.pdx.edu/upload_docs/1248894238mPTJ0AxnYW.pdf, date of query 2010-05-03)

Pohlheim, H. (2000): Evolutionäre Algorithmen – Verfahren, Operatoren und Hinweise für die Praxis. Springer, Berlin

Richards, P.I. (1956): Shockwaves on the Highway. In: Operations Research B (22), pp. 81-101.

Robertson, D.I. (1969): TRANSYT: A Traffic Network Study Tool. RRL Laboratory Report 253, Road Research Laboratory, Crowthorne, Berkshire, UK

Rohde, J. (2006): Schätzung von Herkunfts-/Zielbeziehungen. OptiV - Erschließung von Entscheidungs- und Optimierungsmethoden für die Anwendung im Verkehr. Hannover, 2006 (www.optiv.de, date of query 2009-06-16)

Rohde, J., Friedrich, B., Schüler, T. (2008): Emissions- und Kraftstoffreduzierung im Stadtverkehr durch Versatzzeitoptimierung. Proceedings of Heureka'08, Stuttgart, Germany, March 5-6.

SCATS (2010): SCATS website, www.scats.com.au, date of query: 2010-04-20

Schnabel, W., Lohse, D. (1997): Grundlagen der Straßenverkehrstechnik und der Verkehrsplanung. Verlag für Bauwesen, Berlin

SCOOT (2010): SCOOT website, www.scoot-utc.com, date of query: 2010-04-20

Shannon, C. E. (1948): A Mathematical Theory of Communication. In: The Bell System Technical Journal, 27

Sheffi, Y. (1985): Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods. Prentice-Hall, Inc., Englewood Cliffs, N.J.

Shelby, S.G., Bullock, D.M., Gettman, D. (2006): Transition Methods in Traffic Signal Control. In: Transportation Research Record: Journal of the Transportation Research Board (1778), Transportation Research Board of the National Academies, Washington, D.C., USA, pp. 130-140

Siegloch, W. (1973): Die Leistungsermittlung an Knotenpunkten ohne Lichtsignalanlage. Scientific Series „Straßenbau und Straßenverkehrstechnik", Vol. 154, Federal Ministry of Transportation, Bonn, Germany

SIEMENS (2010): SITRAFFIC MOTION MX, http://www.mobility.siemens.com/shared/data/pdf/www/infrastructure_logistics/sitraffic_motion_mx-der_verkehr_flie_dft.pdf, date of query: 2010-04-21

Sims, A.G., Dobinson, K.W. (1980): The Sydney Coordinated Adaptive Traffic (SCAT) System – Philosophy and Benefits. In: IEEE Transactions on Vehicular Technology, Vol. VT-29 (2), pp. 130-137

Stevanovic, A., Martin, P.T., Stevanovic, J. (2007): VisSim-Based Genetic Algorithm Optimization of Signal Timings. In: Transportation Research Record: Journal of the Transportation Research Board, No. 2035, Transportation Research Board of the National Academies, Washington, D.C., pp. 59-68

Stevanovic, A., Kergaye, C., Martin, P. T. (2008): Field Evaluation of SCATS Traffic Control in Park City, UT. Proceedings of the 15th World Congress on Intelligent Transport Systems, New York, USA, November 16-20

Taale, H., Fransen, W.C.M., Dibbits, J. (1998): The second assessment of the SCOOT system in Nijmegen. Proceedings of the 9th International Conference on Road Transportation Information and Control, pp. 109-113

Teklu, F., Sumalee, A., Watling, D. (2007): A Genetic Algorithm Approach for Optimizing Traffic Control Signals Considering Routing. In: Computer-Aided Civil and Infrastructure Engineering, Vol. 22, pp. 31-43

TRB, Transportation Research Board (2000): Highway Capacity Manual. National Research Council, Washington, D.C., USA

TSS, Transport Simulation Systems (2006): AIMSUN NG User's Manual, Barcelona, Spain

TSS, Transport Simulation Systems (2007): AIMSUN API Manual, Barcelona, Spain

van Katwijk, R.T. (2008): Multi-Agent Look-Ahead Traffic-Adaptive Control. Doctoral thesis at the Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

van Vliet, D. (1982): SATURN – A Modern Assignment Model. In: Traffic Engineering and Control, Vol. 23, pp. 578-581

van Zuylen, H.J. (1979): The Information Minimising Method: its Validity and Applicability to Transportation Planning. In: New Developments in Modelling Travel Demand in Urban Systems, Saxon House. Westmead, Farnborough, Hants, UK, pp. 344-371

van Zuylen, H.J. (1981): Some Improvement in the Estimation of an OD Matrix from Traffic Counts. Proceedings of the 8th International Symposium on Transportation and Traffic Theory. University of Toronto Press, Toronto, Canada

van Zuylen, H.J., Branston, D.M. (1982): Consistent Link Flow Estimation From Counts. In: Transportation Research B 16 (6), pp. 473-476

van Zuylen, H.J., Willumsen, L.G. (1980): The Most Likely Trip Matrix Estimated from Traffic Counts. In: Transportation Research B 14, pp. 281-293

Walsh, J.A., Webber, M.J. (1977): Information Theory: some Concepts and Measures. In: Environment and Planning A, 9, pp. 395-417

Wang, Y. (2008). Optimierung der Quelle-Ziel-Matrixschätzung hinsichtlich Redundanzstörung sich verändernder Verkehrszustände. Doctoral thesis at the Institute of Transport, Road Engineering and Planning, Leibniz Universität Hannover, Hannover, Germany

Wardrop, J.G. (1952): Some Theoretical Aspects of Road Traffic Research. Proceedings of the Institute for Transportation Studies, University of Leeds, City of Leeds, UK

Webster, F.V. (1958): Traffic Signal Settings. Road Research Technical Paper No. 39. Department of Scientific and Industrial Research, Road Research Laboratory, London, UK

Wietholt, T. (2009): Einsatzbereiche Grüner Wellen und Verkehrsabhängiger Steuerungen. Series Lehrstuhl für Verkehrswesen, Ruhr-Universität Bochum, Vol. 33

Wood, K. (1993): Urban Traffic Control, System Review. Project Report 41. Traffic and Transport Resource Centre, Transport Research Laboratory, Crowthorne, Berkshire, UK

Yun, I., Park, B. (2006): Stochastic Optimization Method for Coordinated Actuated Signal Systems. Research Report No. UVACTS-15-0-102, Center for Transportation Studies, University of Virginia, Charlottesville, USA

Yun, I., Best, M., Park, B. (2008): Evaluation of Traffic Controller Performance during Time-of-Day Transition at Coordinated Actuated Signal Systems. In: Transportation Research Record: Journal of the Transportation Research Board, No. 2080, Transportation Research Board of the National Academies, Washington, D.C., pp. 92-99

Zhang, L., Yin, Y., Lou, Y. (2010): Robust Signal Timing for Arterials under Day-to-Day Demand Variations. Proceedings of the 89th TRB Annual Meeting, Washington, D.C., USA, January 10-14

List of figures

List of tables

List of abbreviations

API	Application Programming Interface
ATCS	Adaptive Traffic Control System
BALANCE	Balancing Adaptive Network Control Method
BPR	U.S. Bureau of Public Roads
CF	Communality Factor
CTM	Cell Transmission Model
DS	Degree of saturation
FGSV	Forschungsgesellschaft für Straßen- und Verkehrswesen
GA	Genetic Algorithm
GHz	Gigahertz
GPS	Global Positioning System
GUI	Graphical User Interface
HBS	Handbuch für die Bemessung von Straßenverkehrsanlagen
HCM	Highway Capacity Manual
IM	Information Minimization
JGAP	Java Genetic Algorithms and Genetic Programming Package
LWR	Lighthill-Whitham-Richards
OD	Origin-Destination
OPAC	Optimization Policies for Adaptive Control
MAPE	Mean Absolute Percentage Error
MNL	Multinomial Logit
MSA	Method of Successive Averages
MSE	Mean Squared Error

MOTION	Method for the Optimization of Traffic Signals in Online Controlled Networks
PGA	Parallel Genetic Algorithm
PI	Performance Index
RiLSA	Richtlinien für Lichtsignalanlagen
RMSE	Root Mean Square Error
RRMSE	Relative Root Mean Square Error
SCATS	Sydney Coordinated Adaptive Traffic System
SCOOT	Split, Cycle and Offset Optimization Technique
SE	Sequential Enumeration
SGA	Serial Genetic Algorithm
SUE	Stochastic User Equilibrium
T7F	TRANSYT 7F
TOD	Time-of-Day
TRANSYT	Traffic Network Study Tool
TSS	Transport Simulation Systems
TUC	Traffic-responsive Urban Control
UTC	Urban Traffic Control
UTOPIA	Urban Traffic Optimization by Integrated Automation
veh/h	vehicles per hour
veh·s	vehicles seconds
VFC-OPAC	Virtual-Fixed-Cycle OPAC
XML	Extensible Markup Language

Appendices

A Validation of the CTM by comparison with AIMSUN

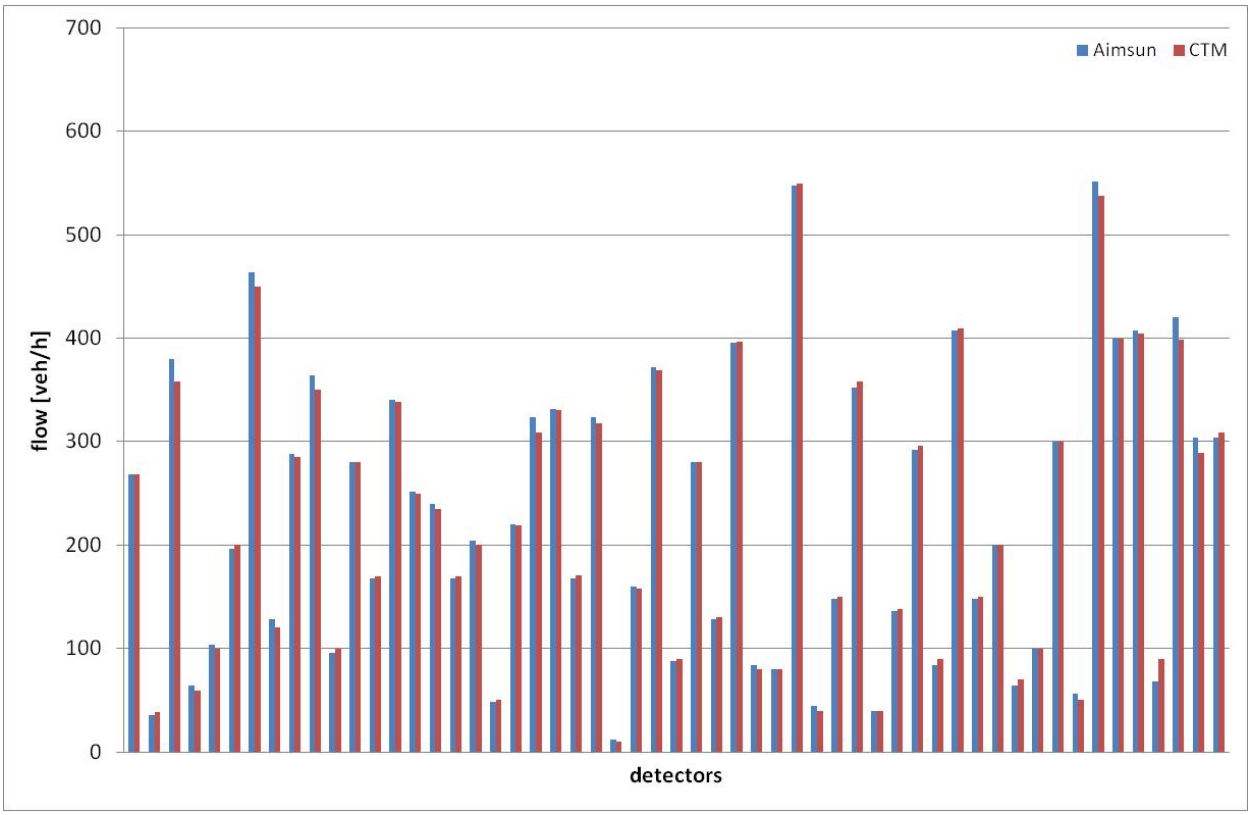


Figure A-1: Flow, morning peak (100 percent)

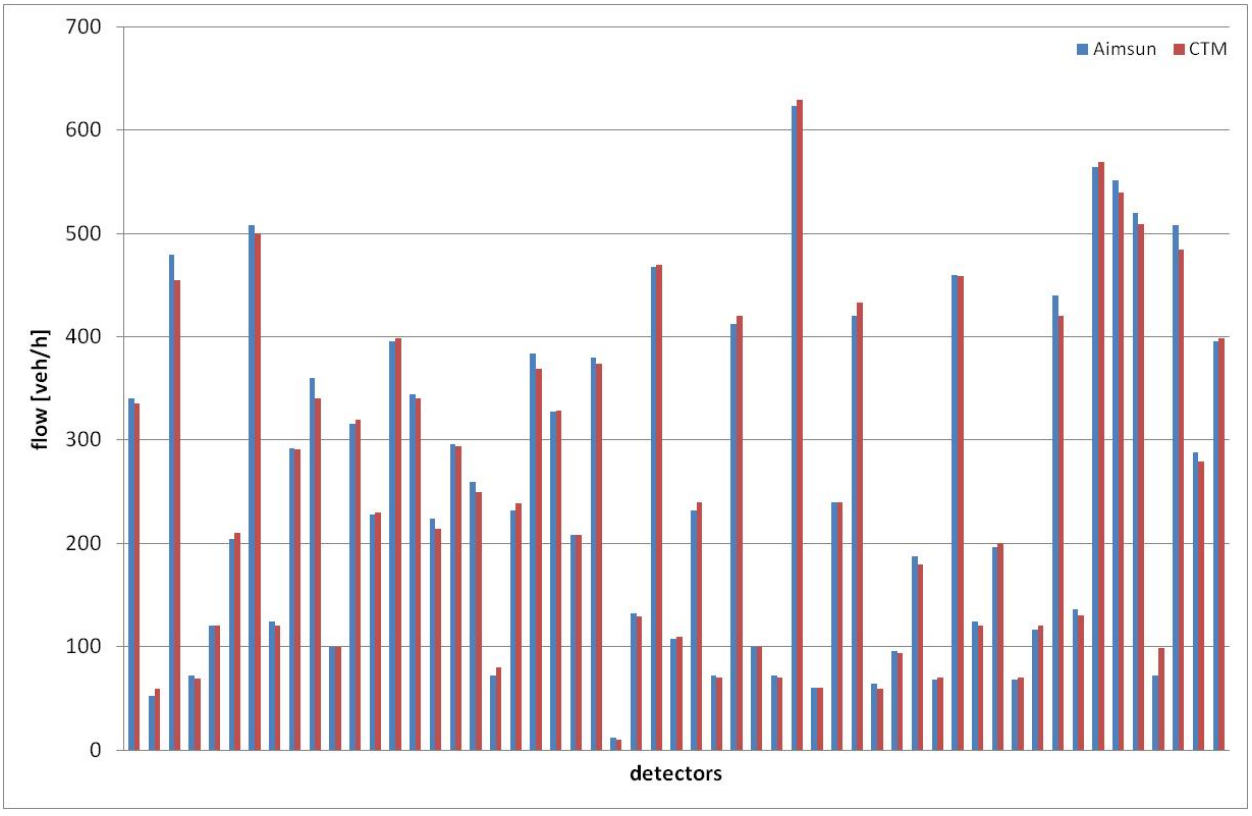


Figure A-2: Flow, afternoon peak (100 percent)

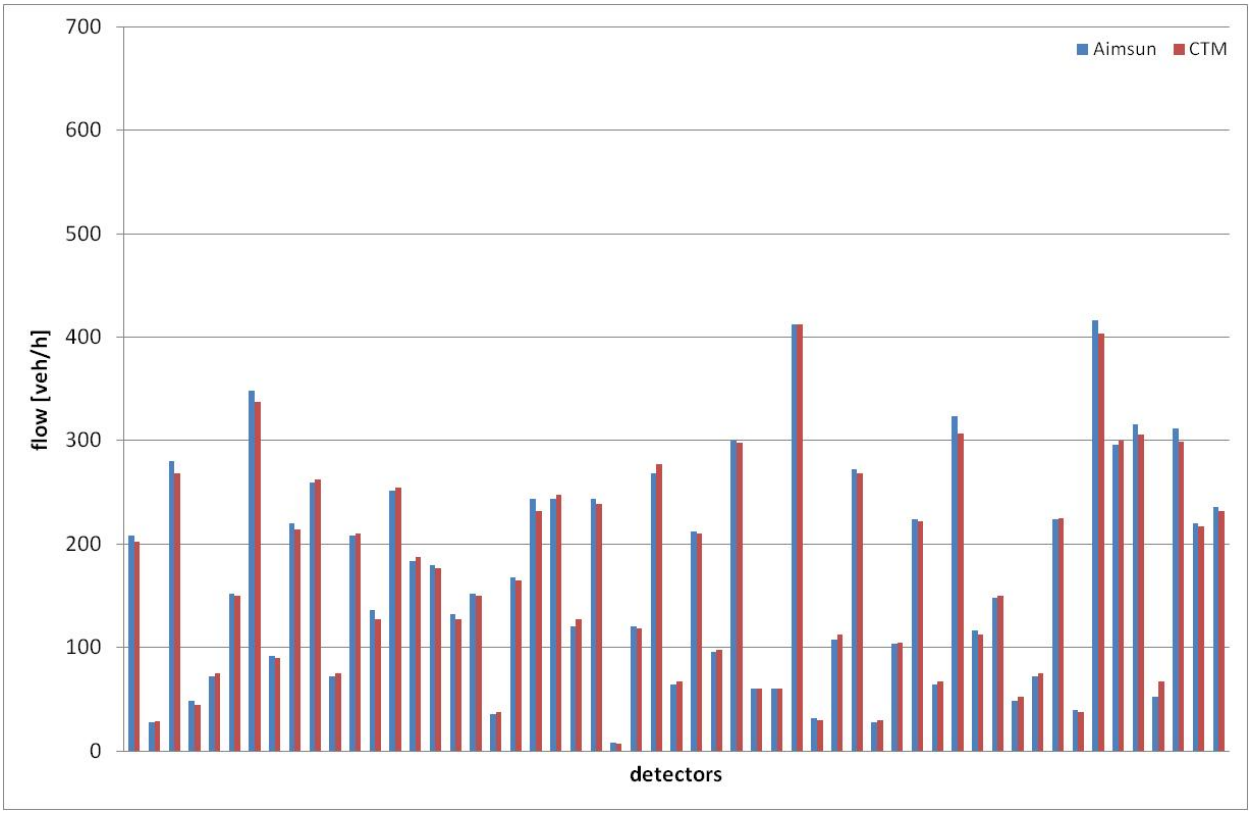


Figure A-3: Flow, morning peak (75 percent)

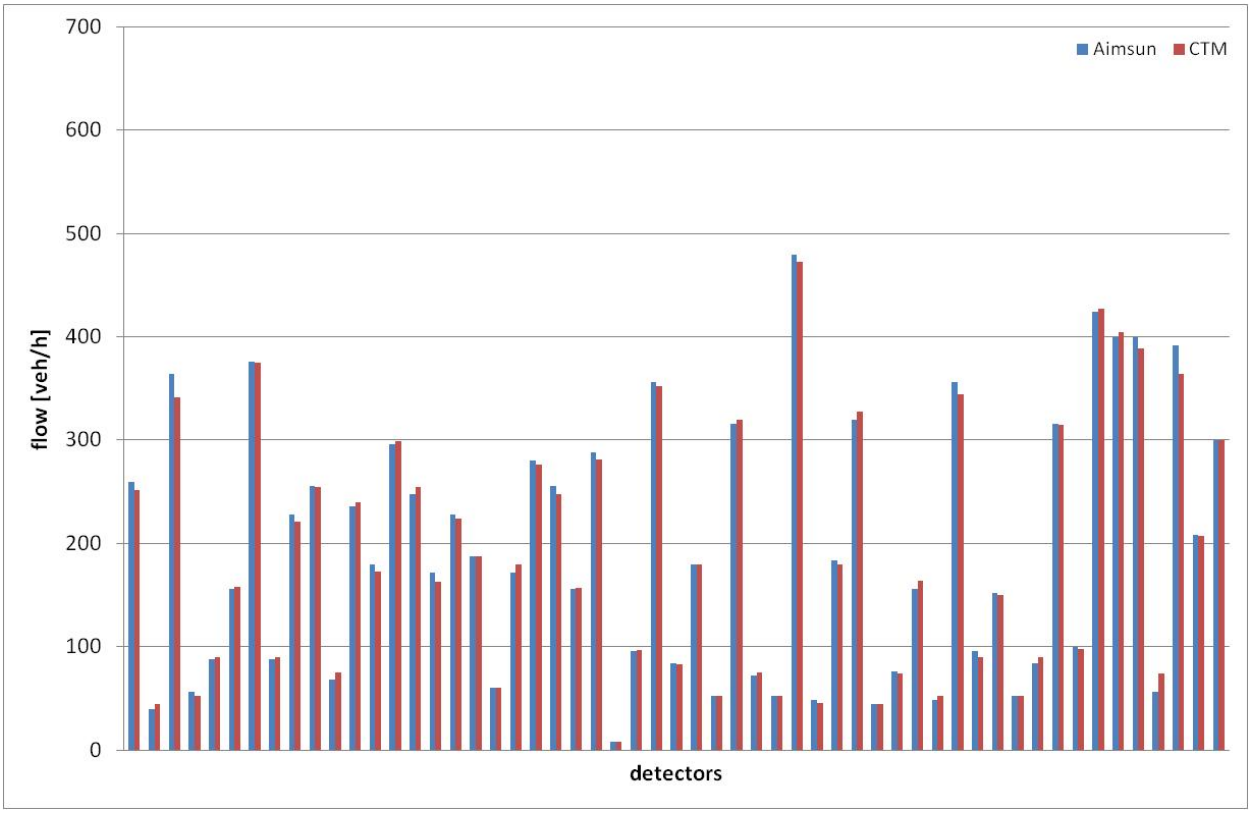


Figure A-4: Flow, afternoon peak (75 percent)

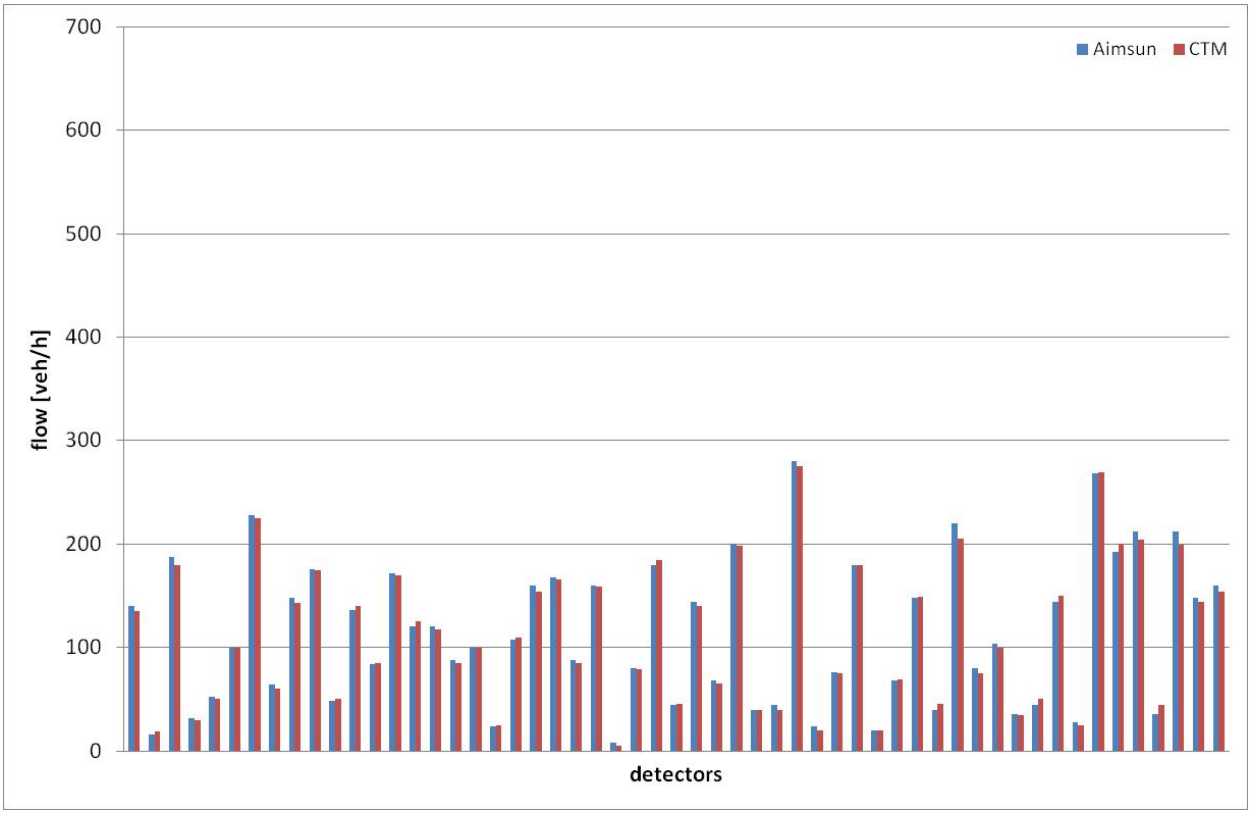


Figure A-5: Flow, morning peak (50 percent)

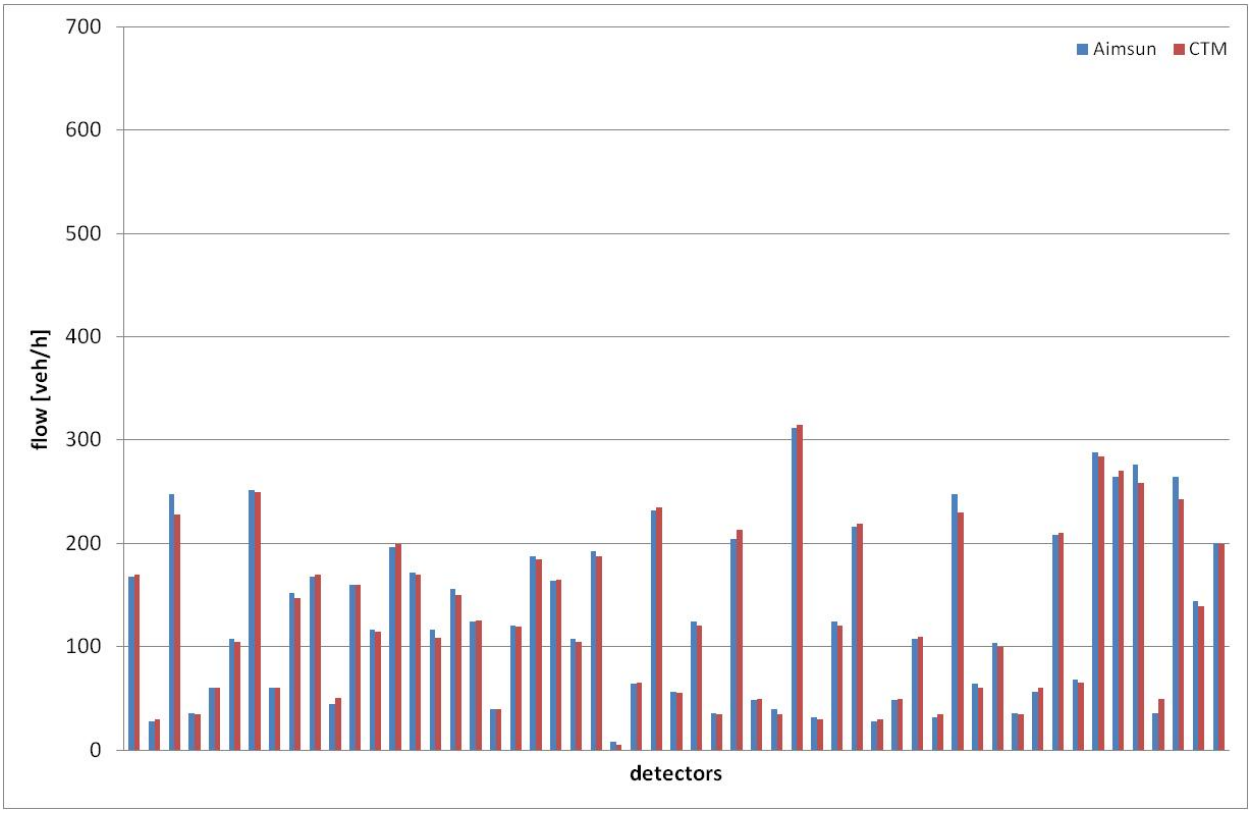


Figure A-6: Flow, afternoon peak (50 percent)

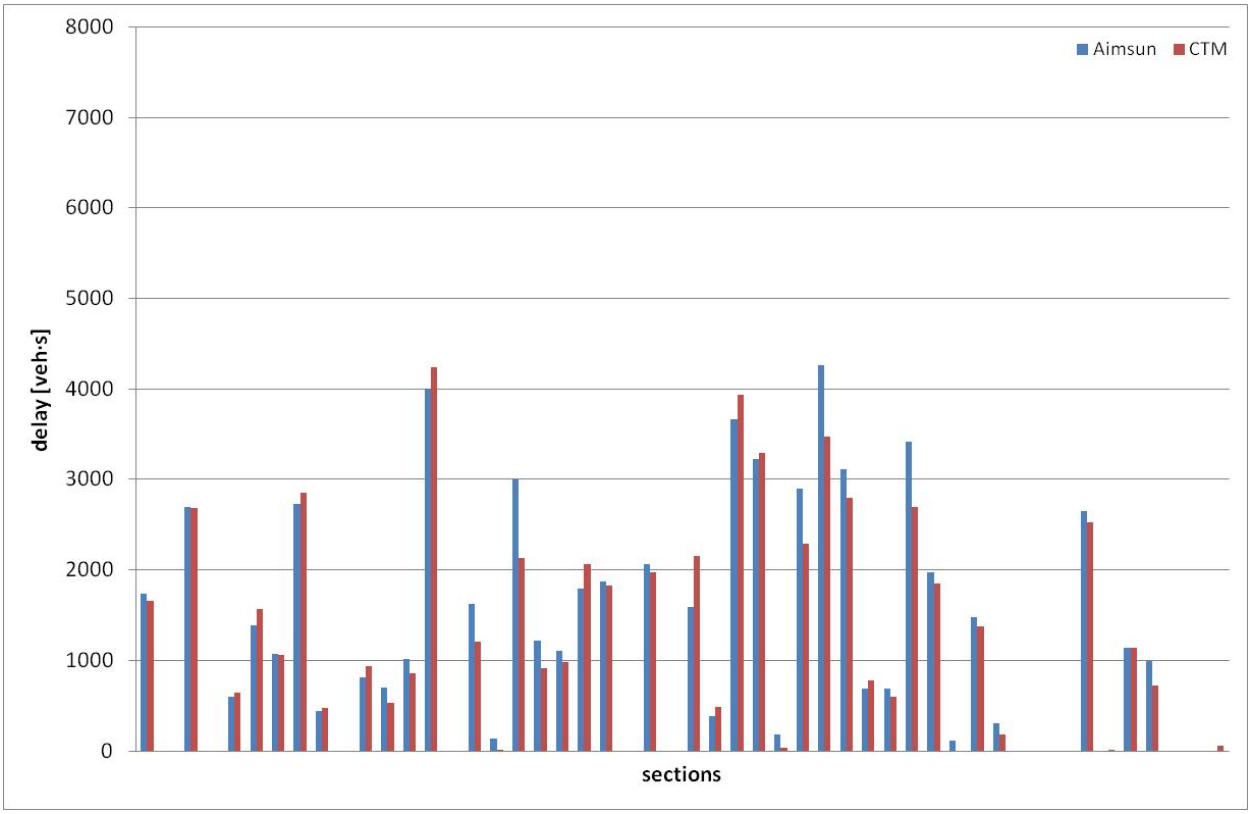


Figure A-7: Delay, morning peak (100 percent)

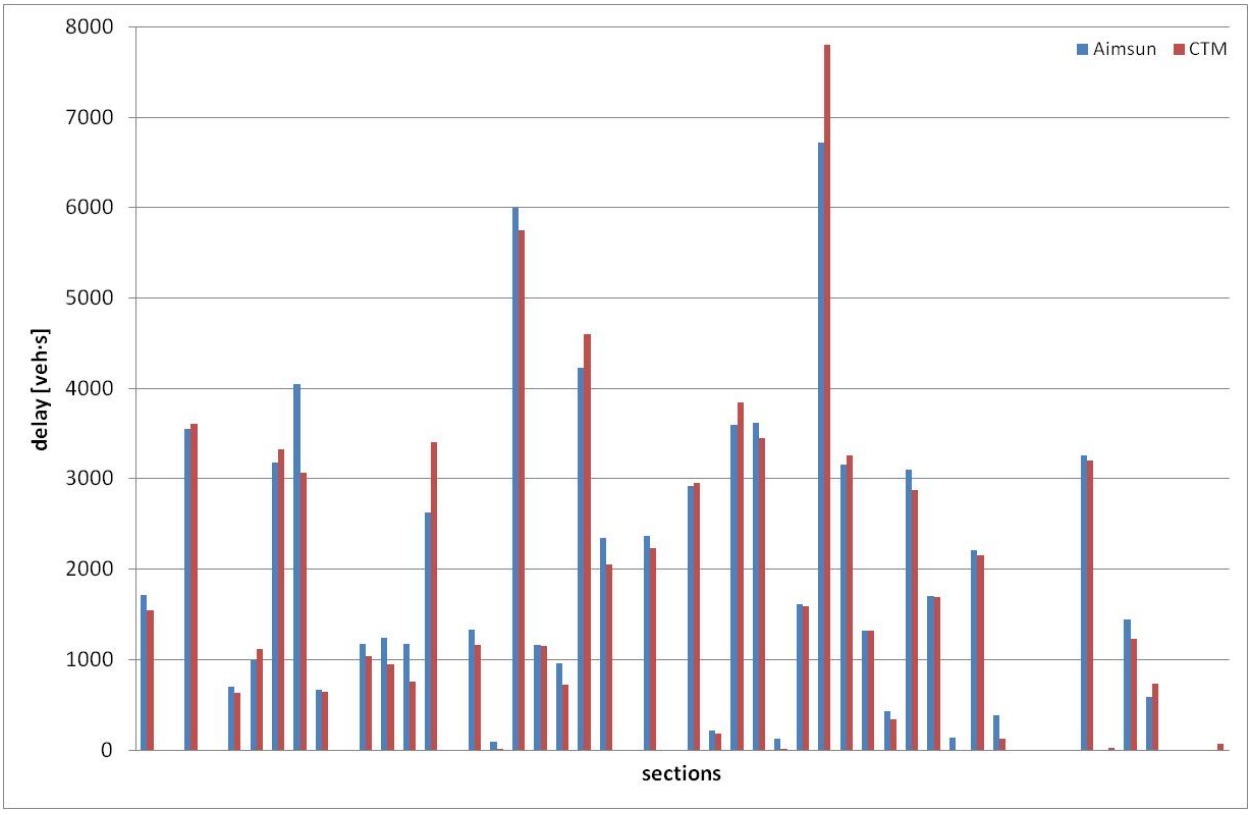


Figure A-8: Delay, afternoon peak (100 percent)

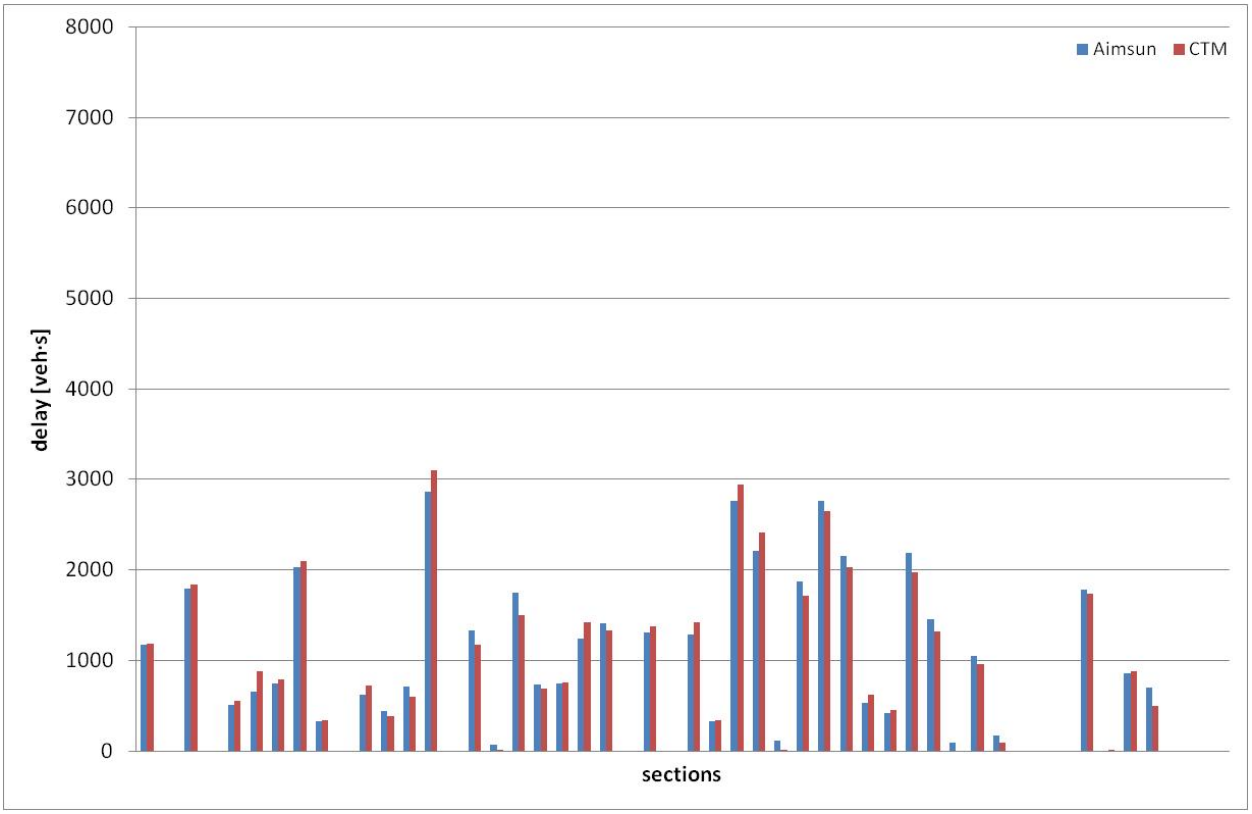


Figure A-9: Delay, morning peak (75 percent)

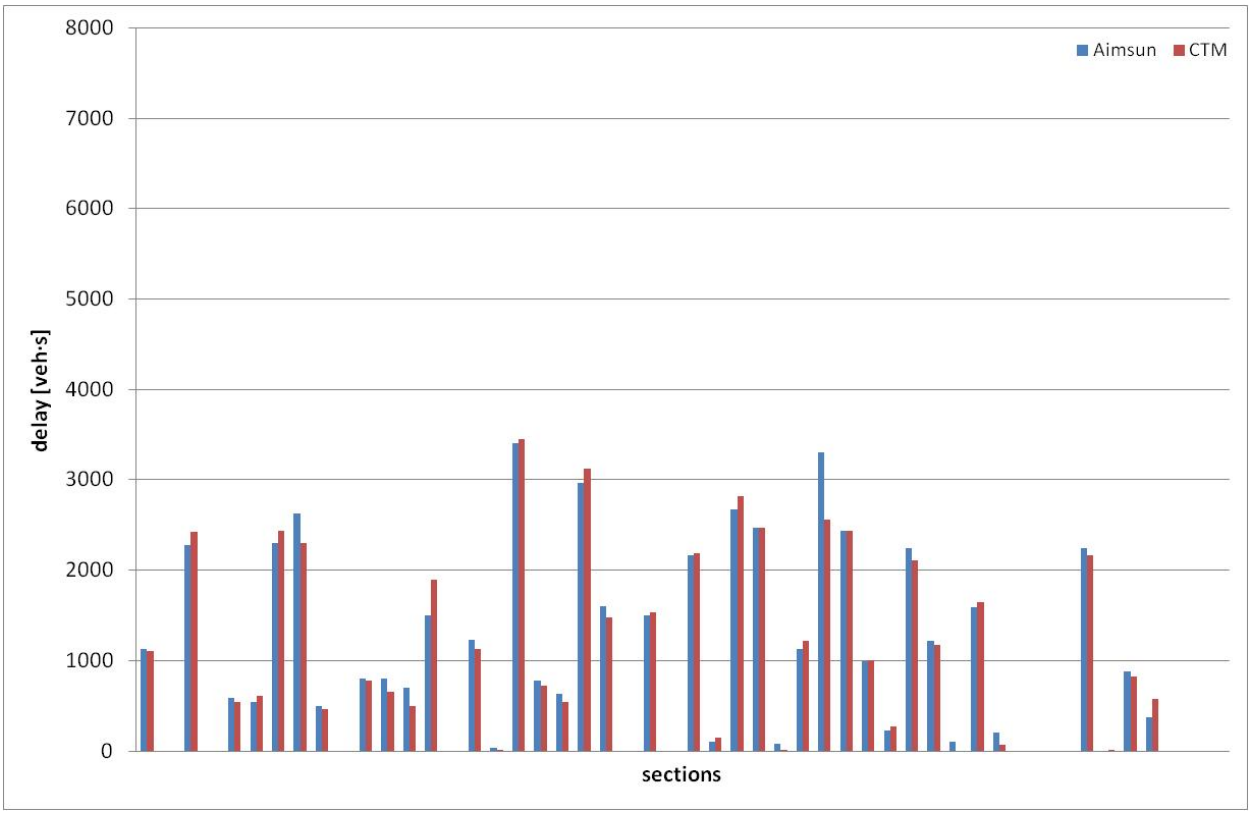


Figure A-10: Delay, afternoon peak (75 percent)

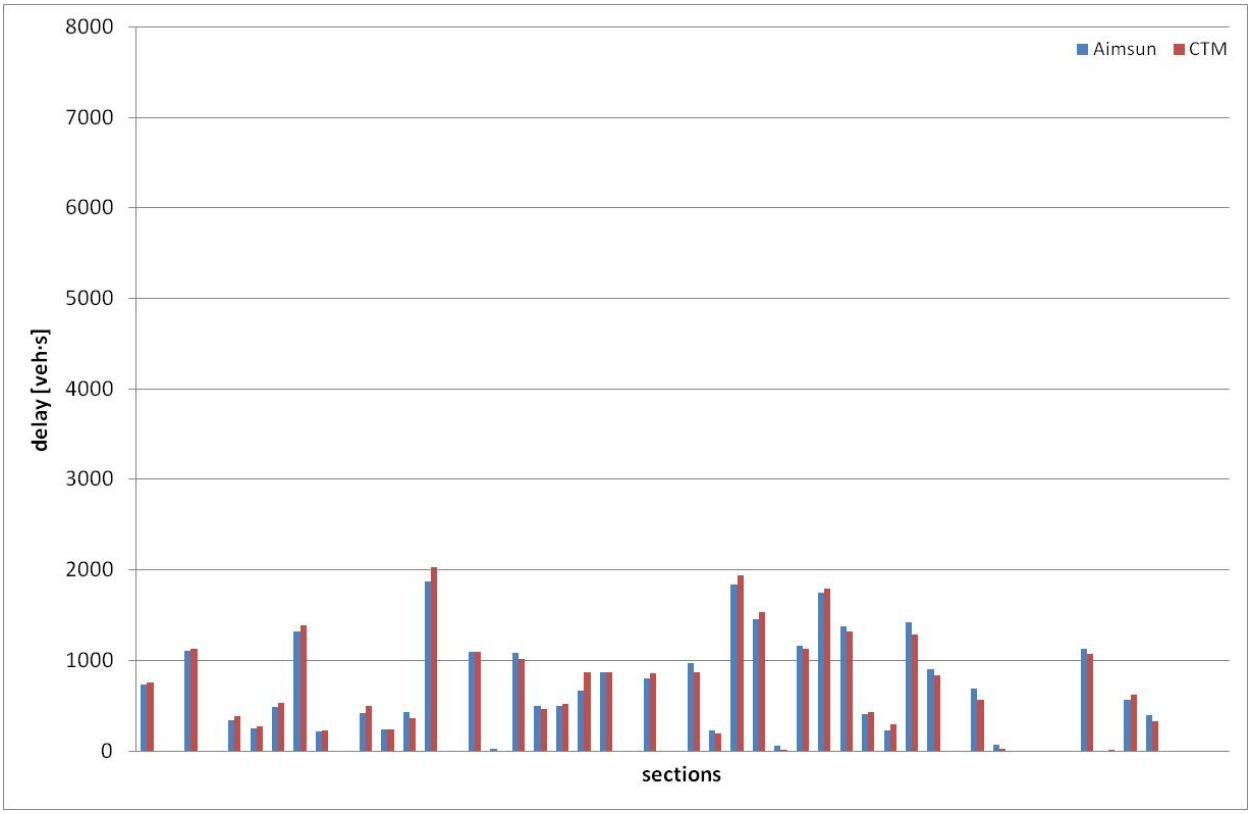


Figure A-11: Delay, morning peak (50 percent)

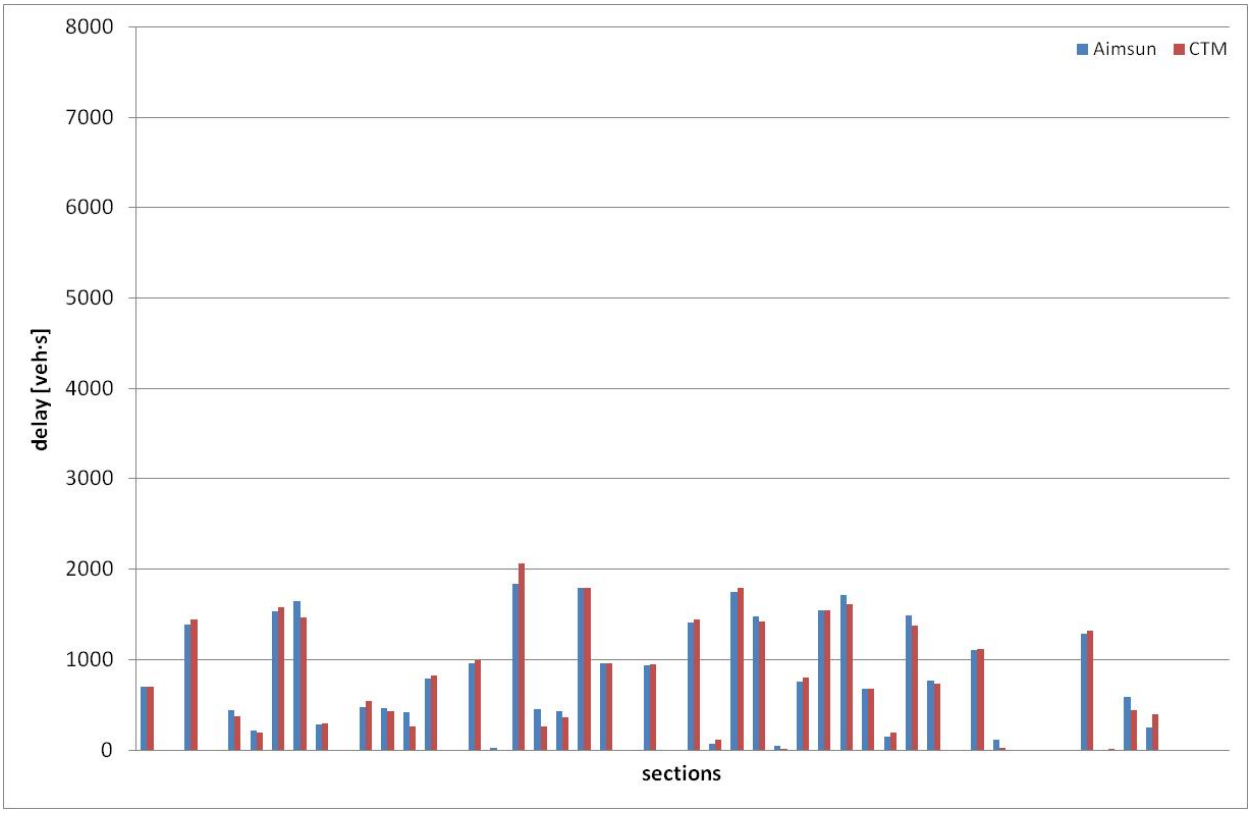


Figure A-12: Delay, afternoon peak (50 percent)

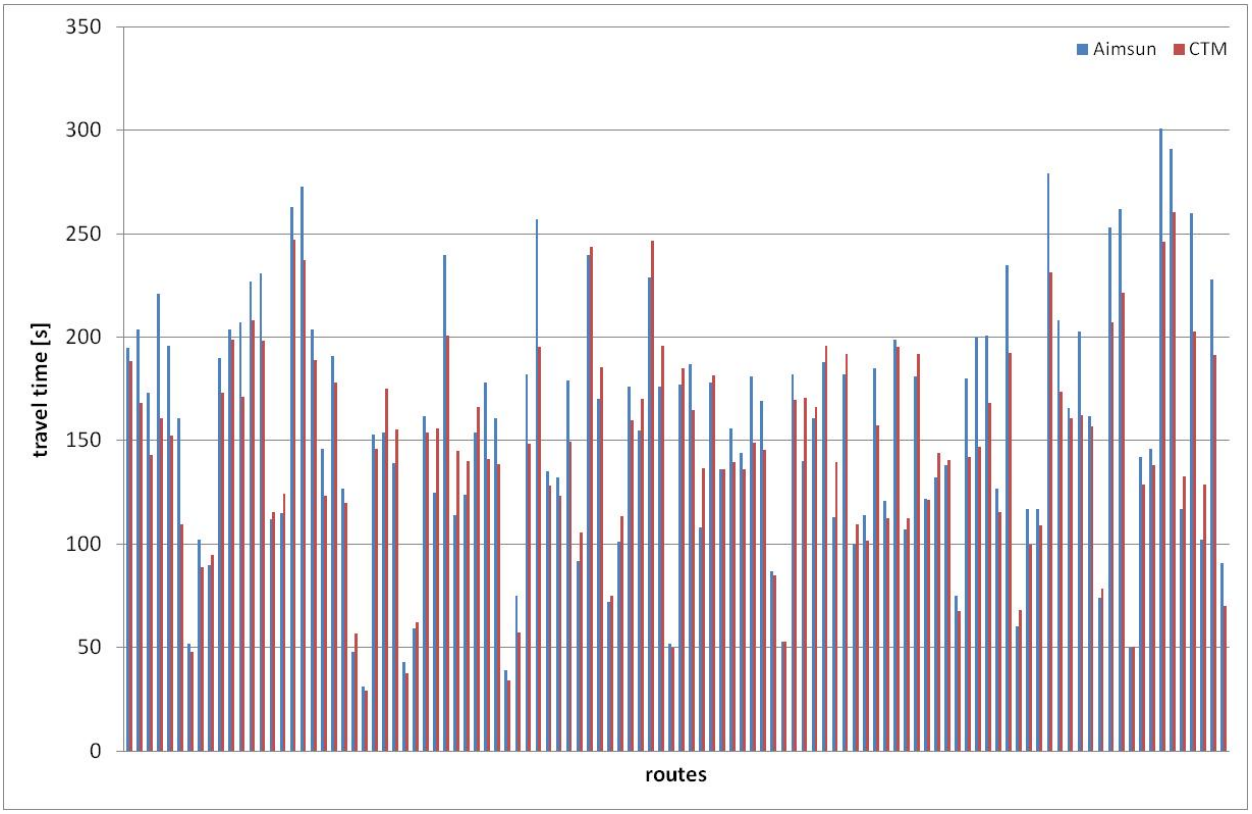


Figure A-13: Travel time, morning peak (100 percent)

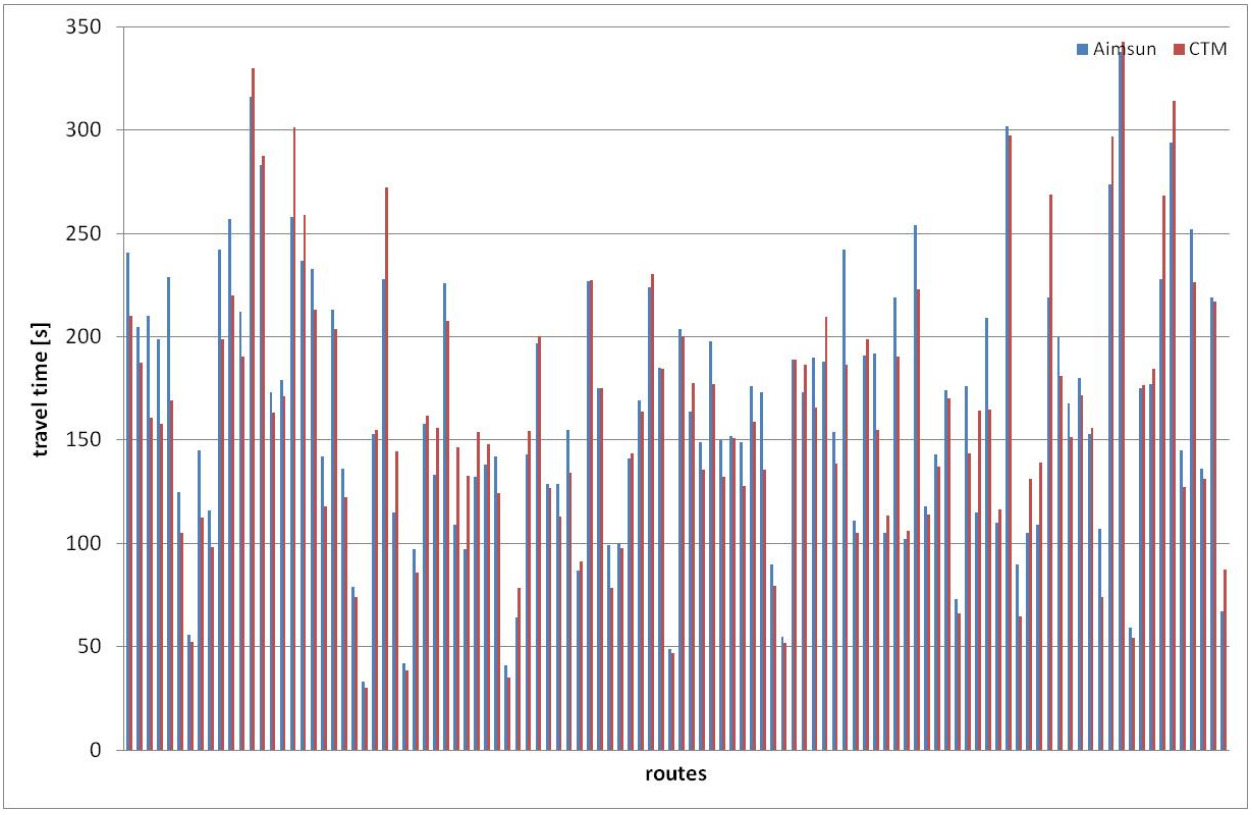


Figure A-14: Travel time, afternoon peak (100 percent)

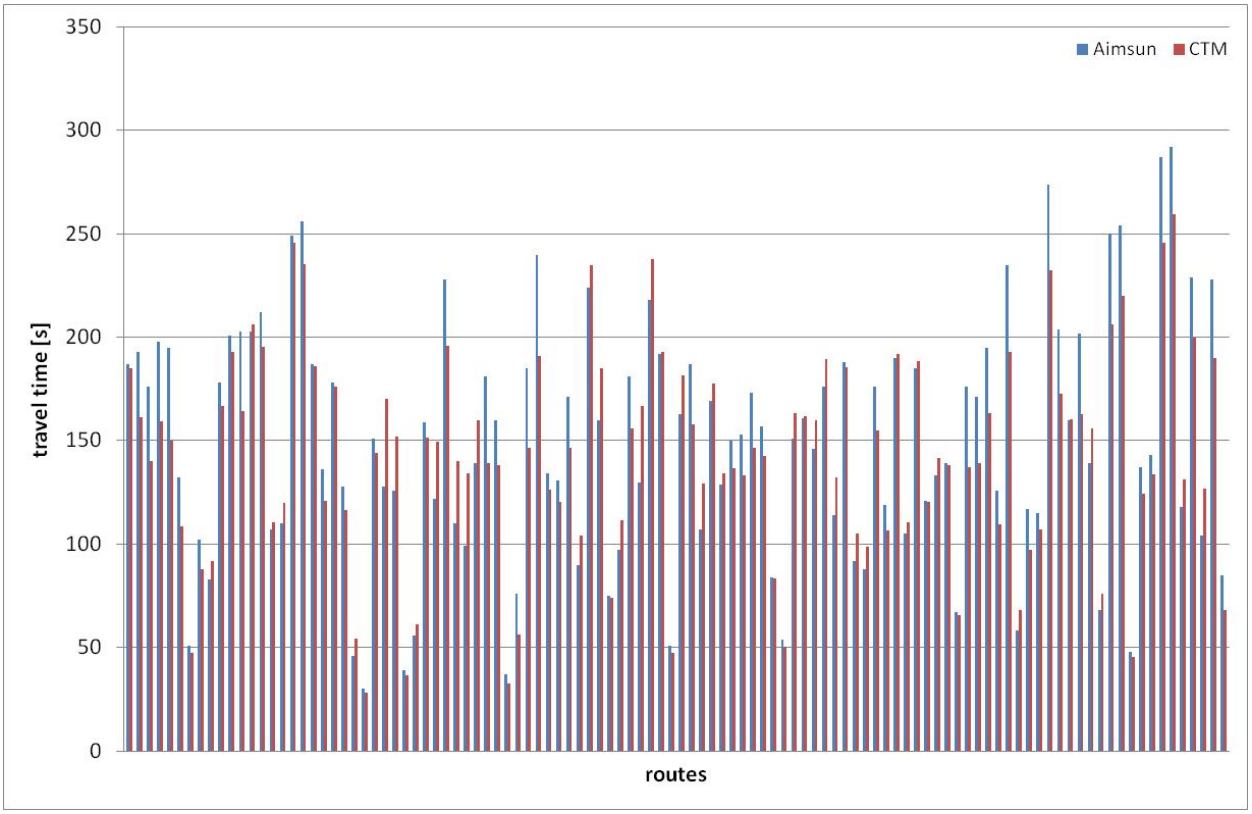


Figure A-15: Travel time, morning peak (75 percent)

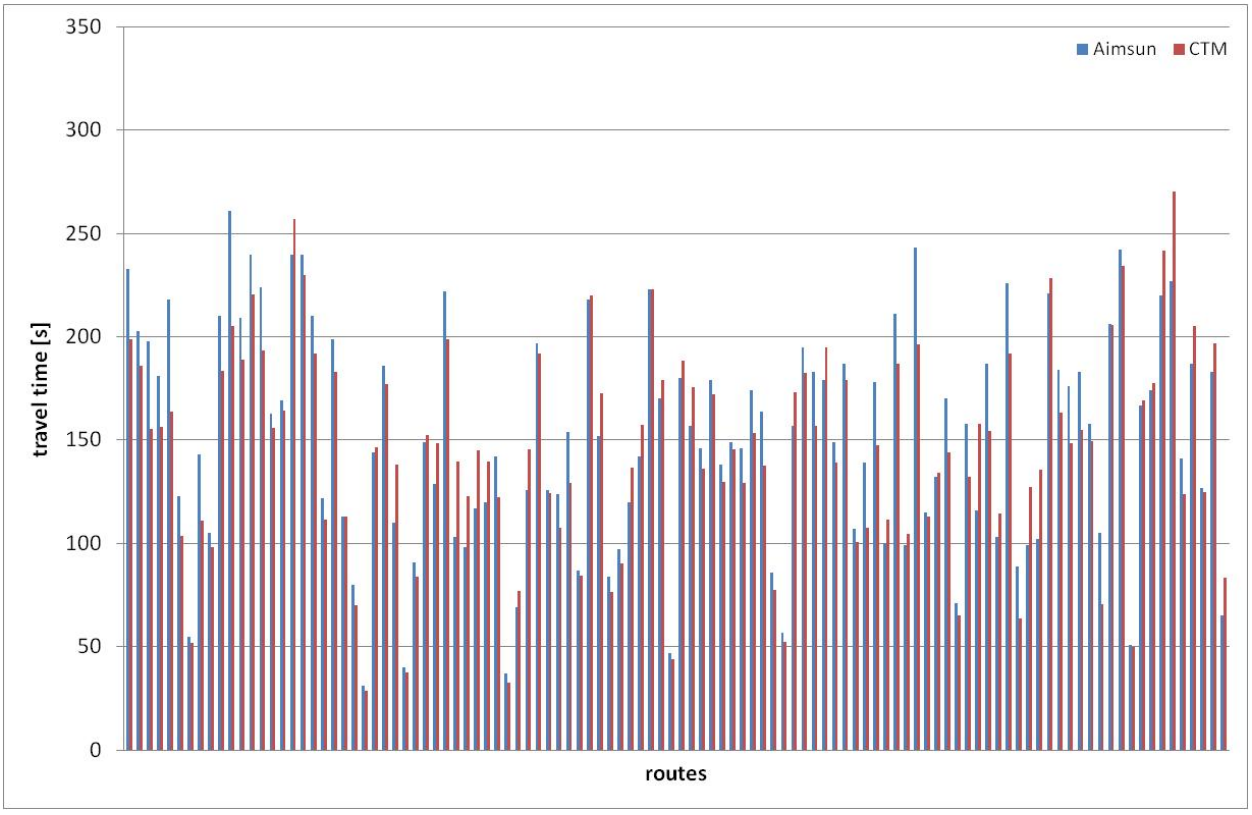


Figure A-16: Travel time, afternoon peak (75 percent)

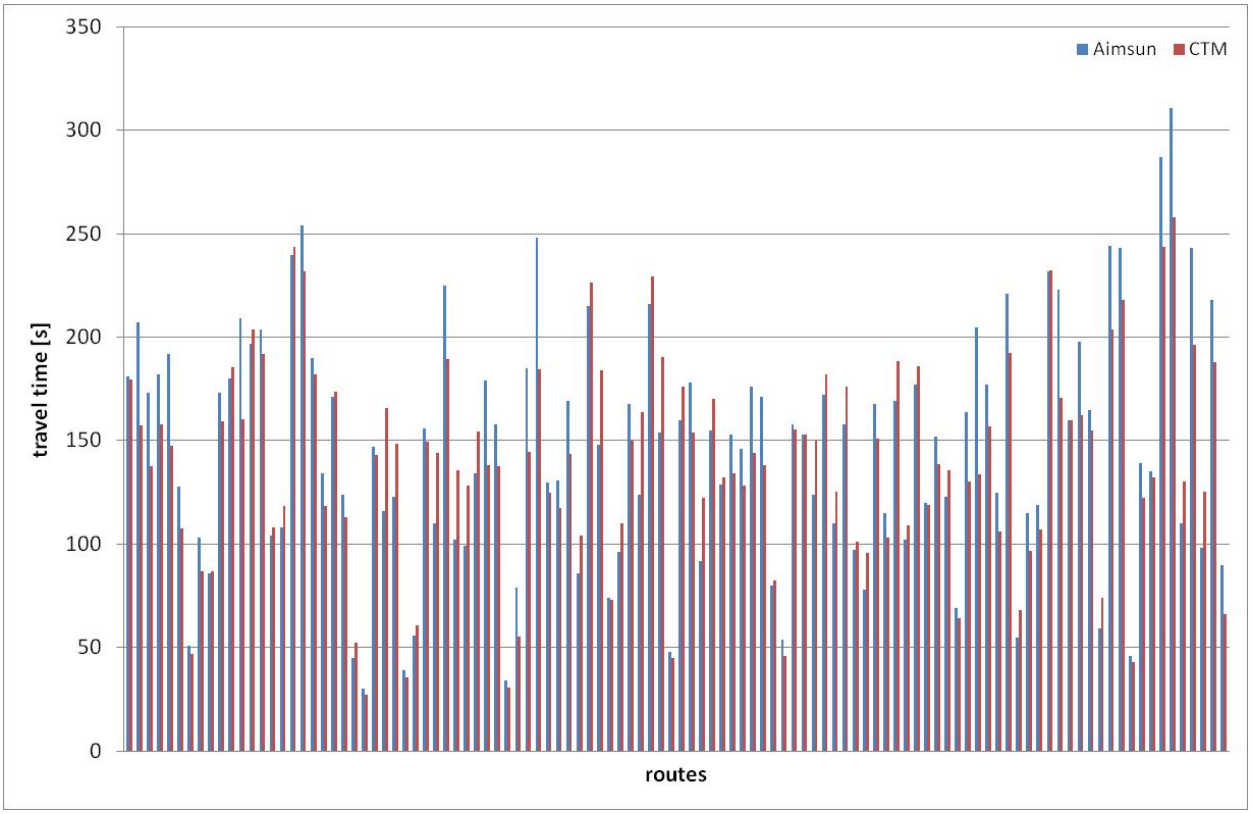


Figure A-17: Travel time, morning peak (50 percent)

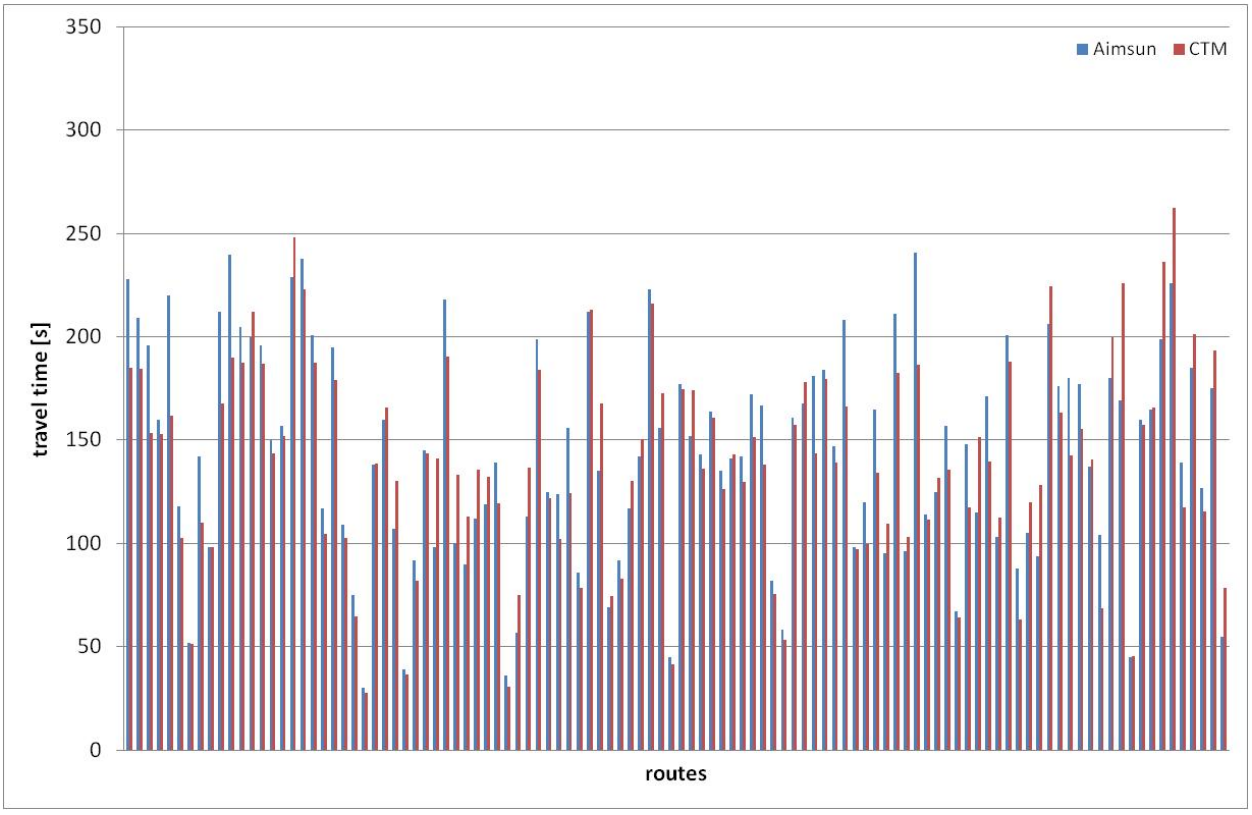


Figure A-18: Travel time, afternoon peak (50 percent)

B Theory and derivation of the Information Minimization model

This appendix provides an overview of the information theory which is the basis of the information minimization model. The specific problem of OD matrix estimation has been described in paragraph 5.3.3. It will be shown how equation 5-6 has been derived by VAN ZUYLEN/WILLUMSEN (1980) based on this theory. Finally, the approach will be discussed shortly.

B.1 Information theory

Information according to Brillouin

To understand what information according to information theory is, some first explanations by BRILLOUIN (1956) are helpful. In a system where P_0 different states or outcomes are possible with the same *a priori* probability, it cannot be said for sure without any further information which of the P_0 possibilities has finally arisen. The less certain the outcome is, i.e. the more possibilities P_0 exist, the more information is needed to know the exact current state of the system. BRILLOUIN (1956) defines the necessary quantity of information to

$$I_1 = K \cdot \ln P_0 . \qquad \text{(B-1)}$$

K is a scale factor and can be omitted (K = 1).

Using the logarithm of P_0 leads to additivity of information, which is very convenient. If there are two mutually independent problems, each of which with P_{01} and P_{02} possible and equally probable states, a combination of both problems has

$$P_0 = P_{01} \cdot P_{02} \qquad \text{(B-2)}$$

possible outcomes. Substituting B-2 in B-1 leads to

$$I_1 = K \cdot \ln P_0 = K \cdot \ln\left(P_{01} \cdot P_{02}\right) = K \cdot \ln P_{01} + K \cdot \ln P_{02} = I_{11} + I_{12} . \qquad \text{(B-3)}$$

The amount of information that is needed to know the final outcome of the combined problem is the sum of the information needed for each individual problem.

The base of the logarithm can be chosen arbitrarily according to the problem. In a binary system for instance one would use base 2. In this case the quantity of information $\log_2 P_0$ corresponds to the number of bits (“binary digits”) that are necessary to describe all possible outcomes P_0. According to WALSH/WEBBER (1977) the information in a binary system can be interpreted as the number of “binary” questions that have to be asked to identify the current state of the system. A “binary” question asks for the state of one specific bit, which can be either 0 or 1. In a system with 8 bits one would have to ask 8 “binary” questions to know which of the 2^8 possible states has occurred. If a more general base k is used, the quantity of information corresponds to the number of necessary “k-nary” questions. Information does thus reduce the uncertainty of a system. Equation B-1 can also be interpreted as a measurement of uncertainty. The more uncertain a specific state of a system is, the higher is the necessary quantity of information.

Information according to Shannon

In equation B-1 all P_0 different states occur with same probability. SHANNON (1948) who must be considered to be the originator of information theory considers the more general case where

different states x_1, x_2, …, x_i, … x_M occur with known *a priori* probabilities p_1, p_2, …, p_i, … p_M, so that holds

$$\sum_{j=1}^{M} p_j = 1. \qquad \text{(B-4)}$$

He defines a measurement of uncertainty of a specific outcome x_j, which is equivalent to the quantity of information or entropy $H(p_1, p_2, \ldots, p_i, \ldots p_M)$ of a system.

$$H = I = -K \cdot \sum_{j=1}^{M} p_j \log_k p_j \; . \qquad \text{(B-5)}$$

Again, base k of the logarithm is arbitrary.

WALSH/WEBBER (1977) point out that equation B-5 is an axiom that has been postulated by SHANNON (1948). The definition of information or entropy had to fulfil some requirements. H or I should be continuous at all p_j of its domain. And it should reach its maximum at the point where all p_j are equal (i.e. all states x_j have the same probability of occurrence) because in this case uncertainty about a specific state of the system is highest. Equation B-5 fulfils all of these (and some other) requirements.

BRILLOUIN (1956) showed how his definition of information (equation B-1) can be transformed into the definition by SHANNON (1948) (equation B-5). Details about this transformation can be found there. In addition, WALSH/WEBBER (1977) made some remarks on the formal difference between the two equations. However, this information is dispensable in the context of this thesis.

Interpretation of information

Equation B-5 expresses the quantity of information of an arbitrary state of a system where all p_j of all possible states x_j are known. This quantity of information is an average measurement of information contained in a system, no matter which state x_j finally occurs. As has been said, the quantity of information I is at its maximum at the point where all p_j are equal. Figure B-1 shows the information contained in a system with two possible states for varying probabilities p_1 and $p_2 = 1 - p_1$.

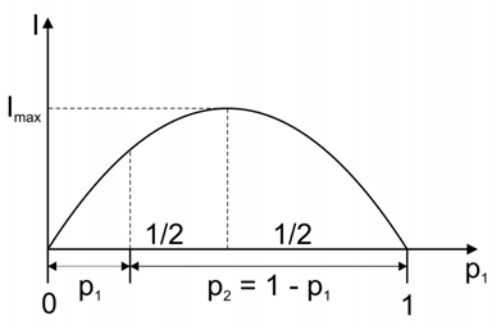


Figure B-1: Information *I* in the case of two possible states with probabilities p_1 and p_2

The maximum information needed to deduce the current state of the system emerges if $p_1 = p_2$, because in this case both states are equally likely. If one case is less likely than the other, uncertainty of the state and thus information is reduced. If one of the two probabilities is

equal to zero, the final outcome is absolutely certain and the system contains no information at all.

As BRILLOUIN (1956) showed, even the information contained in a system with three possible states can be illustrated graphically. The left part of figure B-2 shows a triangular area. Each point on this area represents one possible combination of the three probabilities p_1, p_2 and p_3, so that holds: $p_1 + p_2 + p_3 = 1$. The right part of figure B-2 shows the according information of each point, plotted vertically onto the area. The curves along the edges of the area are equal to the curve in figure B-1. The maximum of I can be found at the centroid of the triangle, where $p_1 = p_2 = p_3 = 1/3$.

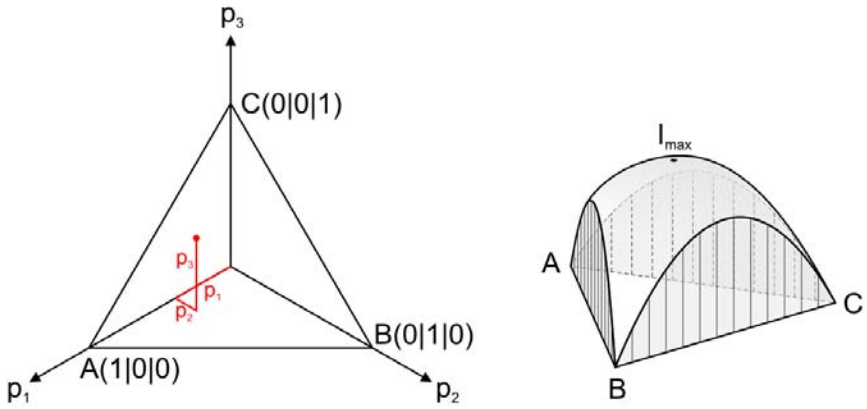


Figure B-2: Information *I* in the case of three possible states with probabilities p_1, p_2 and p_3

In general it can be said that if the state of a system can be predicted easily, it contains only little information. If it is hard to predict, it contains a lot of information. Or in the words of BRILLOUIN (1956): „If a situation is scarce, it contains information."

Information according to Goldmann

According to VAN ZUYLEN (1979) the information minimization model is based on definition B-1 by BRILLOUIN (1956). However, it is more practical to use a third definition of information to derive the model as will be argued later. This third definition is the one by GOLDMANN (1953). It can easily be transformed into the definitions by BRILLOUIN (1956) and SHANNON (1948).

GOLDMANN (1953) defines the information that is contained in the observation of a specific state x_j to

$$I(X = x_j) = \log_k \left(\frac{p_1(X = x_j)}{p_0(X = x_j)} \right), \qquad \text{(B-6)}$$

where $p_0(X = x_j)$ is the probability of occurrence of state x_j *before* the observation and $p_1(X = x_j)$ is the same probability *after* the observation. According to WALSH/WEBBER (1977) the latter of the two probabilities is 1, given an error-free observation, because state x_j has definitely occurred. The first probability is equal to the *a priori* probability p_j of state x_j. Equation B-6 can thus be transformed to

$$I(X = x_j) = \log_k \left(\frac{1}{p_j} \right) = -\log_k p_j \,. \qquad \text{(B-7)}$$

WALSH/WEBBER (1977) point out that equation B-7 is mathematically equivalent to equation B-1 by BRILLOUIN (1956). In equation B-1 all P_0 possible states are equally likely. Every state x_j has the same probability

$$p_j = \frac{1}{P_0} \,. \qquad \text{(B-8)}$$

Substitution of equation B-8 in equation B-7 leads to

$$I(X = x_j) = -\log_k \left(\frac{1}{P_0} \right) = \log_k P_0 \,. \qquad \text{(B-9)}$$

If the natural logarithm is used, equation B-9 is equal to equation B-1 with K = 1.

WALSH/WEBBER (1977) also show that SHANNON's definition can be easily deduced from equation B-7 as well. SHANNON's equation gives the average information that is obtained by observation of an arbitrary state x_j out of M possible states. Thus, the contents of information of all possible states have to be weighted according to their probability of occurrence as follows:

$$I = \sum_{j=1}^{M} p_j \cdot I(X = x_j) = -\sum_{j=1}^{M} p_j \cdot \log_k p_j \,. \qquad \text{(B-10)}$$

This corresponds to equation B-5 with K = 1.

Figure B-3 shows the amount of information according to GOLDMANN (1953) which is contained in the observation of a specific state x_j with probability p_j. The same comments as before can be made. Observation of a rare state contains more information. Observation of a state with probability p_j = 1 contains no information at all. Its prediction is trivial.

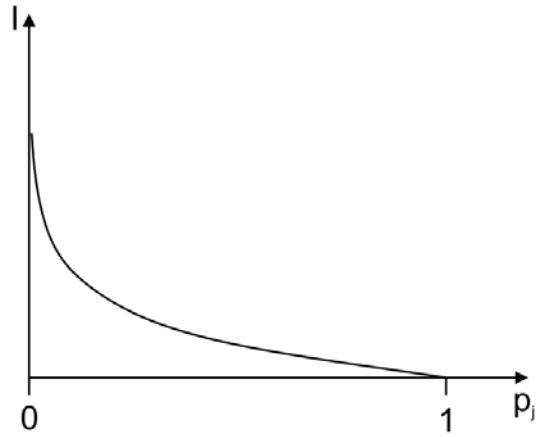


Figure B-3: Information according to the definition by GOLDMAN (1953)

B.2 Information Minimization model

Basic argumentation

The problem to be solved by the IM model has been described in paragraph 5.3.3: find a matrix of OD flows f_{ij} which satisfy the constraints in the form of equation 5-5. Since in general the according linear system of equations is under-determined, no unique solution exists. VAN ZUYLEN (1979) and VAN ZUYLEN/WILLUMSEN (1980) therefore choose the matrix that contains the least information, i.e. the most likely matrix according to information theory.

VAN ZUYLEN (1979) argues that the IM model is based on the definition of information by BRILLOUIN (1956). He states that the information contained in a series of G observations where state x_j has been observed N_j times is

$$I = \ln\left[G! \cdot \prod_{j=1}^{M} \frac{\hat{N}_j^{N_j}}{N_j!} \right]. \qquad \text{(B-11)}$$

It has to be pointed out that VAN ZUYLEN (1979) mistakenly uses the measurement of information defined by BRILLOUIN (1956) with a negative algebraic sign. This fact will be discussed later. In this appendix, the correct version with positive sign is used.

$\hat{N}_j$ is the absolute frequency of state x_j in an earlier sequence of observations. This historic information is used to estimate the probability of occurrence of each state x_j. It has to be highlighted at this point that according to the reasoning of VAN ZUYLEN (1979) $\hat{N}_j$ is in fact an integer number of occurrence, and not a normalized probability p_j.

Since VAN ZUYLEN (1979) explicitly refers to equation B-1, the square bracket term in equation B-11 corresponds to the number P_0 of all possible outcomes of a system in a given situation. To illuminate this fact, the square bracket term is transformed as follows:

$$P_0 = G! \cdot \prod_{j=1}^{M} \frac{\hat{N}_j^{N_j}}{N_j!} = \frac{G!}{\prod_{j=1}^{M} N_j!} \cdot \prod_{j=1}^{M} \hat{N}_j^{N_j} \; . \qquad \text{(B-12)}$$

The urn model enables better understanding of equation B-12. An urn contains M balls of different colors. Each color represents a different state x_j. The respective numbers $\hat{N}_j$ of balls of a specific color in the urn are equal to the numbers observed in an earlier drawing. One ball at a time is drawn from the urn, its color is noted and the ball is placed back into the urn. The number Z of possible drawings that result in a specific and unique sequence of colors x_j in which each color occurs N_j times is:

$$Z = \prod_{j=0}^{M} \hat{N}_j^{N_j} \; . \qquad \text{(B-13)}$$

Note that in a unique sequence of colors, which can be drawn in Z different ways, all balls of one color can switch their positions without changing the sequence.

Now that the number of possible drawings that produce a specific and unique sequence of colors is known, the number of distinguishable sequences that can be produced from G balls

which comprise N_j balls of each color j has to be determined. According to the laws of combinatorics this number is:

$$P = \frac{G!}{\prod_{j=1}^{M} N_j!}. \qquad \text{(B-14)}$$

The product of equation B-13 and B-14 is the total number P_0 of all possible ways to draw and place back balls from the urn so that they result in an arbitrary sequence of G balls in which every color j occurs N_j times. This is expressed in equation B-12. A drawing of G balls that comprise each color with absolute frequency N_j in any order can be realized in P_0 different ways which are all equally likely.

VAN ZUYLEN (1979) claims that the historic frequencies $\hat{N}_j$ in equation B-11 do not necessarily have to be transformed to normalized probabilities p_j of occurrence. He argues that this would just result in a different scaling factor in equations B-12 and B-11. Against this statement it is nevertheless helpful to use probabilities p_j so that

$$p_j = \frac{\hat{N}_j}{\hat{G}} \qquad \text{(B-15)}$$

where

$$\hat{G} = \sum_{j=1}^{M} \hat{N}_j \,. \qquad \text{(B-16)}$$

A normalized version of equation B-12 can also be obtained by dividing the number P_0 of all possible drawings that result in sequences with frequencies N_j of each state by the total number P_{tot} of all possible drawings. The latter is

$$P_{tot} = \hat{G}^G \,, \qquad \text{(B-17)}$$

because each of the G balls is drawn out of $\hat{G}$ balls. Dividing equation B-12 by equation B-17 and considering equations B-15 and B-16 gives

$$\begin{aligned} P_{0,norm} &= \frac{\frac{G!}{\prod_{j=1}^{M} N_j!} \cdot \prod_{j=1}^{M} \hat{N}_j^{N_j}}{\hat{G}^G} = \frac{G!}{\prod_{j=1}^{M} N_j!} \cdot \prod_{j=1}^{M} \frac{\hat{N}_j^{N_j}}{\hat{G}^{N_j}} = \frac{G!}{\prod_{j=1}^{M} N_j!} \cdot \prod_{j=1}^{M} \left(\frac{\hat{N}_j}{\hat{G}} \right)^{N_j} \\ &= \frac{G!}{\prod_{j=1}^{M} N_j!} \cdot \prod_{j=1}^{M} p_j^{N_j}. \end{aligned} \qquad \text{(B-18)}$$

Substitution of equation B-18 in equation B-1 leads to a normalized version of equation B-11 postulated by VAN ZUYLEN (1979). As has been said, VAN ZUYLEN – knowingly (?) – uses a wrong negative algebraic sign in equation B-11. This results in a negative sign in front of the normalized version of this equation, which is correct as will be shown in the next paragraph. If

however VAN ZUYLEN had used equation B-1 correctly, i.e. with a positive sign, the normalized version would be wrong.

Equation B-18 is the so-called multinomial distribution. In this form it does not longer give the number of equally likely sequences in which each state x_j occurs N_j times but the probability that G states x_j with *a priori* probabilities p_j occur N_j times each. According to this, the derivation of the normalized version of equation B-11 can be simplified considerably if the measurement of information by GOLDMANN (1953) given in equation B-7 is used instead of the measurement by BRILLOUIN (1956) given in equation B-1. The probability $p(X_1 = N_1 \wedge X_2 = N_2 \wedge \ldots \wedge X_j = N_j \wedge \ldots \wedge X_M = N_M)$ can be determined directly with equation B-18. Substitution of equation B-18 as probability p_j in equation B-7 (i.e. the finally observed state x_j in equation B-11 corresponds to the case in which an observed sequence of different possible individual states contains each of these possible states with frequency N_j) gives

$$I = -\ln\left[\frac{G!}{\prod_{j=1}^{M} N_j!} \cdot \prod_{j=1}^{M} p_j^{N_j}\right]. \qquad \text{(B-19)}$$

This equation has the necessary negative sign which would not arise from the derivation of VAN ZUYLEN (1979) if he had used equation B-1 correctly.

Starting from equation B-19 (with wrong sign) VAN ZUYLEN/WILLUMSEN (1980) explain the derivation of the IM model. Since this derivation needs a negative sign, they just add it during the first transformation without giving any explication. Obvisouly the original derivation has an inconsistency. Therefore, the derivation using the measurement of information by GOLDMANN (1953) as presented in this appendix should be preferred over the derivation by VAN ZUYLEN (1979) using the measurement of information by BRILLOUIN (1956).

A simple example shows that the correct version of equation B-19 has a negative sign. Two states *A* and *B* occur with probabilities p_A = 0.75 and p_B = 0.25. In this case the bracket term in equation B-19 transforms from multinomial to binomial distribution. Figure B-4 shows the probability and the according measurement of information according to equation B-19 for all frequencies of occurrence N_A of state *A* when there are G = 100 observations.

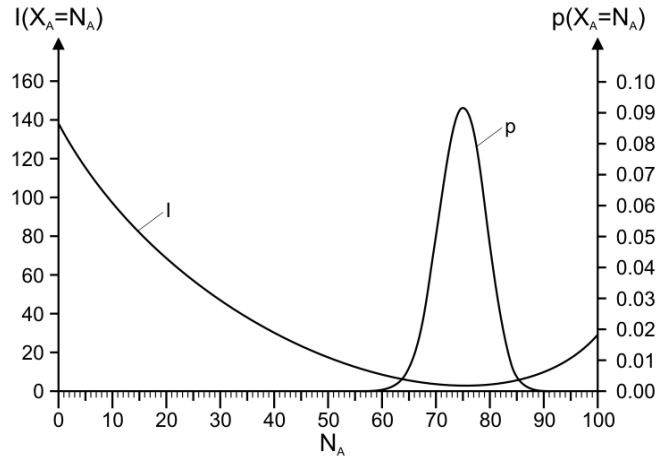


Figure B-4: Probability p and measurement I of information X_A=N_A und G=100

As expected 75 occurrences of state *A* have the highest probability. At this point the information contained in such an observation is lowest. The less likely a frequency of occurrence N_A is, the higher is the information contained in N_A observations of state *A*.

Derivation of an objective function

Until now, only the general case using general states x_j has been presented. VAN ZUYLEN/WILLUMSEN (1980) translate this general case to the case of OD matrix estimation. They want to express the information contained in an observed flow q_a on a link *a* by using equation B-19. The number q_a of observed vehicles on link *a* corresponds to the number *G* of observations, i.e.

$$G = q_a \,. \qquad \text{(B-20)}$$

Every observed vehicle can have a different state x_j, i.e. it belongs to one of the possible relations or OD pairs *ij* that pass link *a*. N_j in equation B-19 becomes

$$N_j = f_{ij} \cdot p_{ij}^a \,. \qquad \text{(B-21)}$$

As defined in paragraph 5.3.3 f_{ij} is the flow from origin *i* to destination *j* and p_{ij}^a is the portion of vehicles passing link *a* while travelling from *i* to *j*.

The *a priori* probability that a vehicle observed on link *a* travels from *i* to *j* has to be estimated from historic observations at the same link. Given the flows f_{ij} of a historic OD matrix, p_j in equation B-19 becomes

$$p_j = \frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a} \,. \qquad \text{(B-22)}$$

Substitution of equations B-20 to B-22 in equation B-19 leads to the amount of information that is contained in an observation of vehicles on link *a*:

$$I_a = -\ln\left[\frac{q_a!}{\prod_{ij}\left(f_{ij} \cdot p_{ij}^a\right)!} \cdot \prod_{ij}\left(\frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a}\right)^{f_{ij} \cdot p_{ij}^a}\right]. \qquad \text{(B-23)}$$

Equation B-23 is transformed as follows (cf. WANG, 2008, ROHDE, 2006):

$$
\begin{aligned}
I_a &= -\left[\ln q_a! - \ln \prod_{ij} \left(f_{ij} \cdot p_{ij}^a \right)! + \ln \prod_{ij} \left(\frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a} \right)^{f_{ij} \cdot p_{ij}^a} \right] \\
&= -\left[\ln q_a! - \sum_{ij} \ln \left(f_{ij} \cdot p_{ij}^a \right)! + \sum_{ij} \ln \left(\frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a} \right)^{f_{ij} \cdot p_{ij}^a} \right] \\
&= -\left[\ln q_a! - \sum_{ij} \ln \left(f_{ij} \cdot p_{ij}^a \right)! + \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(\frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a} \right) \right]
\end{aligned}
\qquad \text{(B-24)}
$$

Equation B-24 can be further transformed using Stirling's approximation

$$
\ln n! \approx n \cdot \left(\ln n - 1 \right), \qquad \text{(B-25)}
$$

so that

$$
I_a \approx -\left[\begin{array}{l} q_a \cdot \ln q_a - q_a - \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(f_{ij} \cdot p_{ij}^a \right) + \\ \sum_{ij} f_{ij} \cdot p_{ij}^a + \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(\frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a} \right) \end{array} \right] \qquad \text{(B-26)}
$$

Substituting equation 5-5 in paragraph 5.3.3 gives:

$$
\begin{aligned}
I_a &\approx -\left(\sum_{ij} f_{ij} \cdot p_{ij}^a \right) \cdot \ln q_a + \sum_{ij} f_{ij} \cdot p_{ij}^a + \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(f_{ij} \cdot p_{ij}^a \right) \\
&\quad - \sum_{ij} f_{ij} \cdot p_{ij}^a - \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(\frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a} \right) \\
&= \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \left[-\ln q_a + \ln \left(f_{ij} \cdot p_{ij}^a \right) - \ln \left(\frac{f_{ij}^0 \cdot p_{ij}^a}{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a} \right) \right] \qquad \text{(B-27)} \\
&= \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(\frac{f_{ij} \cdot p_{ij}^a}{q_a} \cdot \frac{\sum_{ij} f_{ij}^0 \cdot p_{ij}^a}{f_{ij}^0 \cdot p_{ij}^a} \right) = \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(\frac{f_{ij} \cdot \sum_{ij} f_{ij}^0 \cdot p_{ij}^a}{q_a \cdot f_{ij}^0} \right) \\
&= \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln \left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0} \right)
\end{aligned}
$$

where

$$q_a^0 = \sum_{ij} f_{ij}^0 \cdot p_{ij}^a \; . \qquad \text{(B-28)}$$

Equation B-27 expresses the amount of information contained in the observation on one link *a*. Since all links with counts have to be considered as constraints for the matrix estimation, all individual amounts of information of these links have to be summed up.

$$I = \sum_a \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln\left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right) \qquad \text{(B-29)}$$

The most likely matrix according to information theory is the one that contains the least information. Therefore a solution $\{f_{ij}\}$ has to be found that minimizes equation B-29 subject to the constraints given by equation 5-5 in paragraph 5.3.3.

Solution of the problem of minimization

To solve the problem of minimization VAN ZUYLEN/WILLUMSEN (1980) apply the method of Lagrange multipliers. Equation B-29 and all constraints are transformed into a single Lagrange function

$$L = \sum_a \sum_{ij} f_{ij} \cdot p_{ij}^a \cdot \ln\left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right) + \sum_a \lambda_a \cdot \left(\sum_{ij} f_{ij} \cdot p_{ij}^a - q_a\right) \qquad \text{(B-30)}$$

where the λ_a are the Lagrange multipliers for each constraint, i.e. each link *a* with count. Equation B-30 has to be differentiated with respect to all f_{ij}, followed by further transformations (cf. VAN ZUYLEN/WILLUMSEN, 1980, WANG, 2008, ROHDE, 2006).

$$\begin{aligned} \frac{dL}{df_{ij}} &= \sum_a p_{ij}^a \cdot \ln\left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right) + \sum_a f_{ij} \cdot p_{ij}^a \cdot \frac{1}{f_{ij}} + \sum_a \lambda_a \cdot p_{ij}^a \\ &= \sum_a p_{ij}^a \cdot \ln\left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right) + \sum_a p_{ij}^a \cdot (1 + \lambda_a) = 0 \end{aligned} \qquad \text{(B-31)}$$

$$\sum_a p_{ij}^a \cdot \ln\left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right) = -\sum_a p_{ij}^a \cdot (1 + \lambda_a) \qquad \text{(B-32)}$$

$$\sum_a \ln\left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right)^{p_{ij}^a} = -\sum_a p_{ij}^a \cdot (1 + \lambda_a) \qquad \text{(B-33)}$$

$$\ln \prod_a \left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right)^{p_{ij}^a} = -\sum_a p_{ij}^a \cdot (1 + \lambda_a) \qquad \text{(B-34)}$$

$$\prod_a \left(\frac{f_{ij} \cdot q_a^0}{q_a \cdot f_{ij}^0}\right)^{p_{ij}^a} = e^{-\sum_a p_{ij}^a \cdot (1 + \lambda_a)} \qquad \text{(B-35)}$$

$$\prod_a \left(\frac{f_{ij}}{f_{ij}^0}\right)^{p_{ij}^a} \cdot \left(\frac{q_a^0}{q_a}\right)^{p_{ij}^a} = e^{-\sum_a p_{ij}^a \cdot (1+\lambda_a)} \qquad \text{(B-36)}$$

$$\left(\frac{f_{ij}}{f_{ij}^0}\right)^{\sum_a p_{ij}^a} = \left[\prod_a \left(\frac{q_a}{q_a^0}\right)^{p_{ij}^a}\right] \cdot e^{-\sum_a p_{ij}^a \cdot (1+\lambda_a)} \qquad \text{(B-37)}$$

$$\left(\frac{f_{ij}}{f_{ij}^0}\right)^{\sum_a p_{ij}^a} = \prod_a \left[\left(\frac{q_a}{q_a^0}\right)^{p_{ij}^a} \cdot e^{-p_{ij}^a \cdot (1+\lambda_a)}\right] \qquad \text{(B-38)}$$

$$\left(\frac{f_{ij}}{f_{ij}^0}\right)^{\sum_a p_{ij}^a} = \prod_a \left(\frac{q_a}{q_a^0} \cdot e^{-(1+\lambda_a)}\right)^{p_{ij}^a} \qquad \text{(B-39)}$$

$$\frac{f_{ij}}{f_{ij}^0} = \prod_a \left(\frac{q_a}{q_a^0} \cdot e^{-(1+\lambda_a)}\right)^{\frac{p_{ij}^a}{\sum_a p_{ij}^a}} \qquad \text{(B-40)}$$

$$f_{ij} = f_{ij}^0 \cdot \prod_a \left(\frac{q_a}{q_a^0} \cdot e^{-(1+\lambda_a)}\right)^{\frac{p_{ij}^a}{\sum_a p_{ij}^a}} \qquad \text{(B-41)}$$

With

$$g_{ij} = \sum_a p_{ij}^a \qquad \text{(B-42)}$$

and

$$X_a = \frac{q_a}{q_a^0} \cdot e^{-(1+\lambda_a)} \qquad \text{(B-43)}$$

follows

$$f_{ij} = f_{ij}^0 \cdot \prod_a X_a^{\frac{p_{ij}^a}{g_{ij}}} \qquad \text{(B-44)}$$

This is equation 5-6 given in paragraph 5.3.3. In the same paragraph an iterative algorithm has been presented that finds a solution $\{f_{ij}\}$ based on equation B-44.

B.3 Discussion

Equation B-29 implies that the information on individual links *a* that are used as constraints contributes to the overall information independently from other links *a* used as constraints because the overall information is just the sum of the information on individual links *a*. This is the reason why the IM model is sensitive to redundant information on different links *a* which are

partly or entirely linearly dependent on other links. VAN ZUYLEN (1981) described this effect and modified the original model so that it is no longer susceptible to redundant information. WANG (2008) uses the original model and defined several rules to eliminate redundant constraints before starting the matrix estimation.

HELLINGA (1994) investigated the influence of the use of Stirling's approximation (equation B-25) on the quality of demand estimation. This equation provides reliable approximated values if n > 100. These values can be further improved by using a more precise form of the approximation. Any number of summands from the following equation can be used.

$$\ln(n!) = n \cdot \ln(n) - n + \frac{1}{2} \cdot \ln(2\pi n) + \frac{1}{12n} - \frac{1}{360n^3} + \ldots \qquad \text{(B-45)}$$

Figure B-5 shows the relative error of the approximation for the cases that two and three summands are used. The upper curve corresponds to the deviations that arise in the IM model. For n > 48 the relative error is less than 2 percent in the case of two summands. HELLINGA (1994) therefore argues that the imprecision of Stirling's approximation has no significant influence if flows f_{ij} are greater than this value. Furthermore he points out that the derivative of the objective function is of interest for minimization, so that the derivative of Stirling's approximation has to be examined as well. For the case of two and three summands these derivatives are

$$\frac{d}{dn}\left(n \cdot \ln(n) - n\right) = n \cdot \frac{1}{n} + \ln(n) - 1 = \ln(n) \qquad \text{(B-46)}$$

and

$$\frac{d}{dn}\left(n \cdot \ln(n) - n + \frac{1}{2} \cdot \ln(2\pi n)\right) = \ln(n) + \frac{1}{2n}. \qquad \text{(B-47)}$$

The difference between both derivatives is $1/(2n)$ and decreases reciprocally with increasing n or increasing f_{ij} respectively. This implies that relations ij with low flows f_{ij} will have a greater error of estimation. To overcome this problem, observed flows q_a can be converted to a longer period of time, e.g. from vehicles per 15 minutes to vehicles per hour, before starting the estimation. Low flows will then be higher and inaccuracy of the approximation is reduced. After the OD matrix estimation all flows f_{ij} can be transformed back to the original time period.

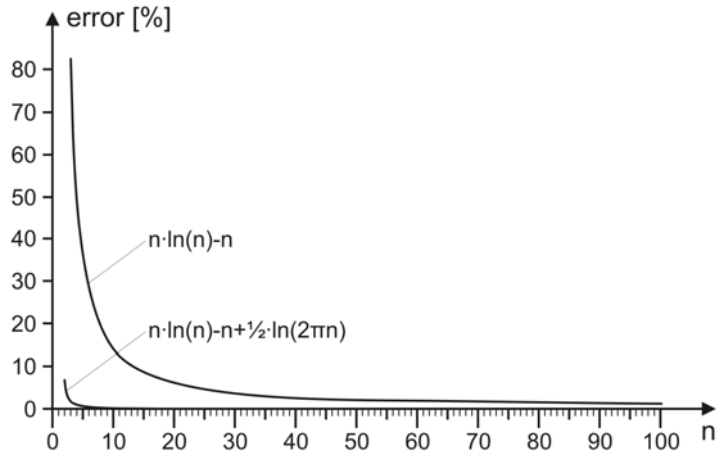


Figure B-5: Relative error of Stirling's approximation

C Elimination of inconsistent constraints (Intermediate steps of derivation)

C.1 Method of Lagrange multipliers

Equation C-1 is to be maximized subject to equations C-2 for each set *i* of nodes.

$$P\left(\{\hat{q}_a\}\right) = \prod_a \frac{q_a^{\hat{q}_a}}{\hat{q}_a!} \cdot e^{-q_a} \qquad \text{(C-1)}$$

$$\sum_a A_{ai} \cdot q_a = 0 \qquad \text{(C-2)}$$

where:

q_a expected number of vehicles in a time interval on link *a* [veh/h]

$\hat{q}_a$ observed number of vehicles in a time interval on link *a* [veh/h]

A_{ai} coefficient relating links *a* to sets *i* of nodes [-]

This sort of problem can be solved by using the method of Lagrange multipliers. In order to facilitate the following transformations, VAN ZUYLEN/BRANSTON (1982) maximize the natural logarithm of equation C-1 instead of equation C-1 itself, i.e.:

$$\begin{aligned} \ln P\left(\{\hat{q}_a\}\right) &= \ln \prod_a \frac{q_a^{\hat{q}_a}}{\hat{q}_a!} \cdot e^{-q_a} = \sum_a \ln \frac{q_a^{\hat{q}_a}}{\hat{q}_a!} \cdot e^{-q_a} = \sum_a \left(\ln e^{-q_a} + \ln q_a^{\hat{q}_a} - \ln \hat{q}_a!\right) \\ &= \sum_a \left(-q_a + \hat{q}_a \cdot \ln q_a - \ln \hat{q}_a!\right) \end{aligned} \qquad \text{(C-3)}$$

This function and all constraints are transformed into a single Lagrange function:

$$L = \sum_a \left(-q_a + \hat{q}_a \cdot \ln q_a - \ln \hat{q}_a!\right) + \sum_i \lambda_i \cdot \sum_a A_{ai} \cdot q_a \qquad \text{(C-4)}$$

where:

λ_i Lagrange multiplier of set *i* of nodes or constraint *i* respectively [-]

The Lagrange function has to be differentiated with respect to all q_a.

$$\frac{dL}{dq_a} = -1 + \hat{q}_a \cdot \frac{1}{q_a} + \sum_i \lambda_i \cdot A_{ai} \qquad \text{(C-5)}$$

To find the maximum of the function, the derivative has to be set to zero, followed by some transformations.

$$-1 + \hat{q}_a \cdot \frac{1}{q_a} + \sum_i \lambda_i \cdot A_{ai} = 0 \qquad \text{(C-6)}$$

$$\frac{\hat{q}_a}{q_a} = 1 - \sum_i \lambda_i \cdot A_{ai} \qquad \text{(C-7)}$$

$$q_a = \hat{q}_a \Big/ \left(1 - \sum_i \lambda_i \cdot A_{ai} \right) \qquad \text{(C-8)}$$

Equation C-8 is equal to equation 5-10 except for the algebraic sign in front of the sigma sign. It is not evident why VAN ZUYLEN/BRANSTON (1982) came to a different sign. However, using a positive sign in equation C-8 will only cause all Lagrange multipliers λ_i to have a positive instead of a negative sign and vice versa. The final solution $\{q_a\}$ will remain unchanged.

C.2 Transformation into Taylor series

Equation C-9 has to be transformed into a Taylor series about the estimates $\tilde{\lambda}_i$ of the Lagrange multipliers λ_i.

$$q_a = \hat{q}_a \Big/ \left(1 + \sum_i \lambda_i \cdot A_{ai} \right) \qquad \text{(C-9)}$$

The Taylor series of a function $f(x_1,..., x_i,..., x_n)$ at a point $\{\tilde{x}_1,...,\tilde{x}_i,...,\tilde{x}_n\}$ is given by equation C-10 for the case that second order terms and above are neglected.

$$\begin{aligned} P_f(x_1,...,x_i,...,x_n) = f(\tilde{x}_1,...,\tilde{x}_i,...,\tilde{x}_n) + (x_1 - \tilde{x}_1) \cdot f_{x_1}(\tilde{x}_1,...,\tilde{x}_i,...,\tilde{x}_n) + ... \\ + (x_i - \tilde{x}_i) \cdot f_{x_i}(\tilde{x}_1,...,\tilde{x}_i,...,\tilde{x}_n) + ... \\ + (x_n - \tilde{x}_n) \cdot f_{x_n}(\tilde{x}_1,...,\tilde{x}_i,...,\tilde{x}_n) \end{aligned} \qquad \text{(C-10)}$$

Applying this scheme to equation C-9 gives:

$$\begin{aligned} q_a &\simeq \hat{q}_a \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right) + \sum_i (\lambda_i - \tilde{\lambda}_i) \cdot \frac{d}{d\lambda_i} \hat{q}_a \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right) \\ &= \hat{q}_a \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right) + \sum_i (\lambda_i - \tilde{\lambda}_i) \cdot (-\hat{q}_a \cdot A_{ai}) \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right)^2 \\ &= \hat{q}_a \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right) - \hat{q}_a \cdot \sum_i (\lambda_i - \tilde{\lambda}_i) \cdot A_{ai} \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right)^2 \end{aligned} \qquad \text{(C-11)}$$

With $\delta_i = \lambda_i - \tilde{\lambda}_i$ equation C-11 becomes

$$q_a \simeq \hat{q}_a \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right) - \hat{q}_a \cdot \sum_i \delta_i \cdot A_{ai} \Big/ \left(1 + \sum_i \tilde{\lambda}_i \cdot A_{ai} \right)^2 \qquad \text{(C-12)}$$

which is equal to equation 5-12.

C.3 Transformation into a linear system of equations

Substitution of equation C-12 in equation C-2 leads to the following steps of transformation.

$$0 = \sum_a A_{ai} \cdot q_a$$

$$= \sum_a A_{ai} \cdot \left[\hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right) - \hat{q}_a \cdot \sum_j \delta_j \cdot A_{aj} \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right)^2 \right] \qquad \text{(C-13)}$$

$$\sum_a A_{ai} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right) = \sum_a A_{ai} \cdot \hat{q}_a \cdot \sum_j \delta_j \cdot A_{aj} \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right)^2 \qquad \text{(C-14)}$$

$$\sum_a A_{ai} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right) = \sum_a \sum_j \delta_j \cdot A_{ai} \cdot A_{aj} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right)^2 \qquad \text{(C-15)}$$

$$\sum_a A_{ai} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right) = \sum_j \delta_j \cdot \sum_a A_{ai} \cdot A_{aj} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right)^2 \qquad \text{(C-16)}$$

$$E_i = \sum_j B_{ij} \cdot \delta_j \qquad \text{(C-17)}$$

where:

$$E_i = \sum_a A_{ai} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right) \qquad \text{(C-18)}$$

$$B_{ij} = \sum_a A_{ai} \cdot A_{aj} \cdot \hat{q}_a \Big/ \left(1 + \sum_k \tilde{\lambda}_k \cdot A_{ak} \right)^2 \qquad \text{(C-19)}$$

Equation C-17 is equal to equation 5-13. One such equation per each set *i* of nodes can be formed.

D Detailed comparison of perfomance indicators

D.1 Introduction

This appendix contains diagrams which compare average travel times and number of stops that resulted from multiple AIMSUN simulation runs. Each diagram compares two cases. The compared cases comprise the original fixed time signal plans, the optimized T7F fixed time signal plans, and traffic control by means of the new ATCS prototype with a variety of different settings and varying accuracy of the input traffic demand used for optimization. The compared cases displayed in each diagram are described in the captions of the figures.

The first two diagrams per comparison illustrate the network-wide performance in terms of average overall travel times [s/veh·km] and number of stops [-/veh·km]. The temporal resolution of the diagrams is 15 minutes which is the duration of the optimization interval. Gray bars highlight intervals with no statistically significant difference between the two cases. A two-tailed t-test with a level of significance α of 5 percent has been used.

The next two diagrams per comparison compare average travel times [s/veh] and number of stops [-/veh] of all available routes in the network. They reveal detailed information on how changes in overall travel times and stops are distributed among the routes. Each row of the diagrams represents one time interval of 15 minutes, each column represents one route. The routes are sorted by traffic volumes. The heaviest loaded routes are on the left side and the less loaded routes on the right side. The upper bar labeled “traffic volumes” shows the extent of traffic load on the routes. The first of the three rows within the bar shows traffic volumes during the morning peak interval by different intensities of purple, the second row shows traffic volumes of the afternoon peak interval, and the third row shows the average volumes of both. The latter have been used to sort the routes.

Each of the two diagrams of route travel times and number of stops shows a pattern of green and red shaded and gray rectangles. The pattern contains one colored rectangle for each route and time interval. Green shaded rectangles indicate that the travel time or number of stops has improved from the first case of the comparison (the “blue line” case in the caption of the figure) to the second case (“red line” case in the caption). Red shaded rectangles show deterioration of travel times or number of stops. Gray rectangles correspond to only minor changes which are not statistically significant according to the t-test mentioned above. In both diagrams the four time intervals of the morning and afternoon peak hour are highlighted.

D.2 Hanover List network

The diagrams on the following pages show detailed results of the Hanover List network.

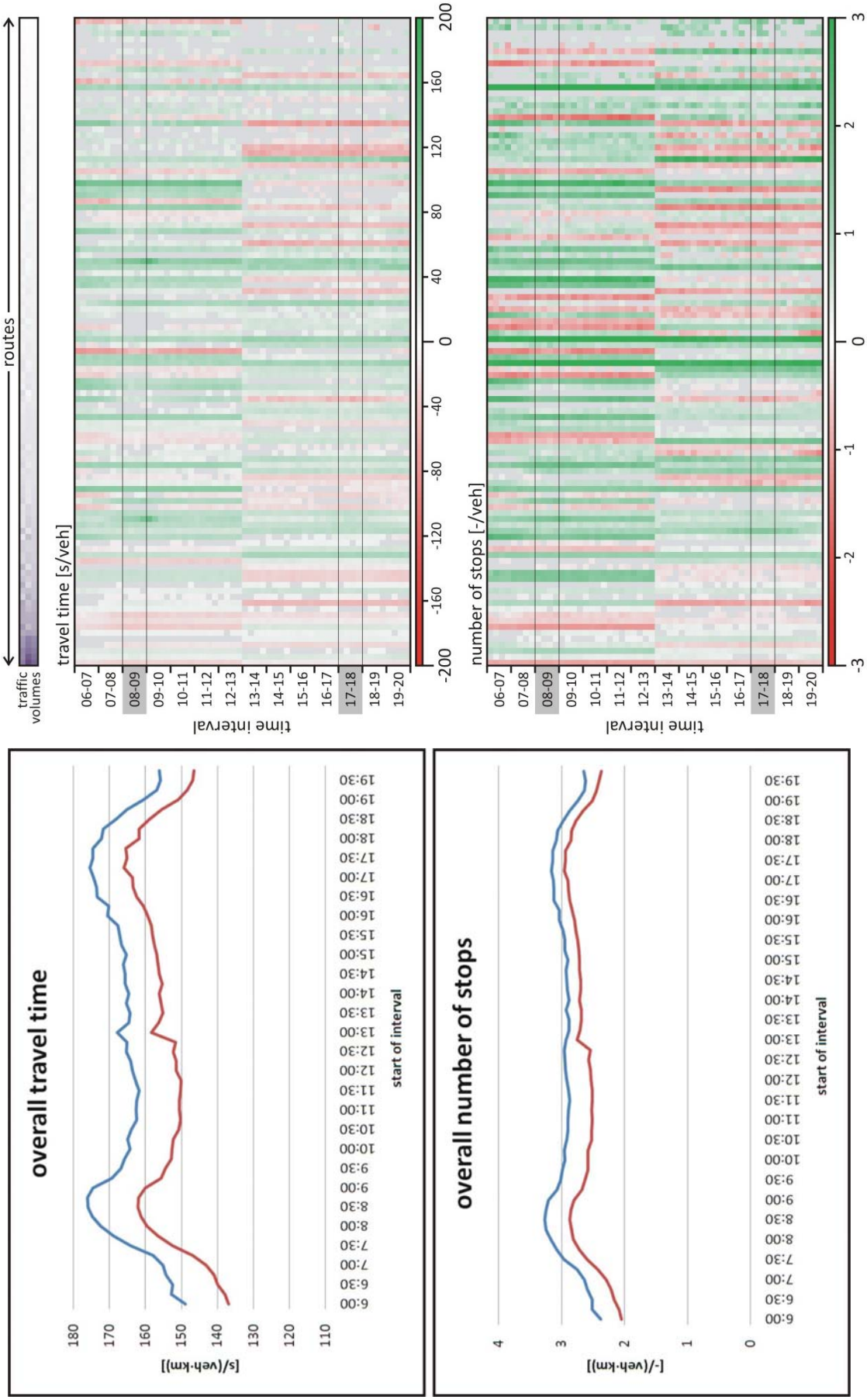

routes
traffic volumes
travel time [s/veh]
time interval
number of stops [-/veh]
overall travel time
[s/(veh·km)]
start of interval
overall number of stops
[-/(veh·km)]
original fixed time control
T7F fixed time control

Figure D-1:

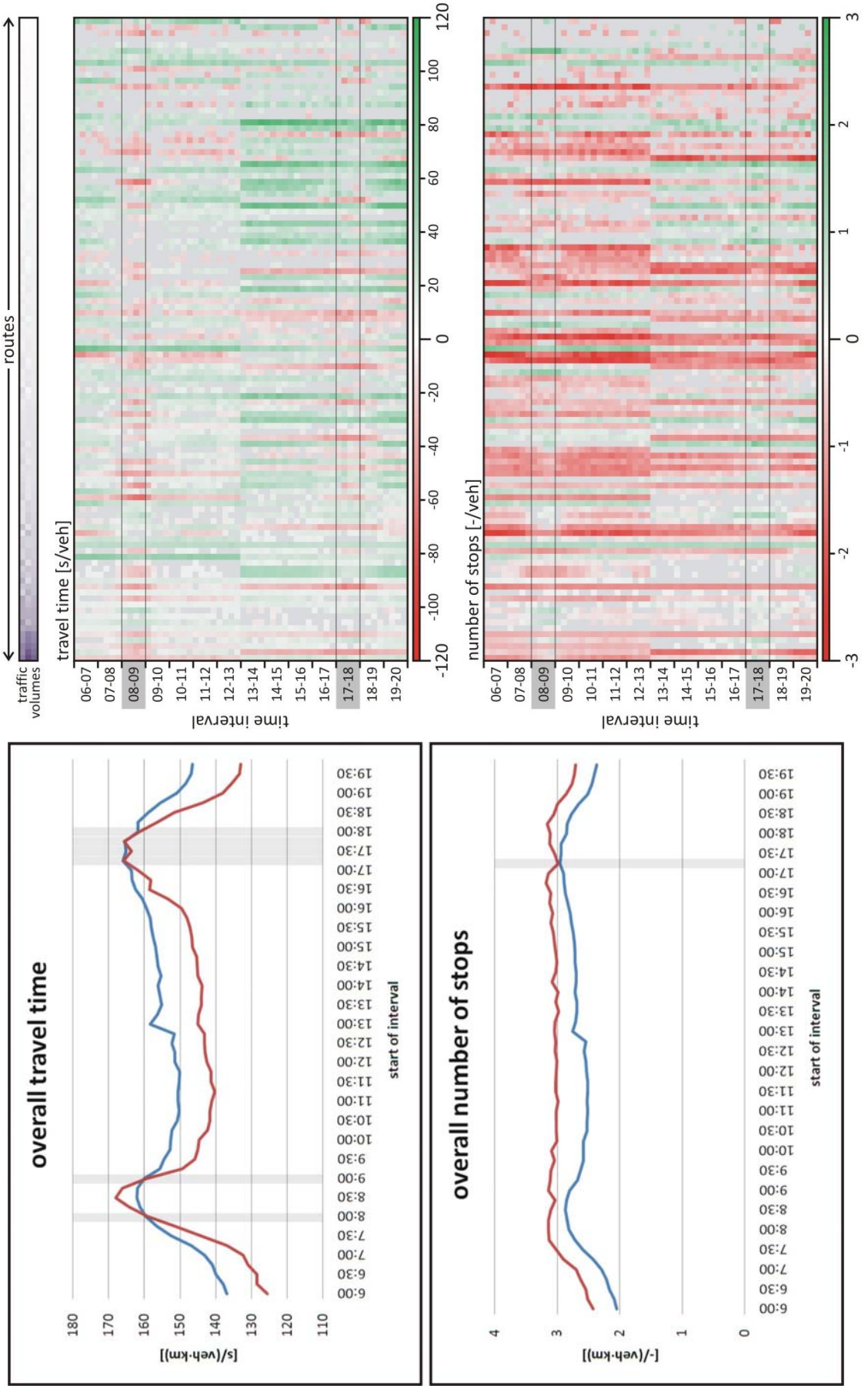


Figure D-2: T7F fixed time control
ATCS, saturation based cycle lengths, no offset optimization, exact demand

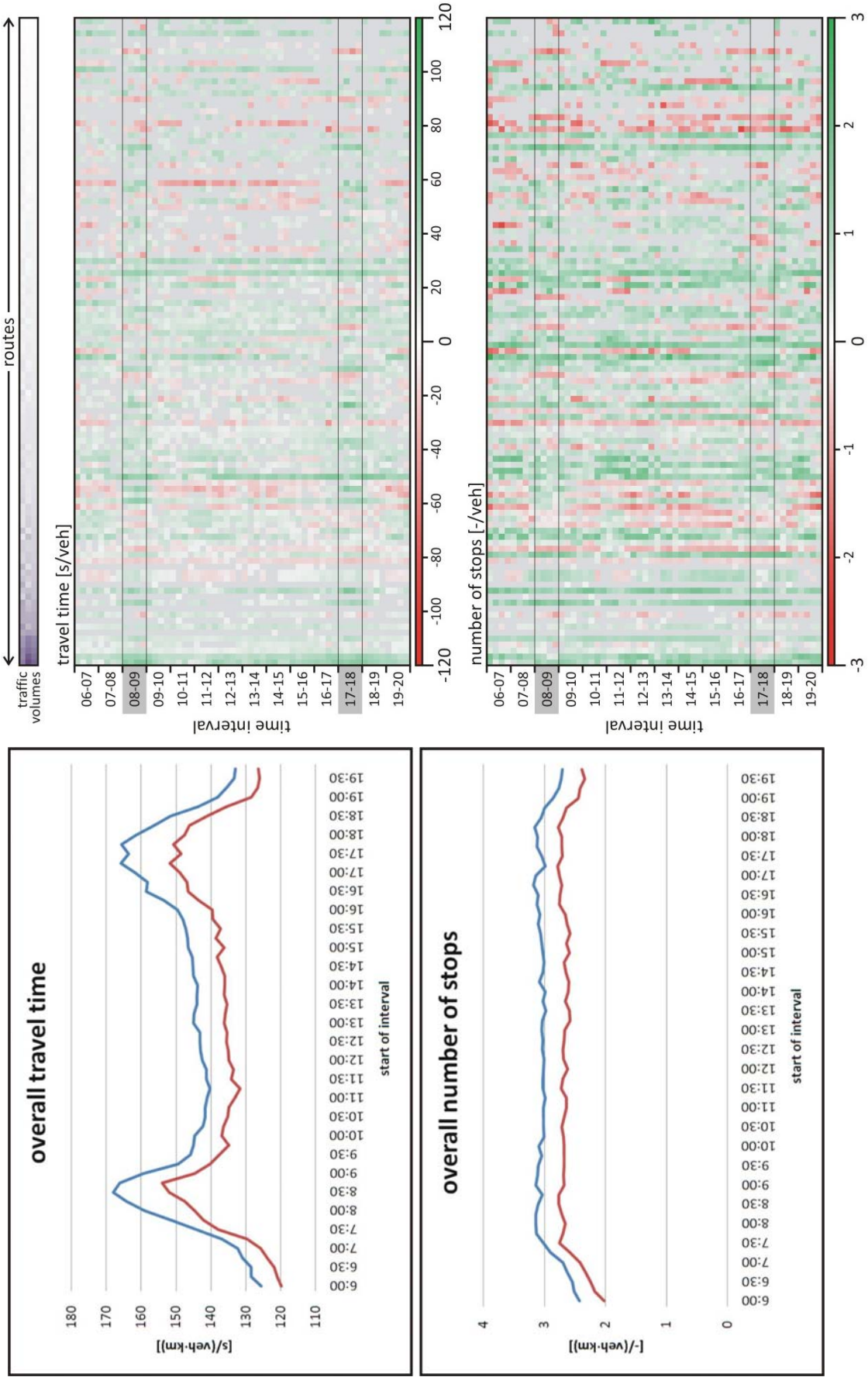


Figure D-3: ATCS, saturation based cycle lengths, no offset optimization, exact demand
ATCS, saturation based cycle lengths, PGA, exact demand

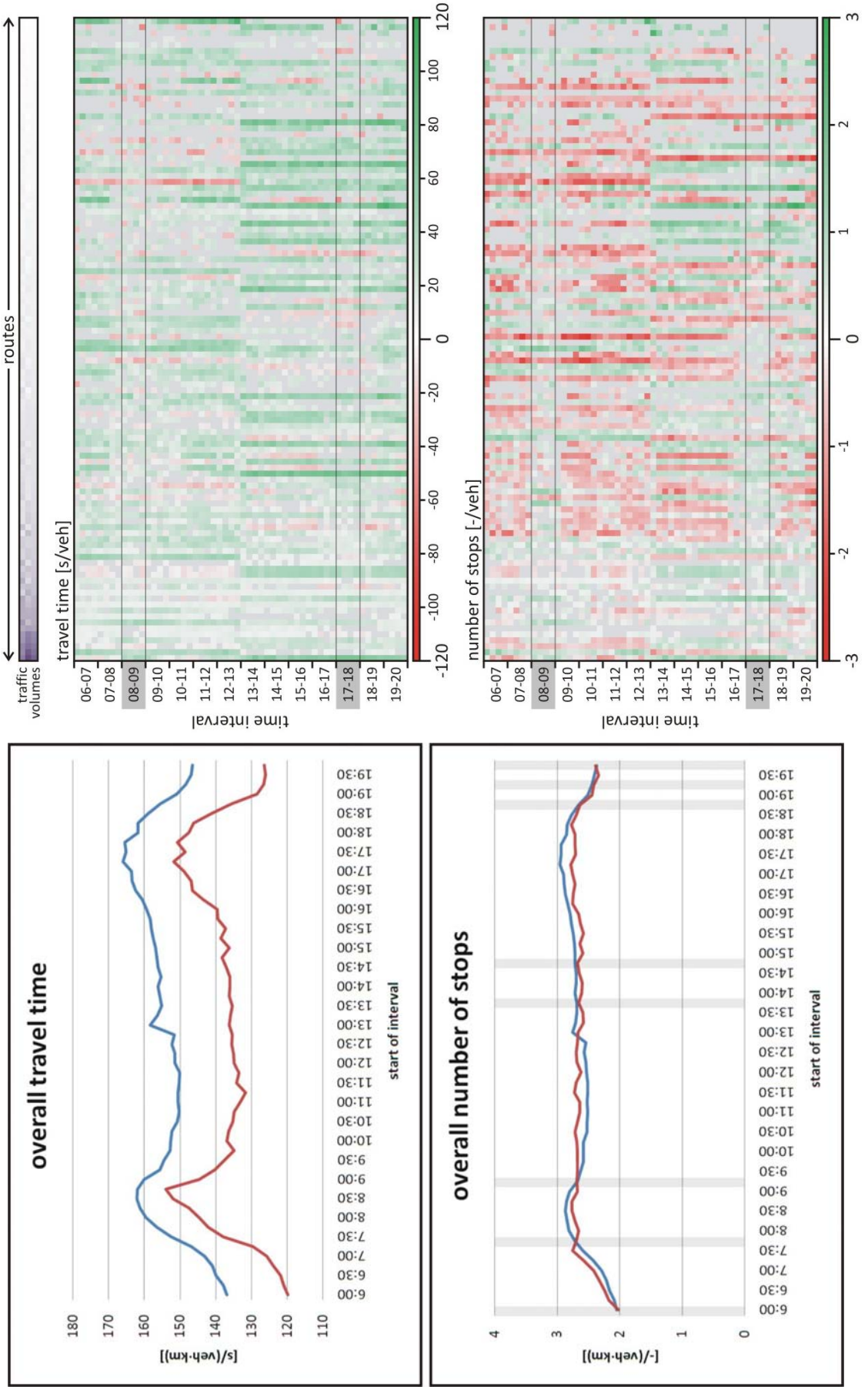


Figure D-4: T7F fixed time control
ATCS, saturation based cycle lengths, PGA, exact demand

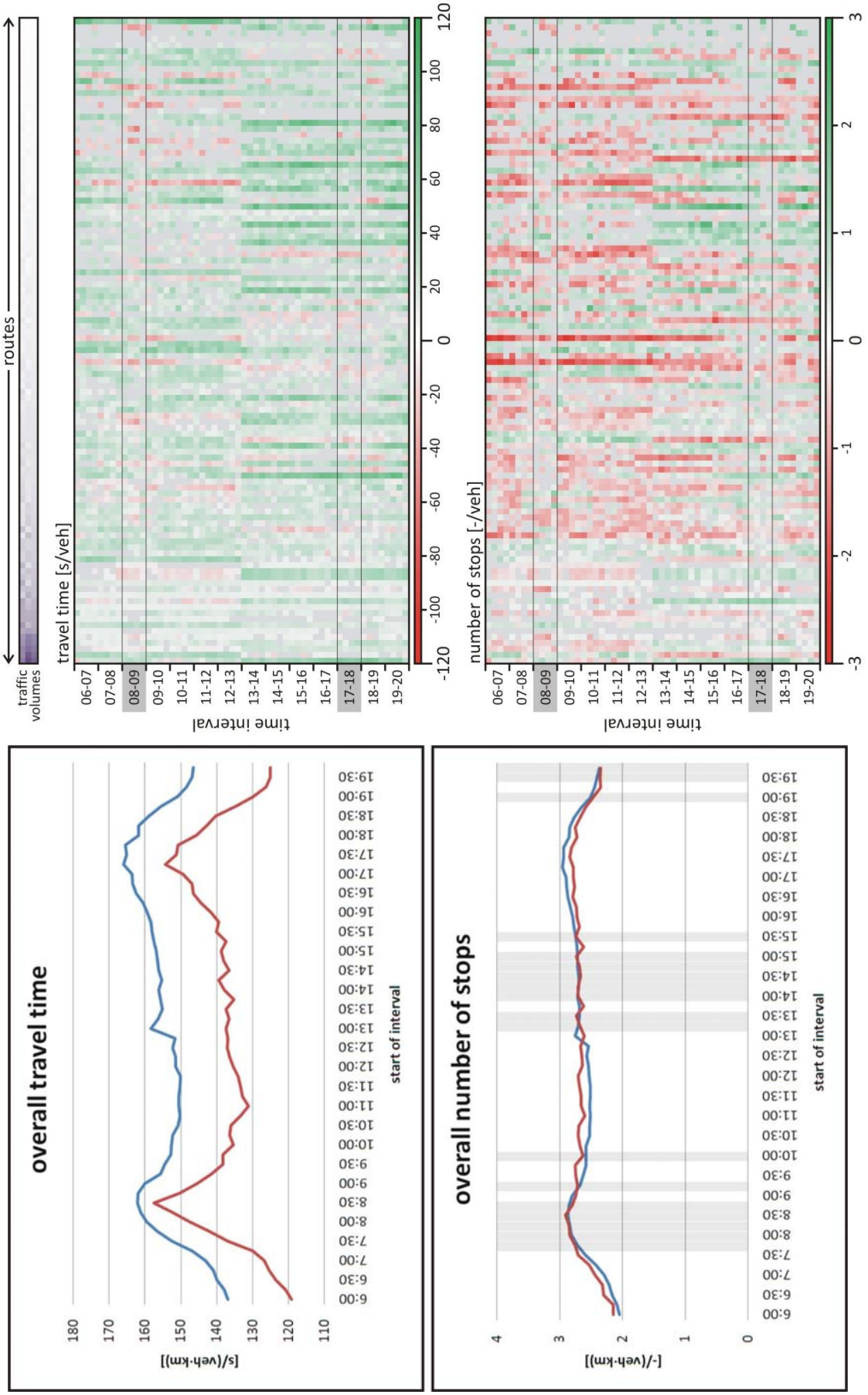

routes
traffic volumes
travel time [s/veh]
time interval
number of stops [-/veh]
overall travel time
[s/(veh·km)]
start of interval
overall number of stops
[-/(veh·km)]
T7F fixed time control
ATCS, saturation based cycle lengths, SGA, exact demand

Figure D-5:

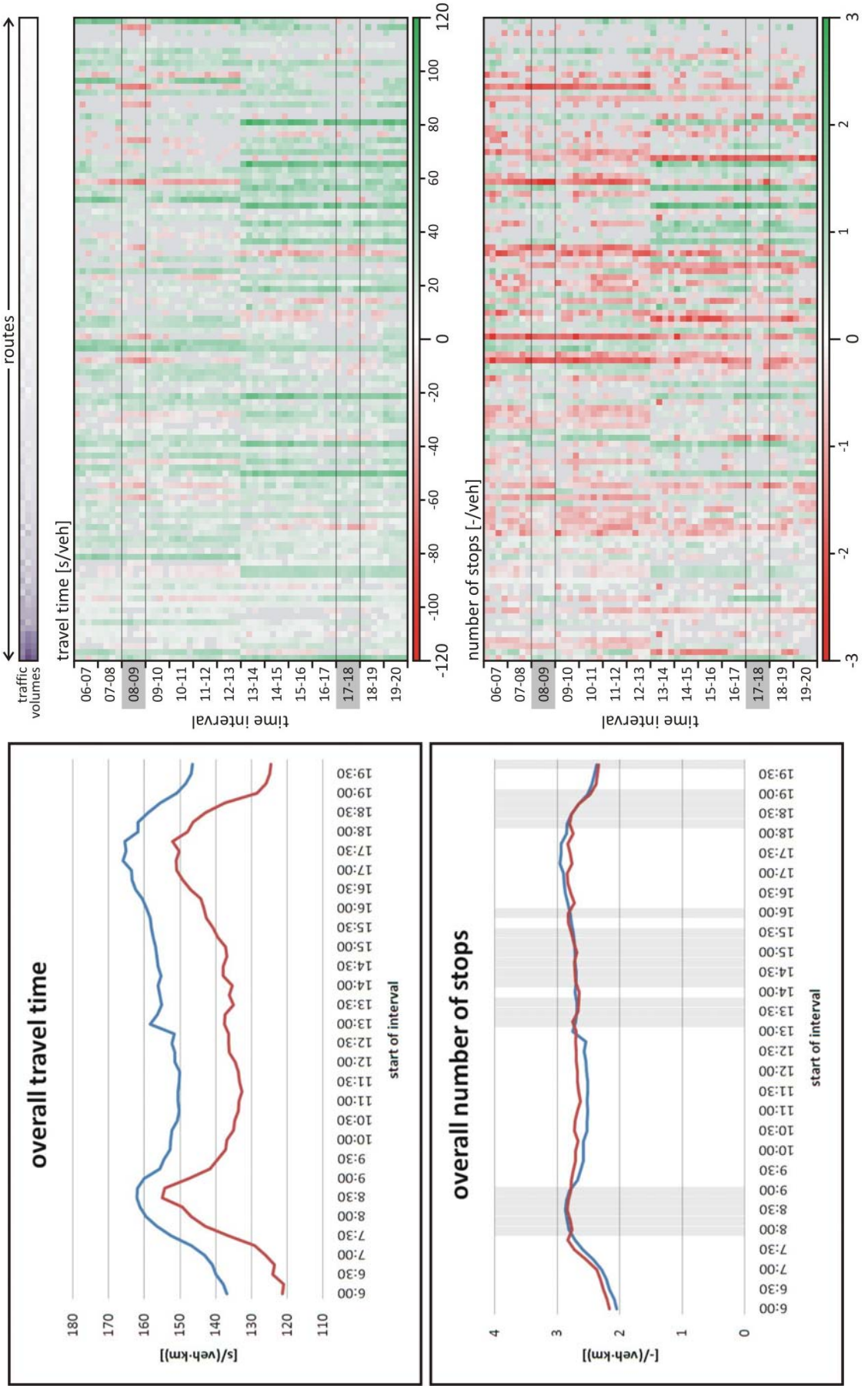


Figure D-6: T7F fixed time control
ATCS, saturation based cycle lengths, SE, exact demand

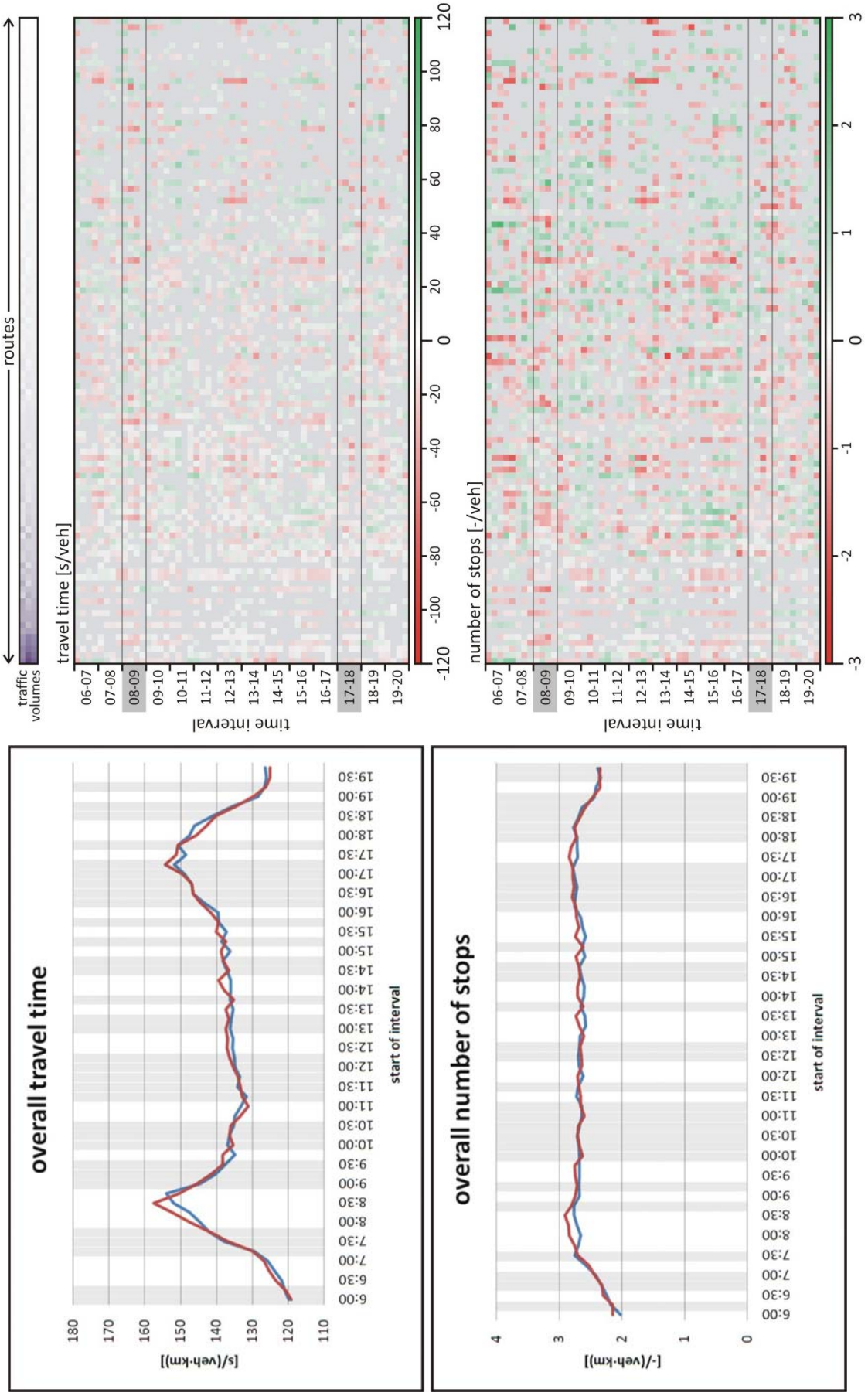


Figure D-7: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, SGA, exact demand

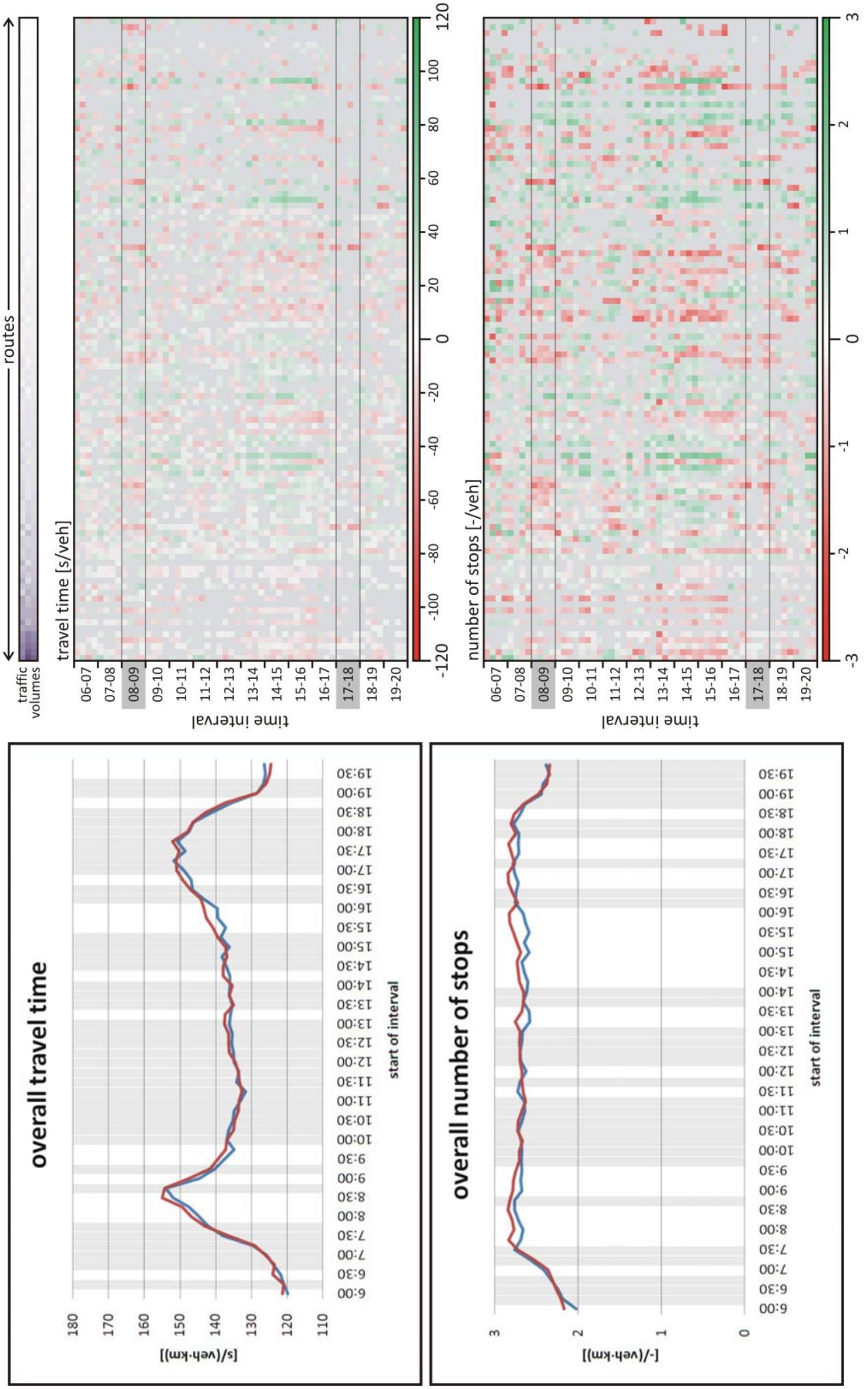


Figure D-8: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, SE, exact demand

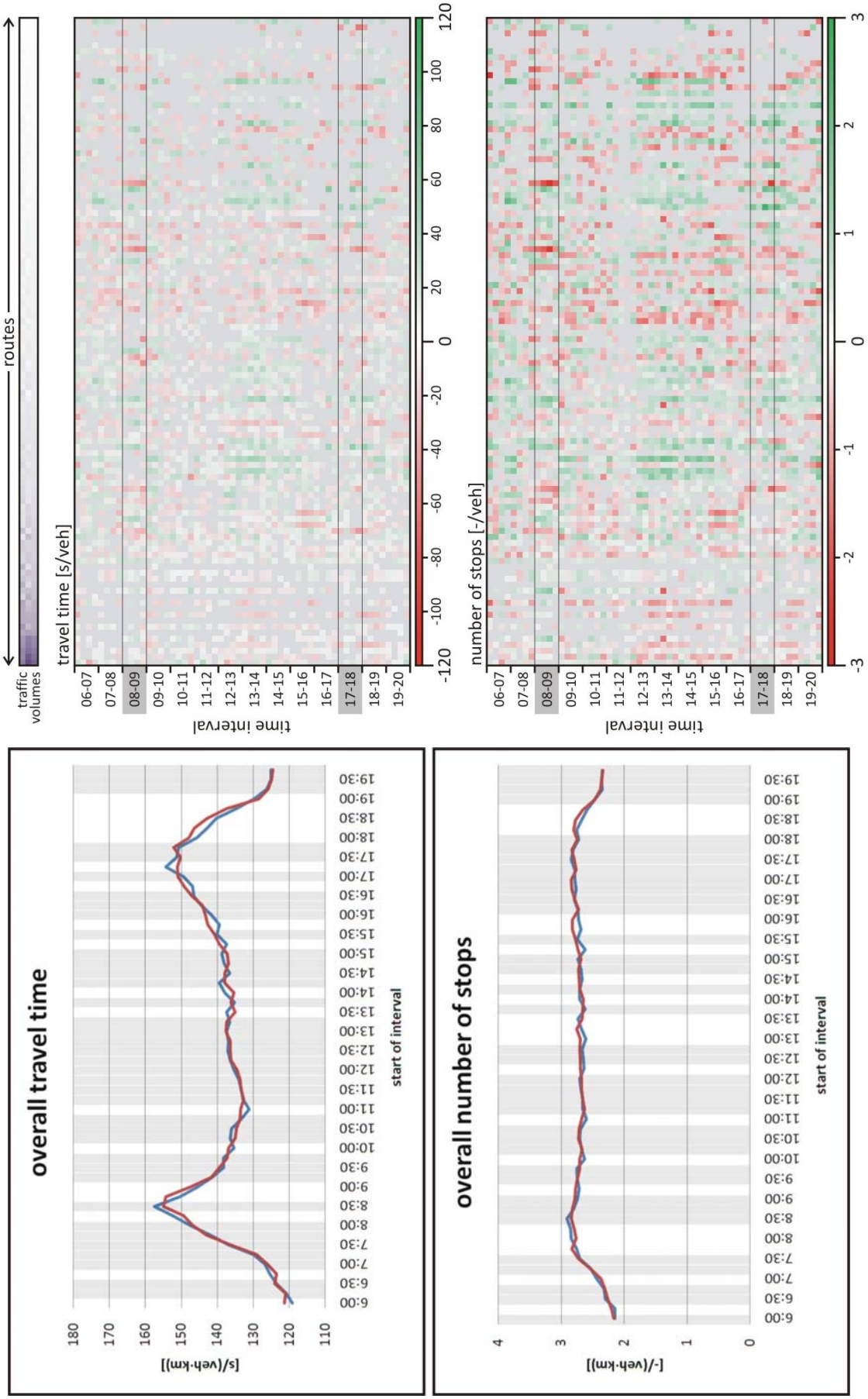

routes
traffic volumes
travel time [s/veh]
time interval
number of stops [-/veh]
overall travel time
[s/(veh·km)]
start of interval
overall number of stops
[-/(veh·km)]
ATCS, saturation based cycle lengths, SGA, exact demand
ATCS, saturation based cycle lengths, SE, exact demand

Figure D-9:

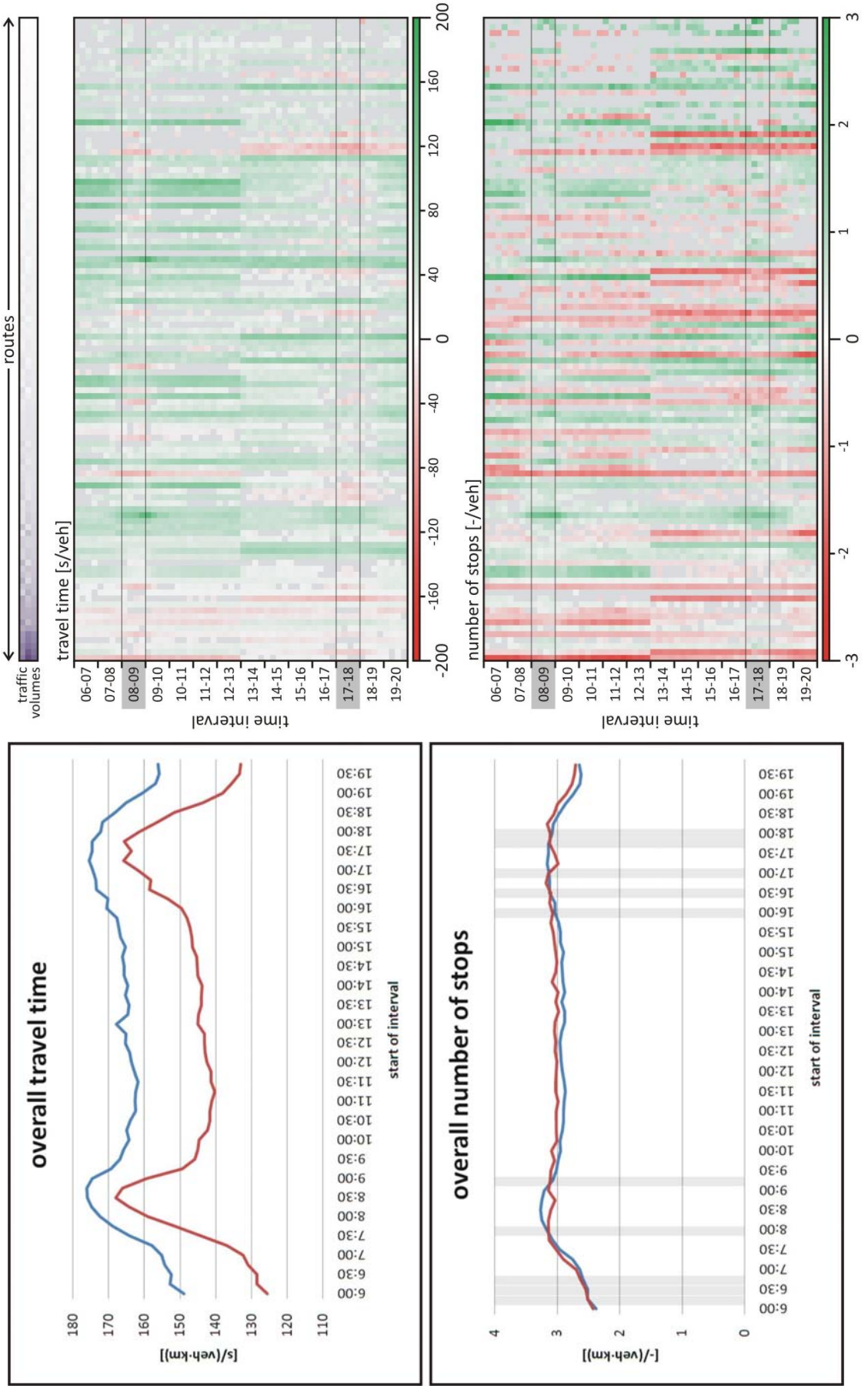


Figure D-10: original fixed time control
ATCS, saturation based cycle lengths, no offset optimization, exact demand

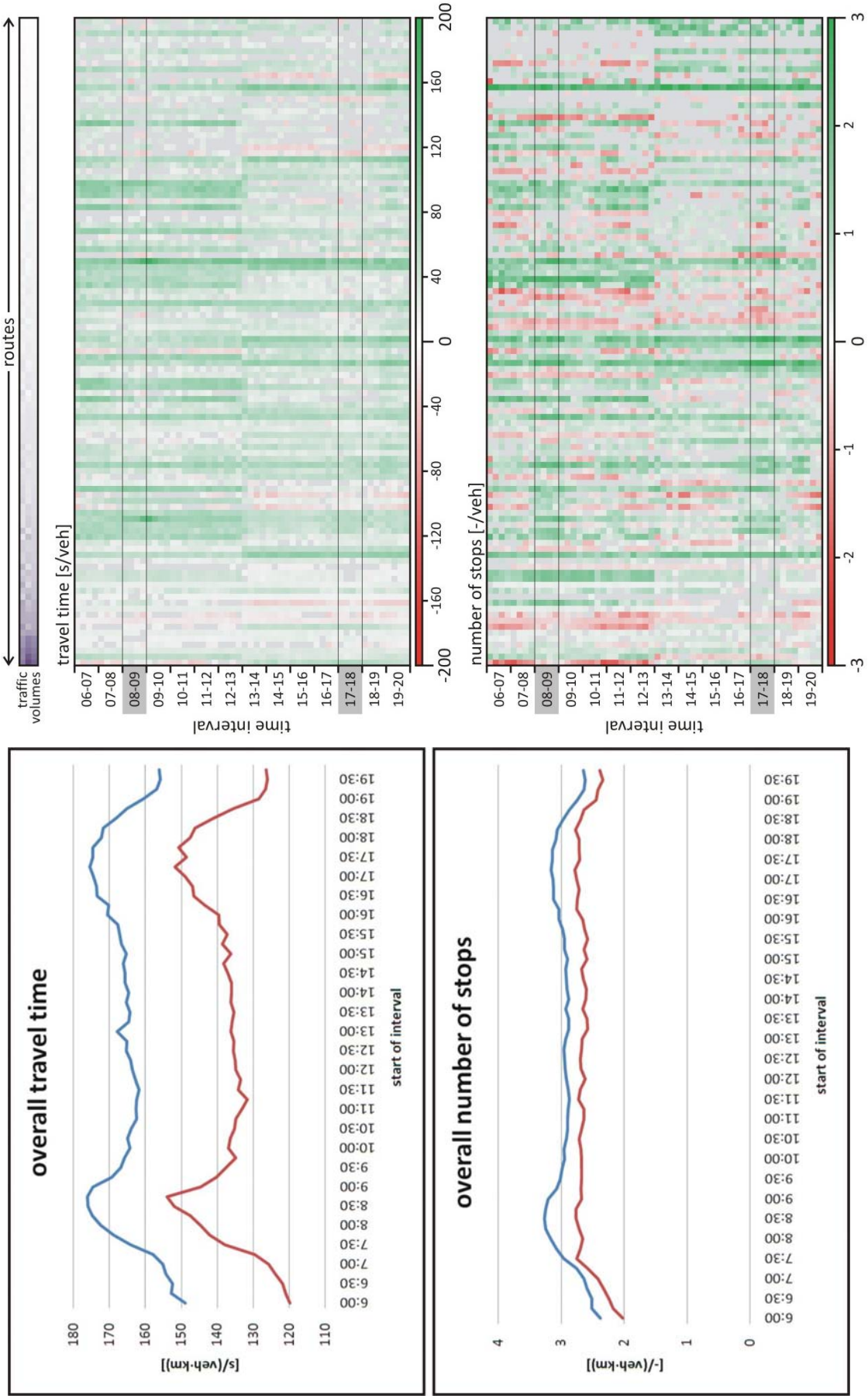


Figure D-11: original fixed time control
ATCS, saturation based cycle lengths, PGA, exact demand

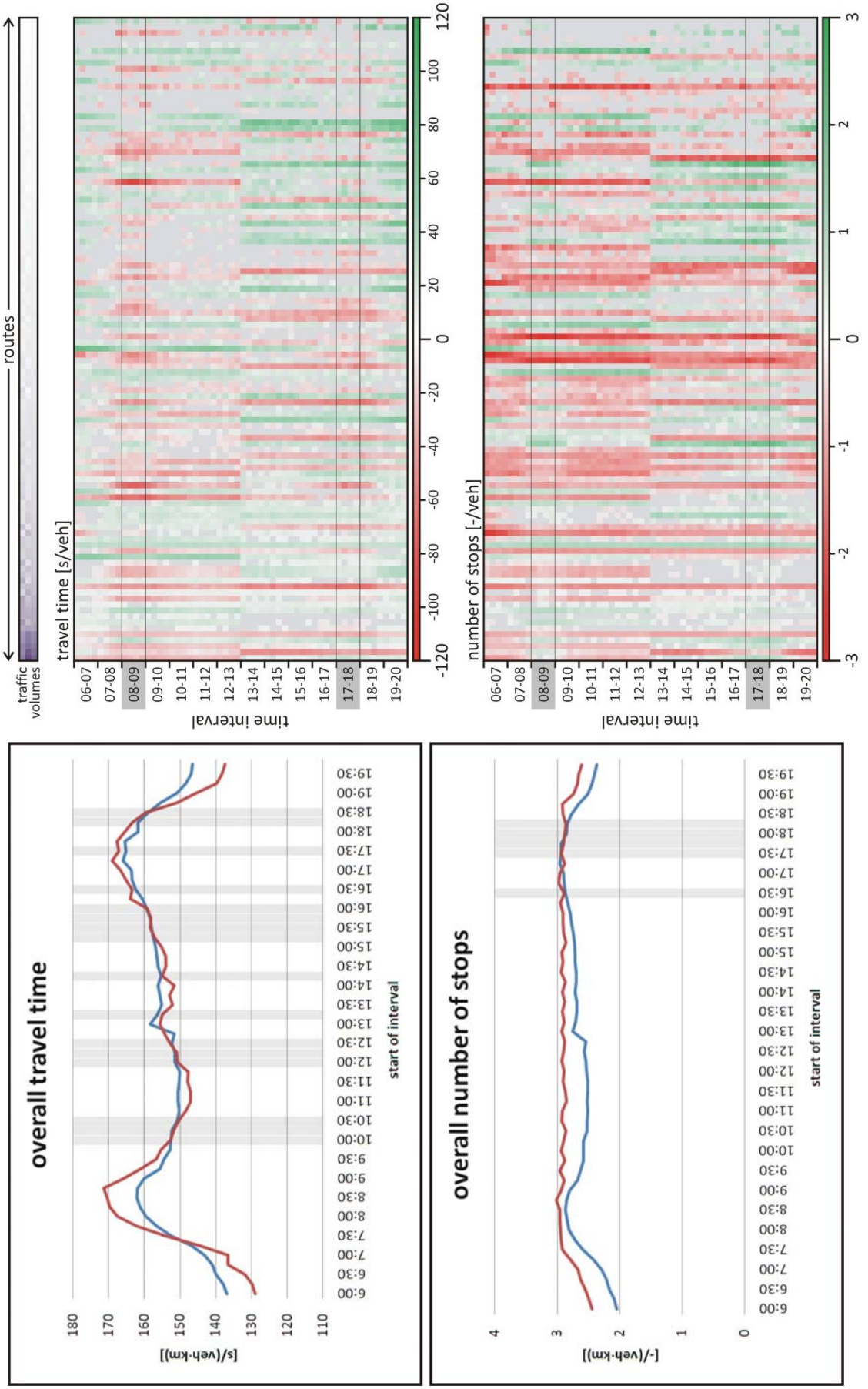


Figure D-12: T7F fixed time control
ATCS, Webster cycle lengths, no offset optimization, exact demand

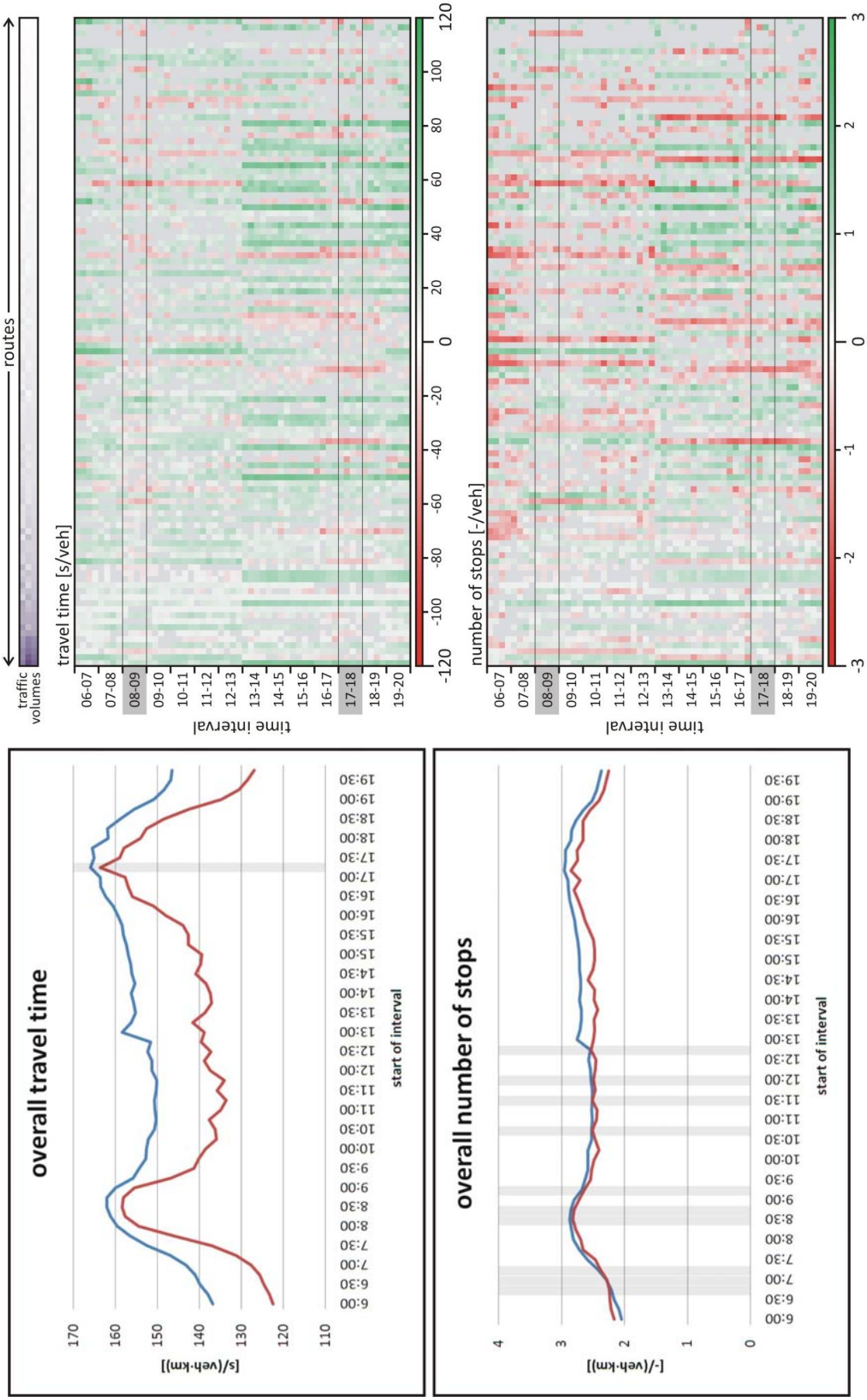


Figure D-13: T7F fixed time control
ATCS, Webster cycle lengths, PGA, exact demand

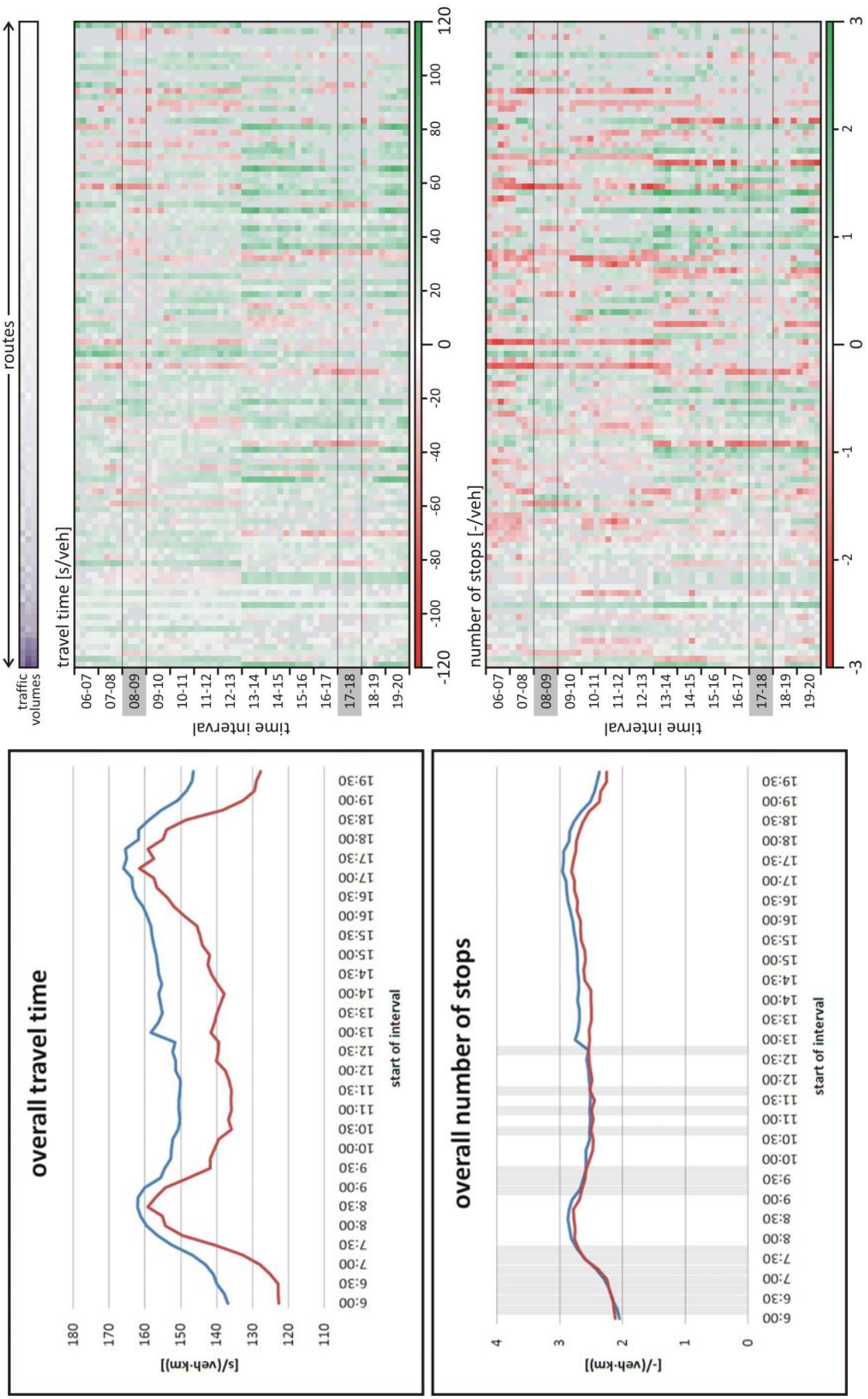


Figure D-14: T7F fixed time control
ATCS, Webster cycle lengths, SGA, exact demand

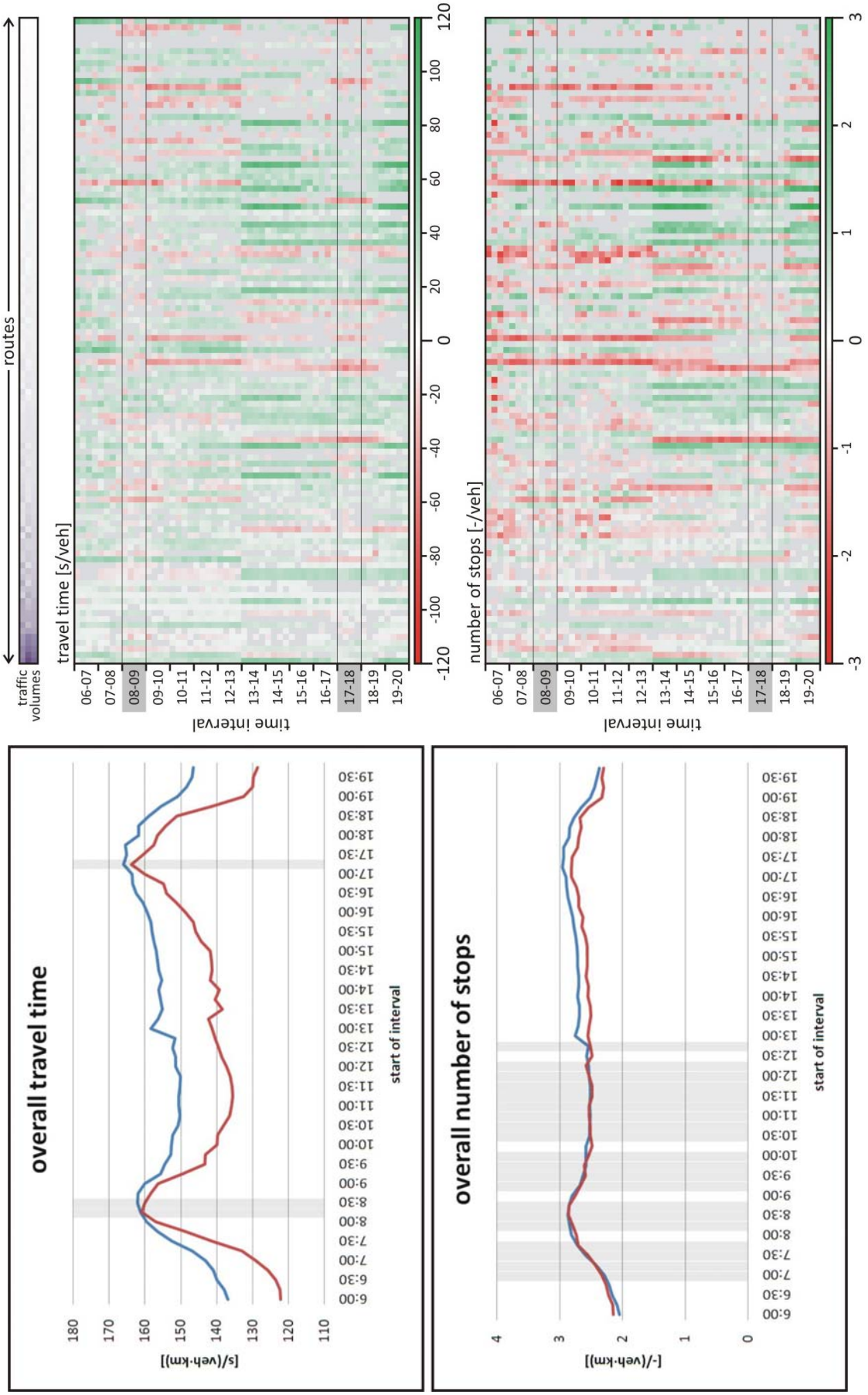


Figure D-15: T7F fixed time control
ATCS, Webster cycle lengths, SE, exact demand

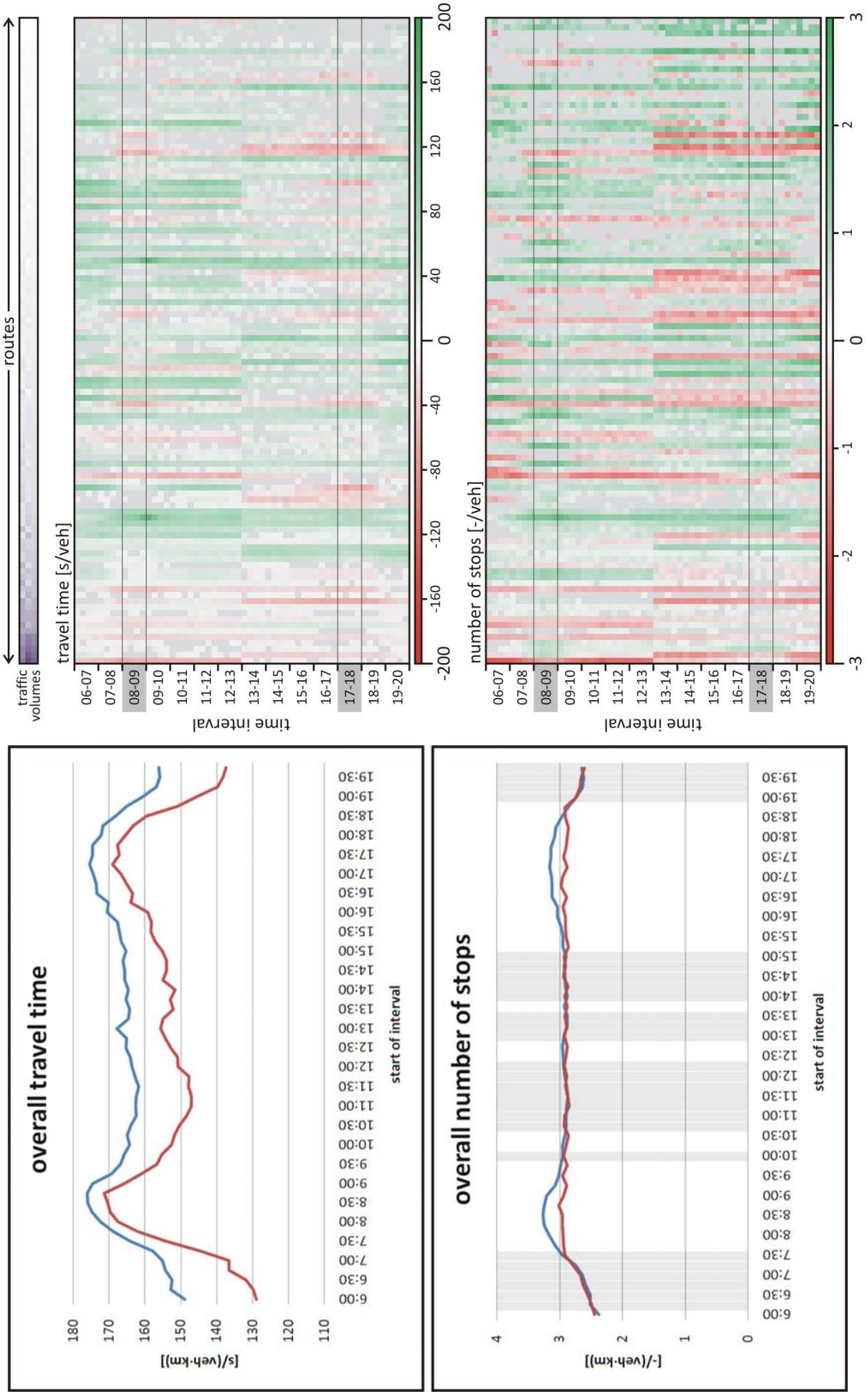


Figure D-16: original fixed time control
ATCS, Webster cycle lengths, no offset optimization, exact demand

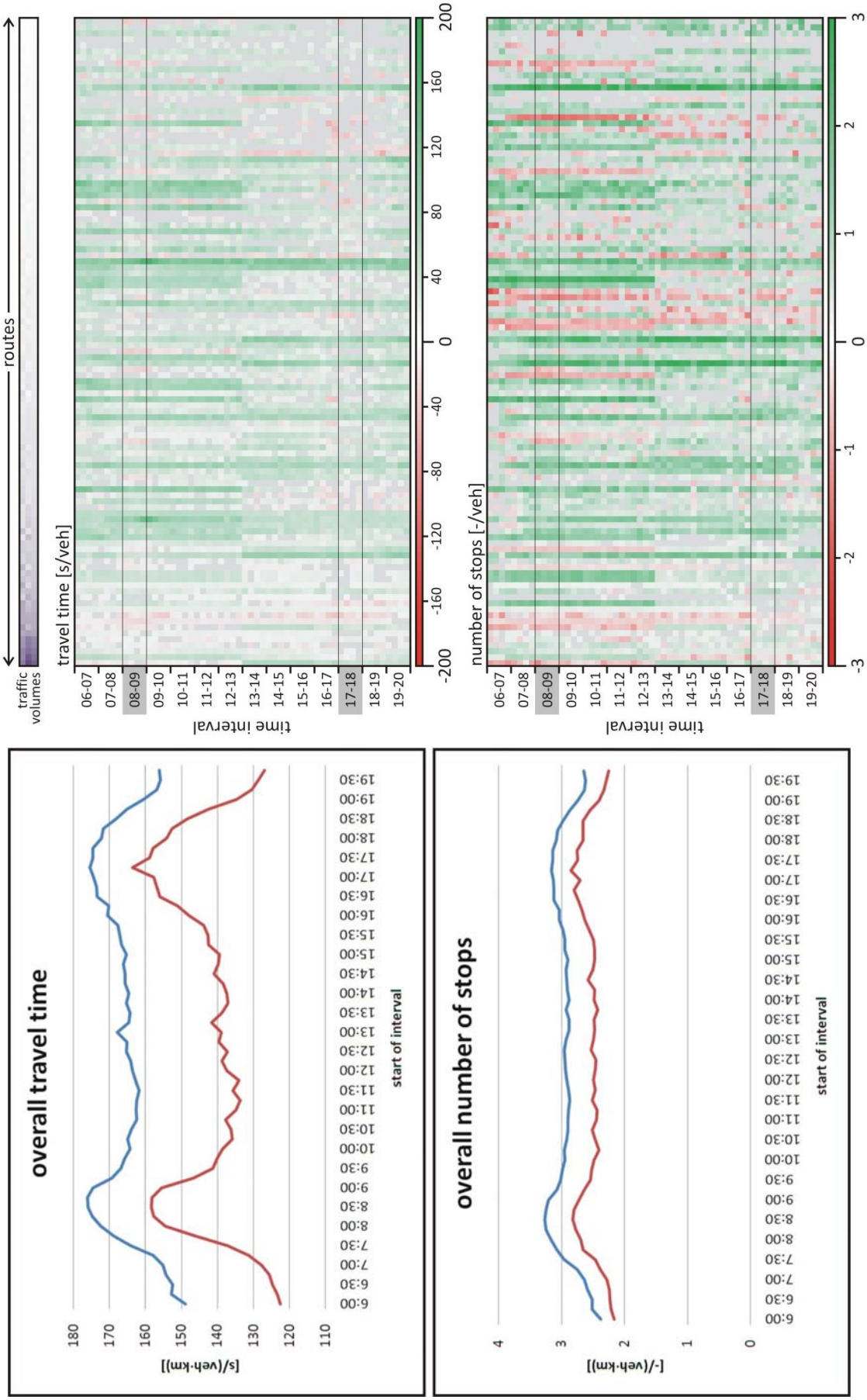


Figure D-17: — original fixed time control
— ATCS, Webster cycle lengths, PGA, exact demand

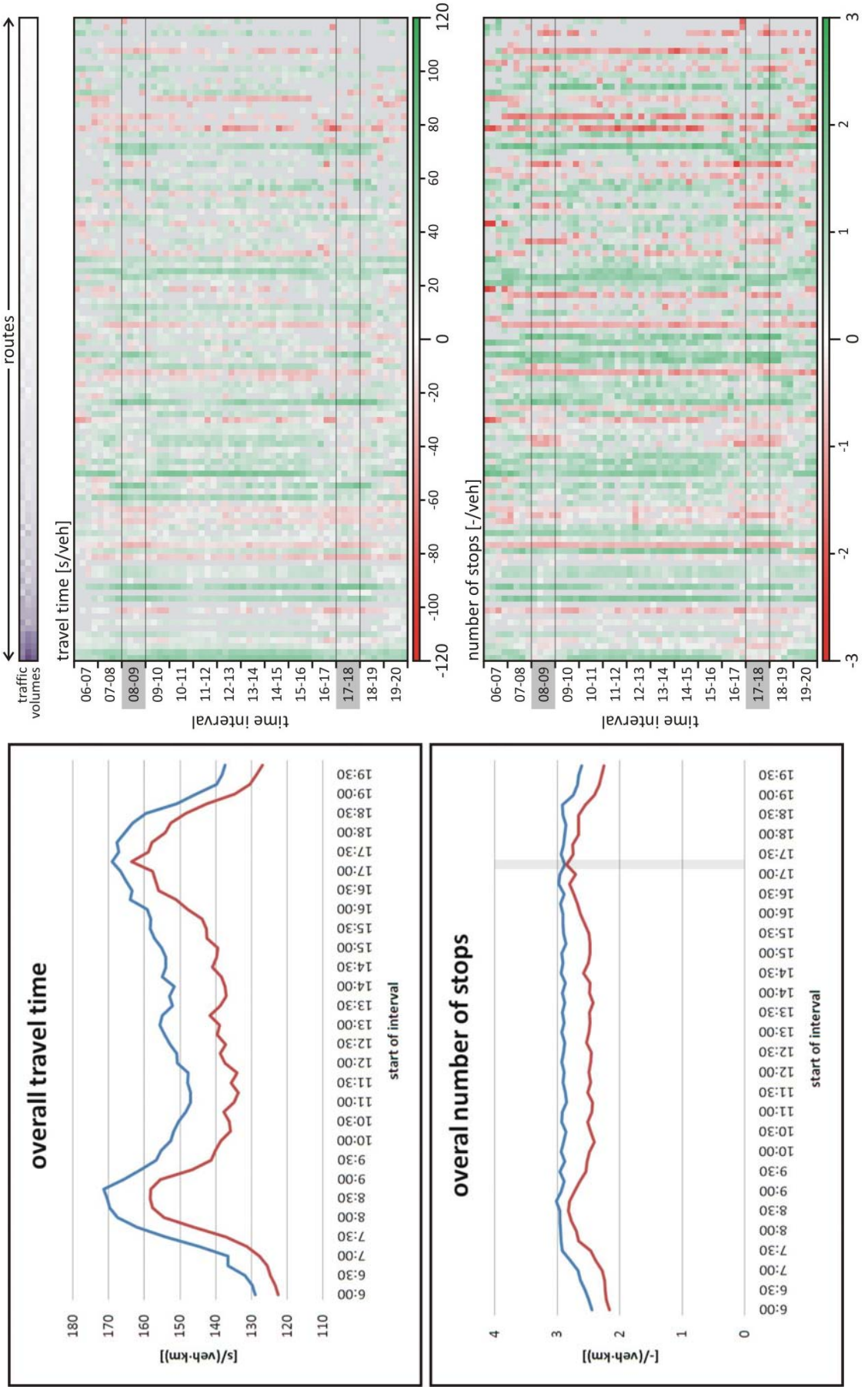


Figure D-18: ATCS, Webster cycle lengths, no offset optimization, exact demand
ATCS, Webster cycle lengths, PGA, exact demand

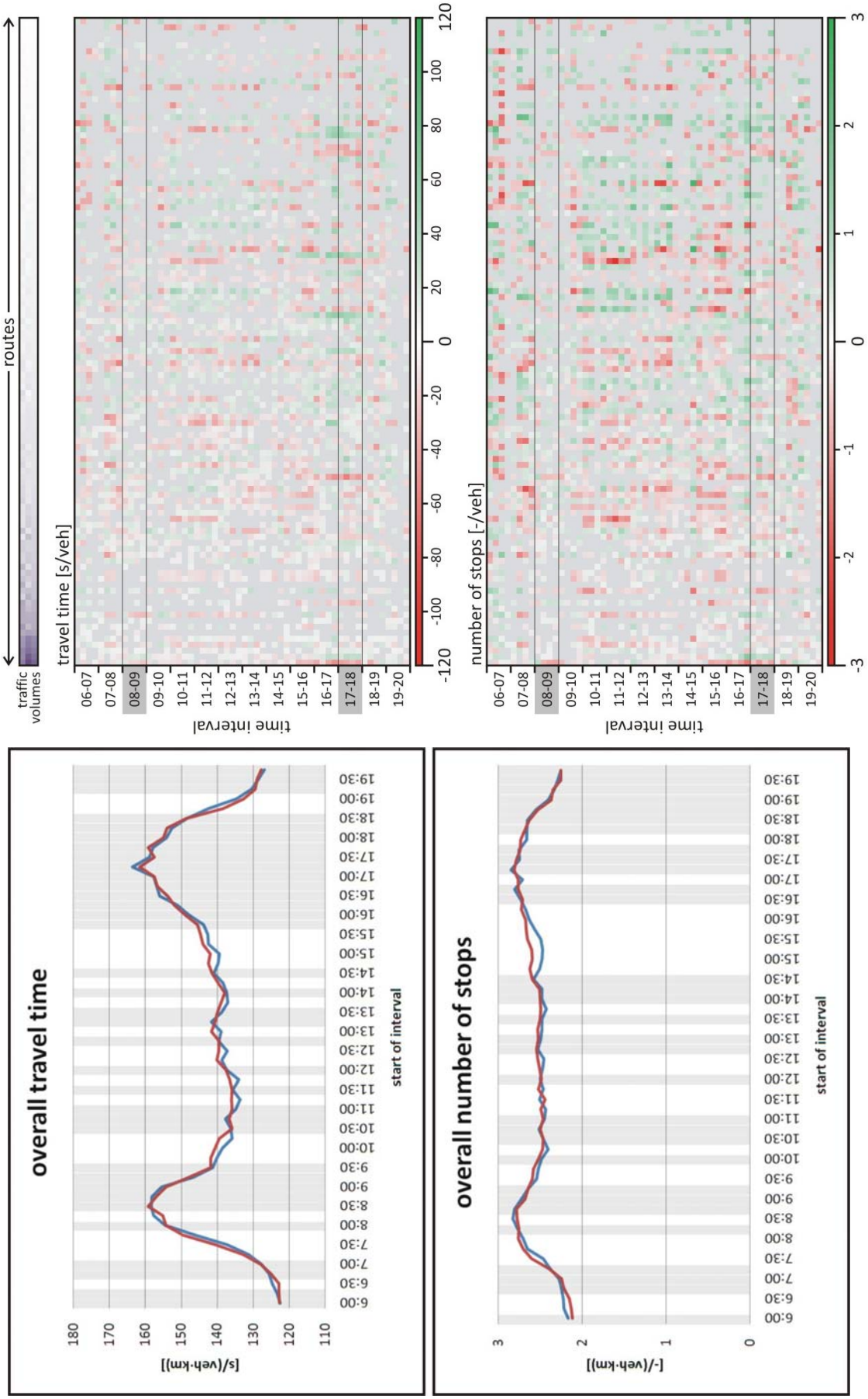


Figure D-19: ATCS, Webster cycle lengths, PGA, exact demand
ATCS, Webster cycle lengths, SGA, exact demand

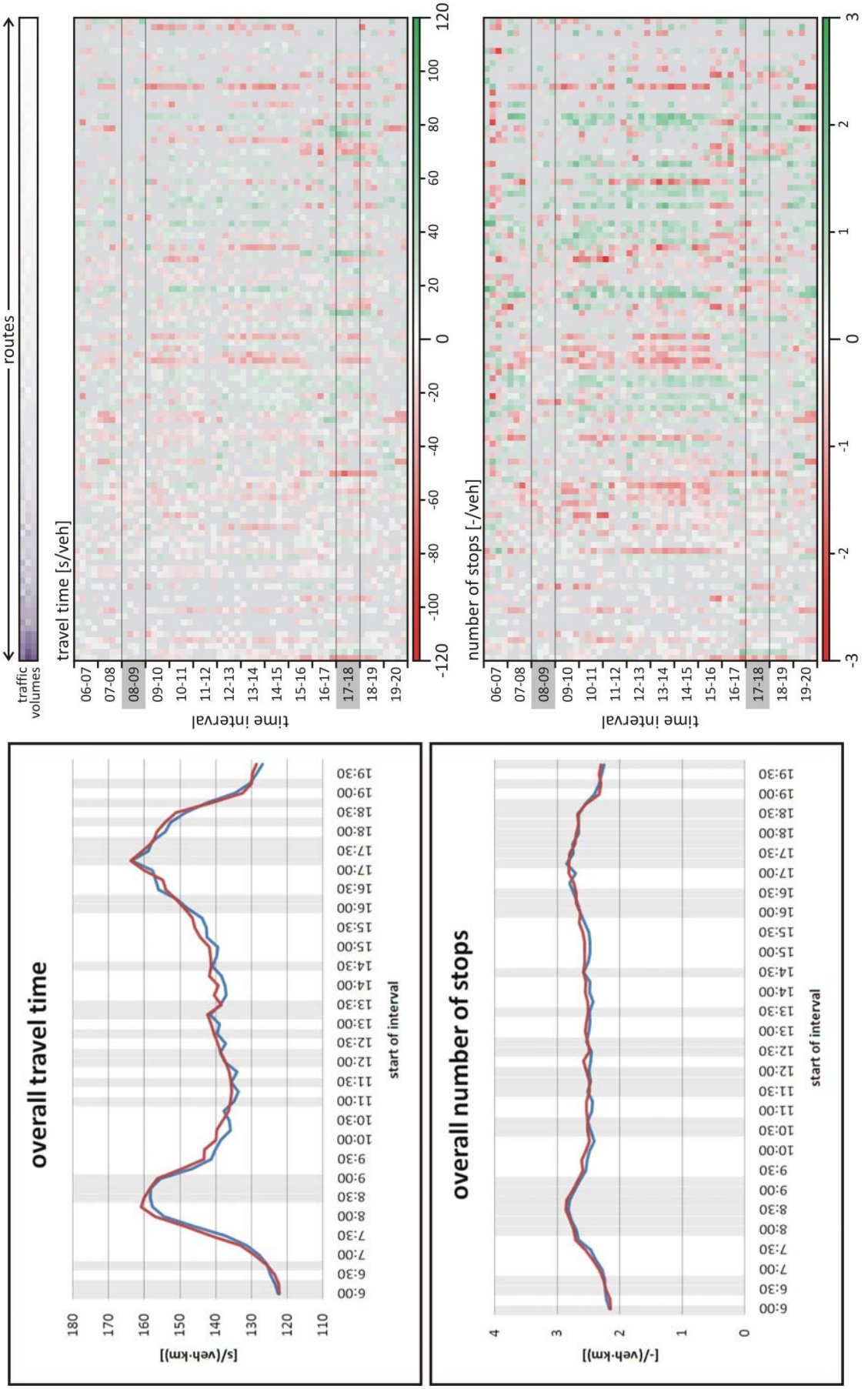


Figure D-20: ATCS, Webster cycle lengths, PGA, exact demand
ATCS, Webster cycle lengths, SE, exact demand

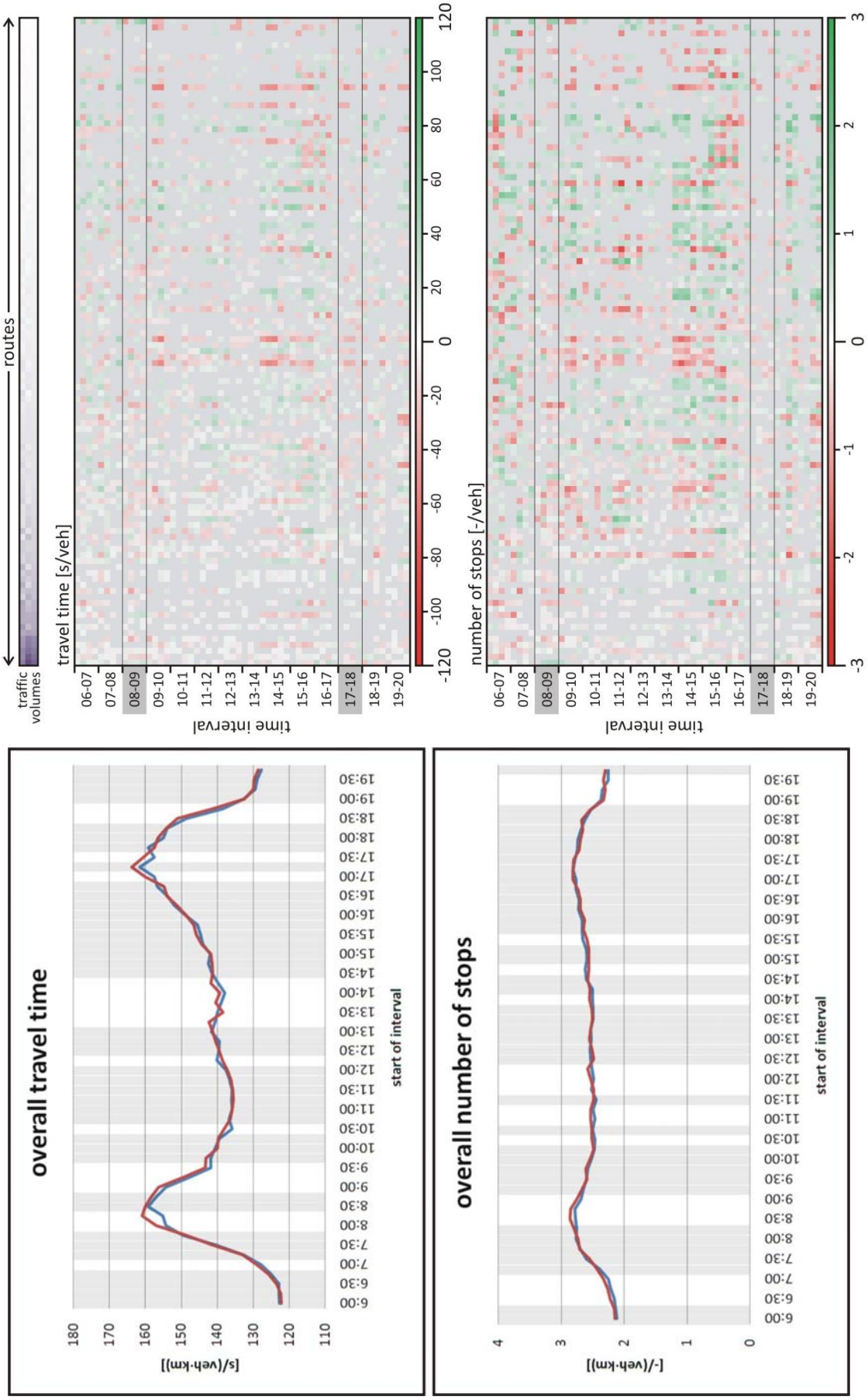


Figure D-21: ATCS, Webster cycle lengths, SGA, exact demand
ATCS, Webster cycle lengths, SE, exact demand

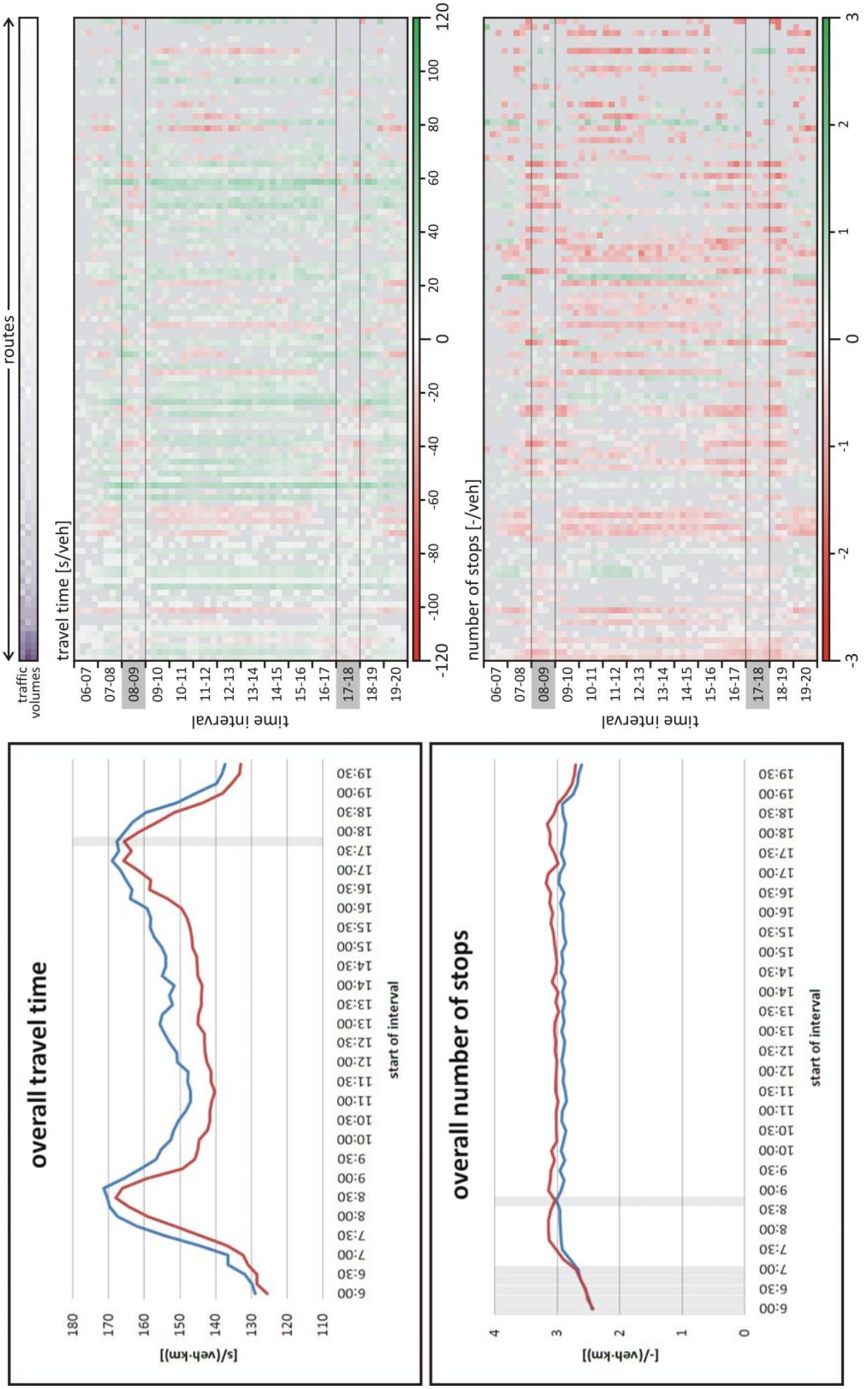


Figure D-22: ATCS, Webster cycle lengths, no offset optimization, exact demand
ATCS, saturation based cycle lengths, no offset optimization, exact demand

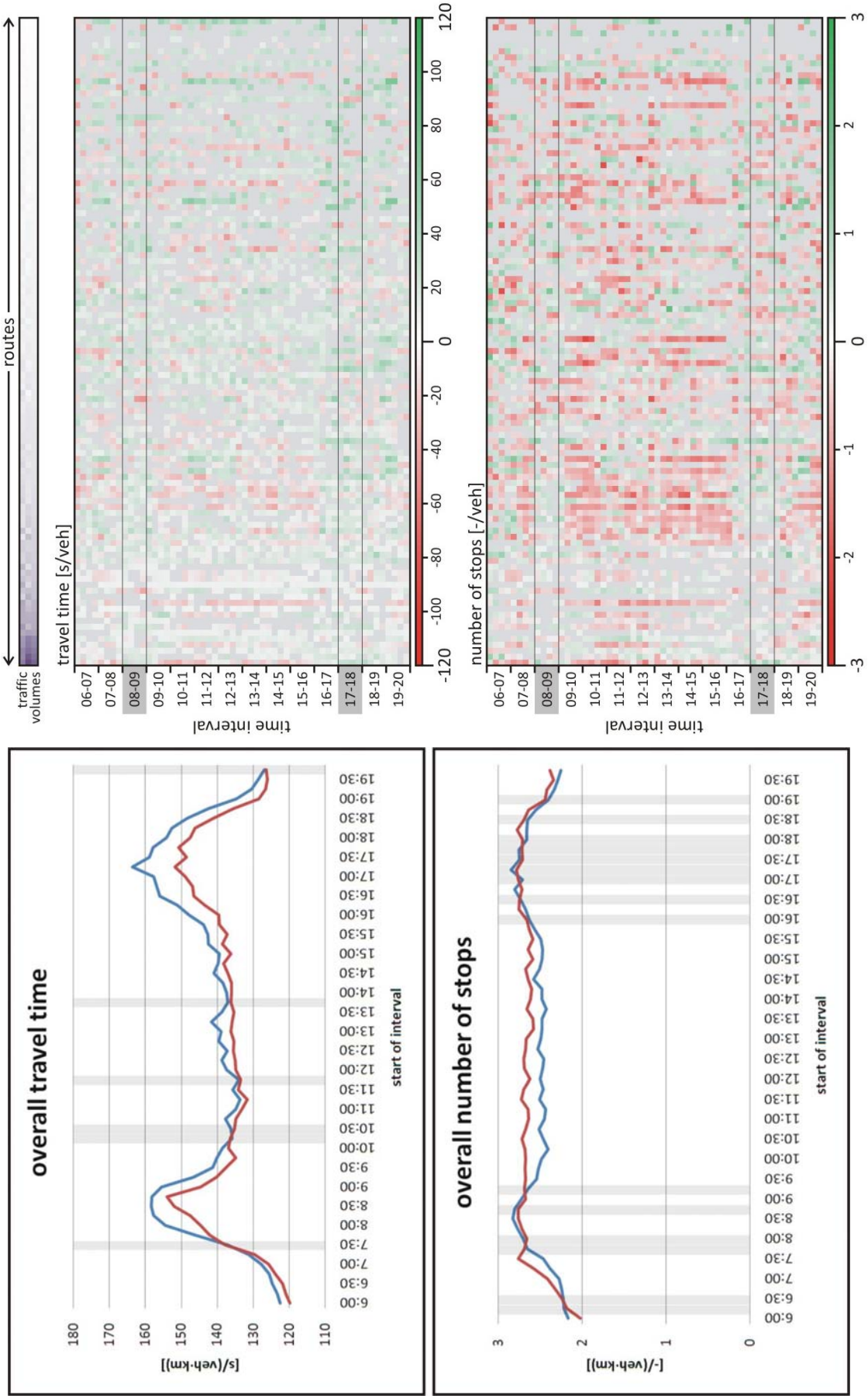


Figure D-23: ATCS, Webster cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, PGA, exact demand

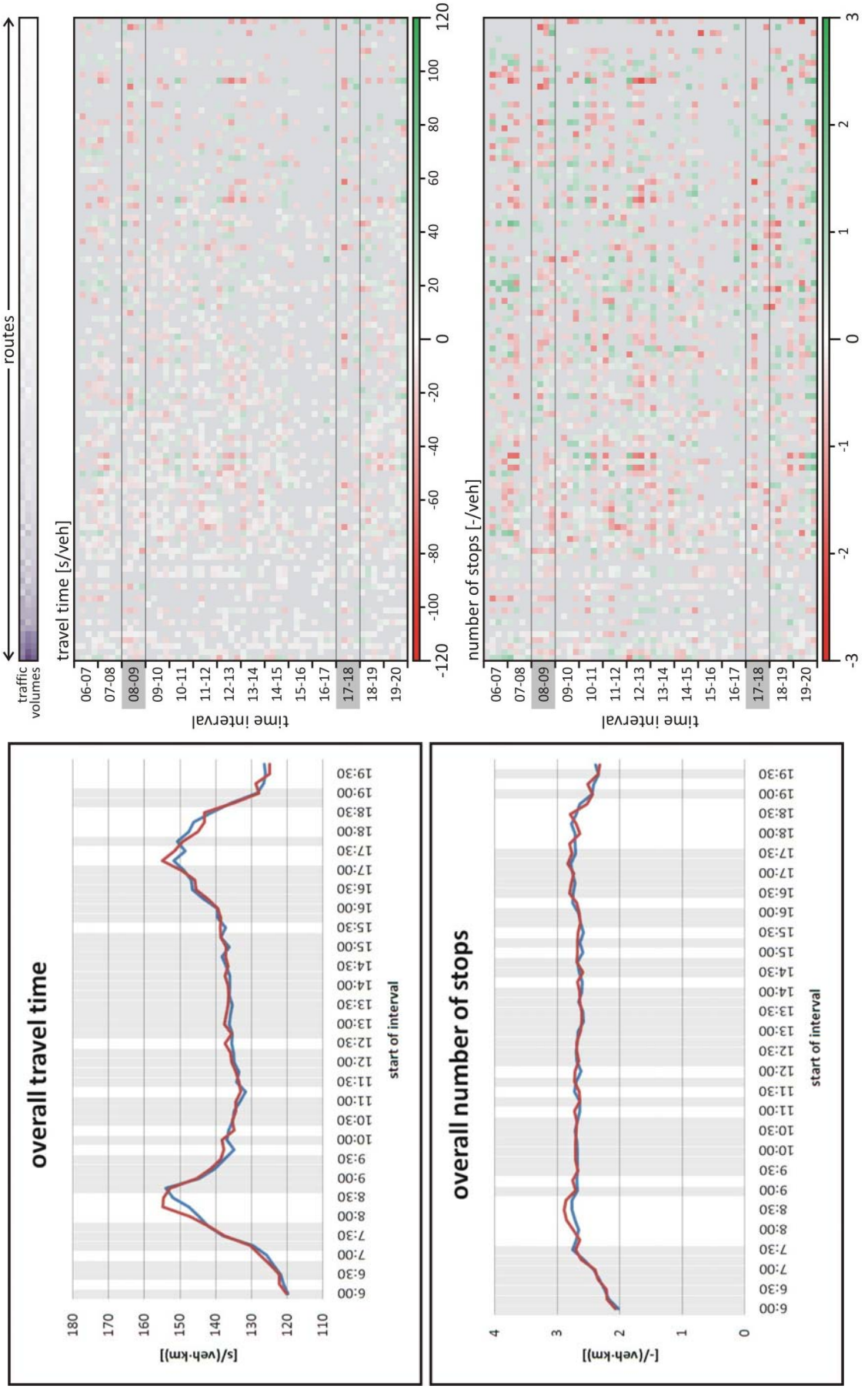


Figure D-24: ATCS, saturation based cycle lengths, PGA with consideration of transition losses, exact demand
ATCS, saturation based cycle lengths, PGA without consideration of transition losses, exact demand

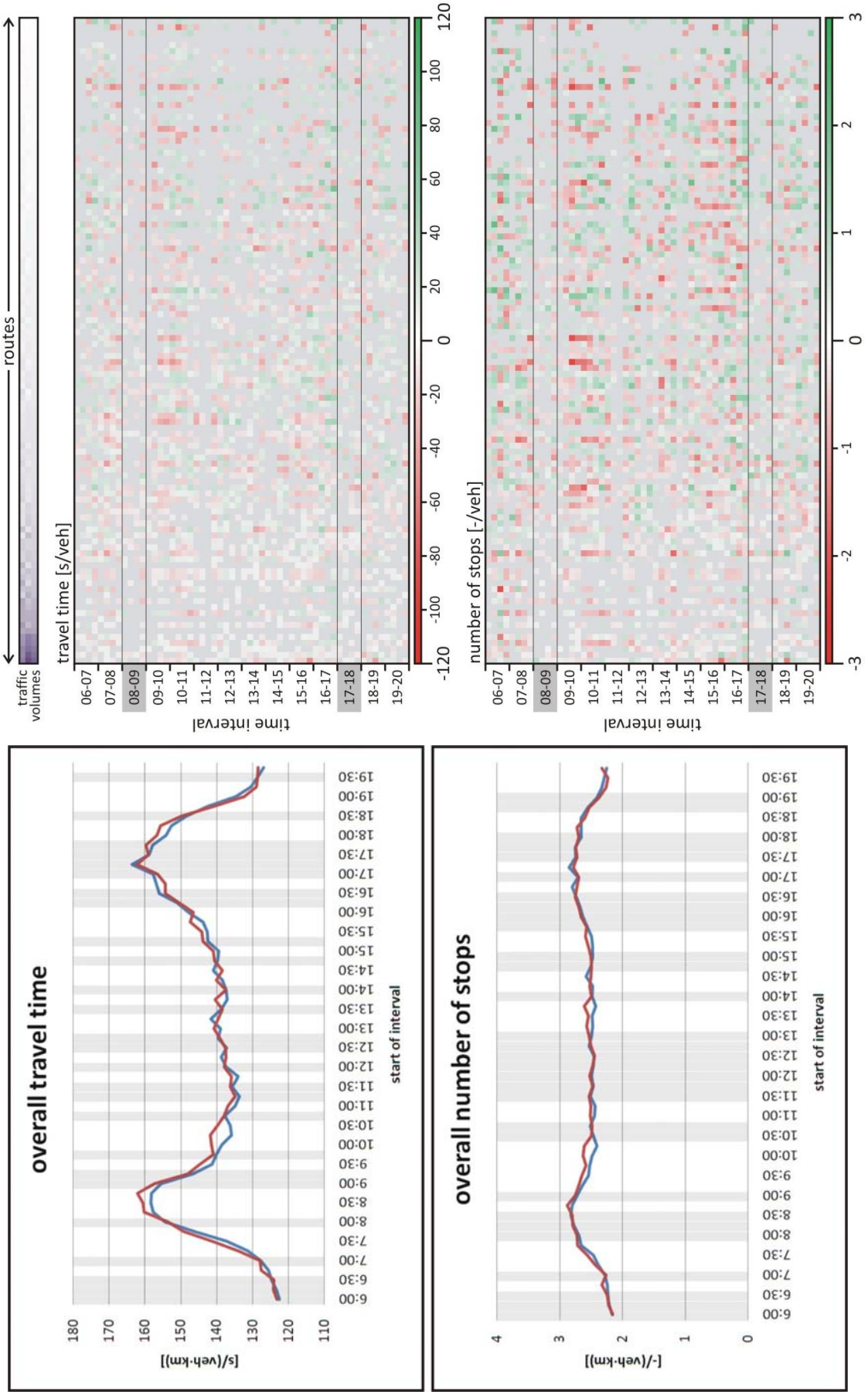


Figure D-25: ATCS, Webster cycle lengths, PGA with consideration of transition losses, exact demand
ATCS, Webster cycle lengths, PGA without consideration of transition losses, exact demand

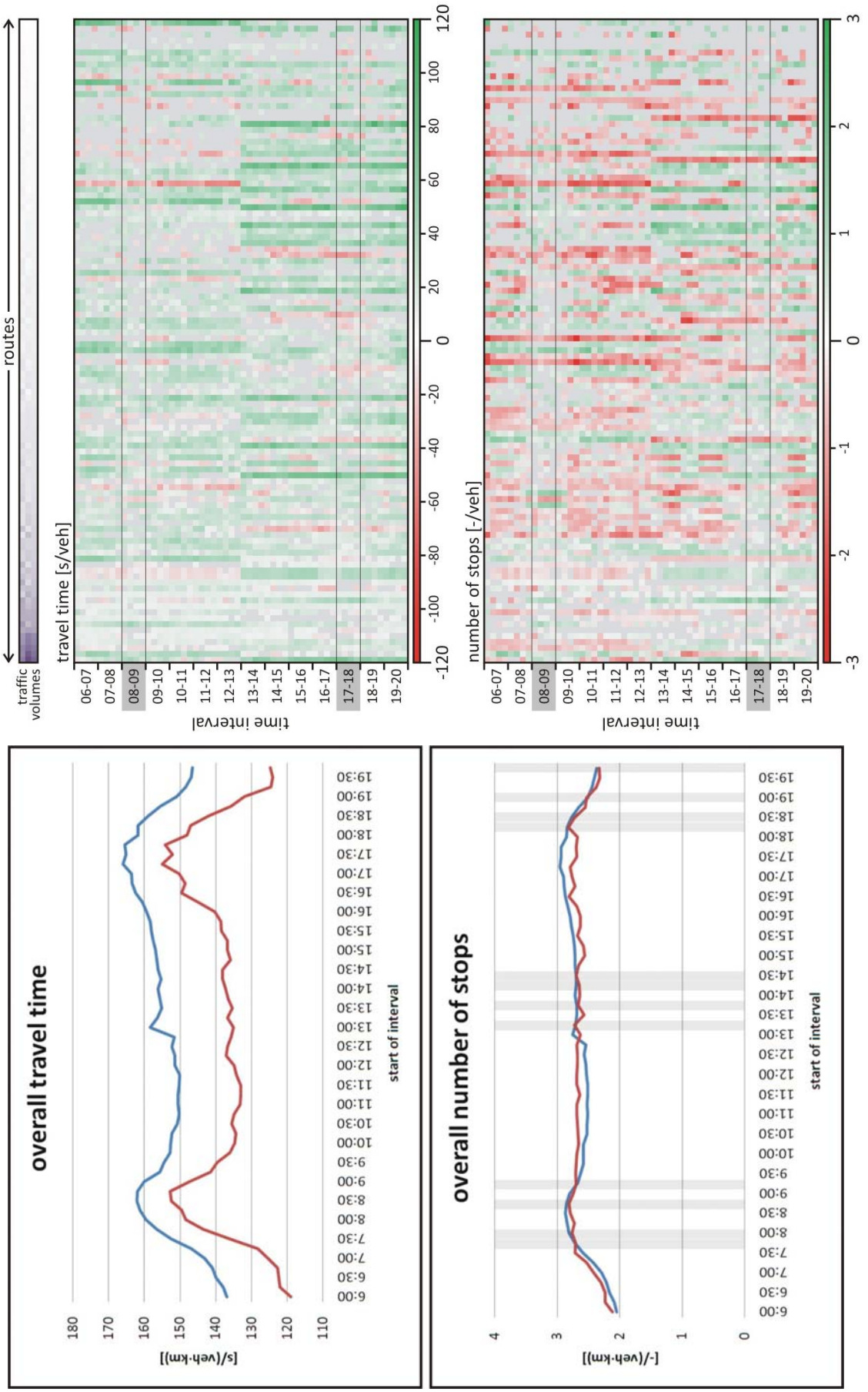


Figure D-26: T7F fixed time control
ATCS, saturation based cycle lengths, PGA, estimated demand (BPR)

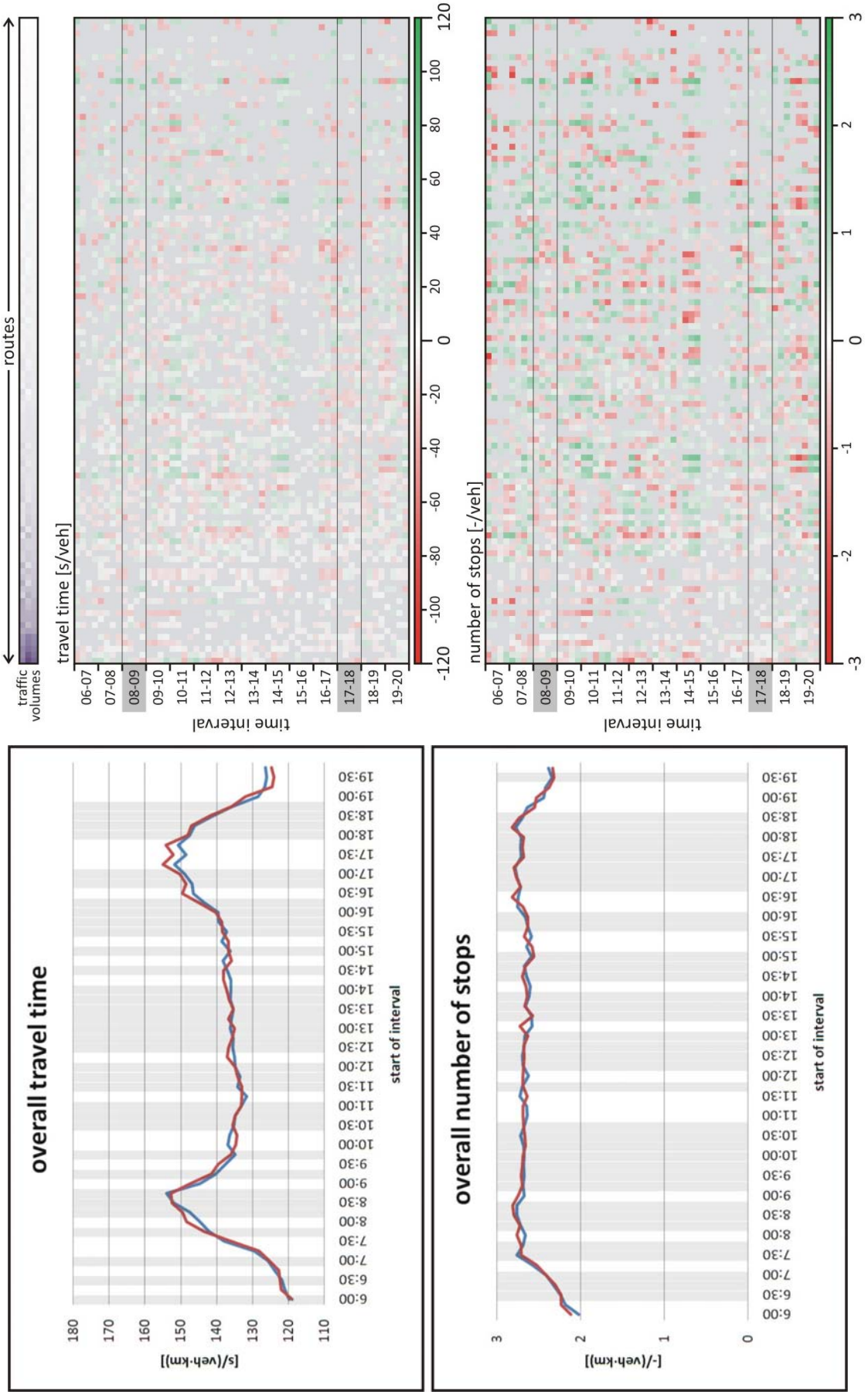


Figure D-27: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, PGA, estimated demand (BPR)

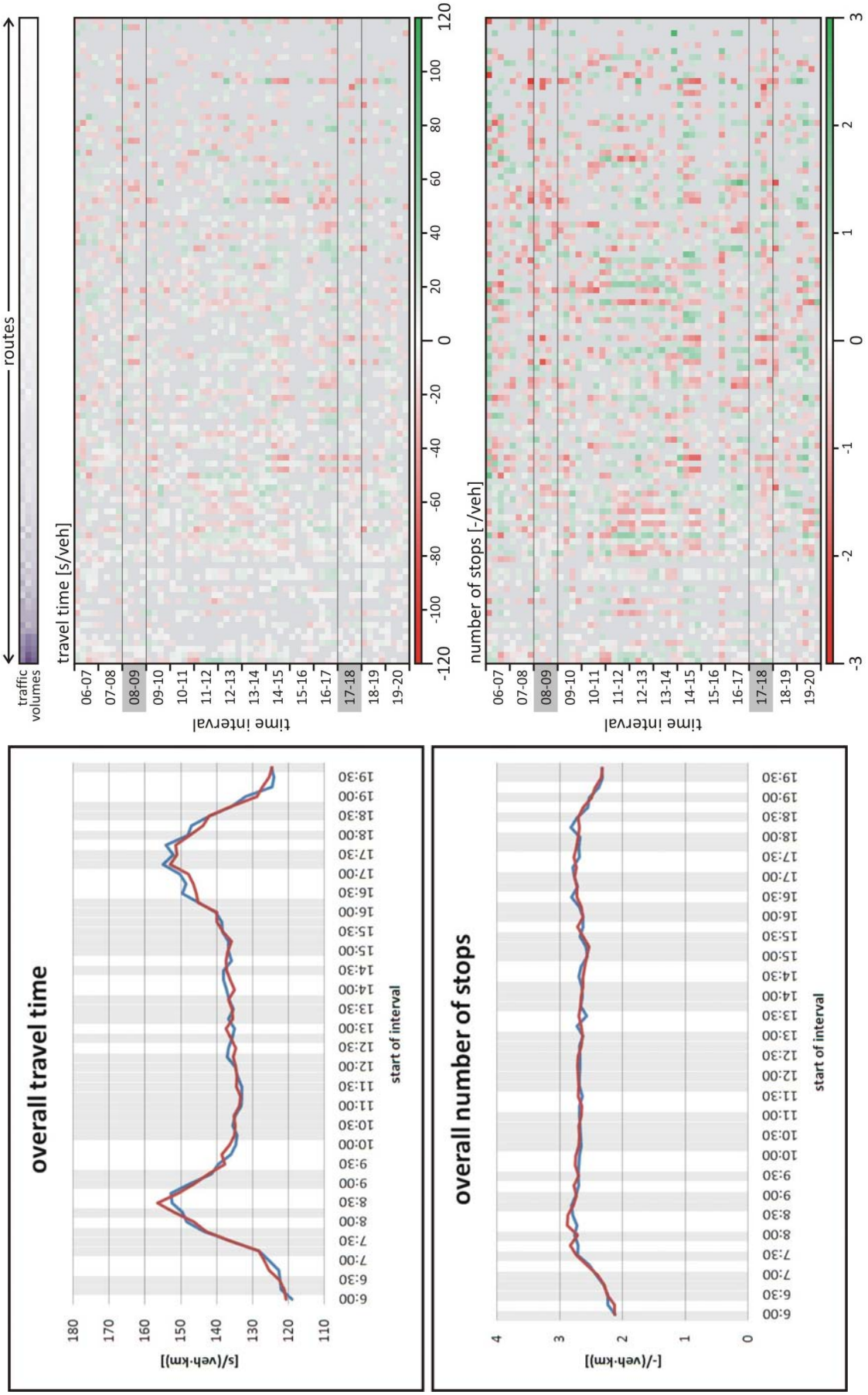


Figure D-28: ATCS, saturation based cycle lengths, PGA, estimated demand (BPR)
ATCS, saturation based cycle lengths, PGA, estimated demand (CTM)

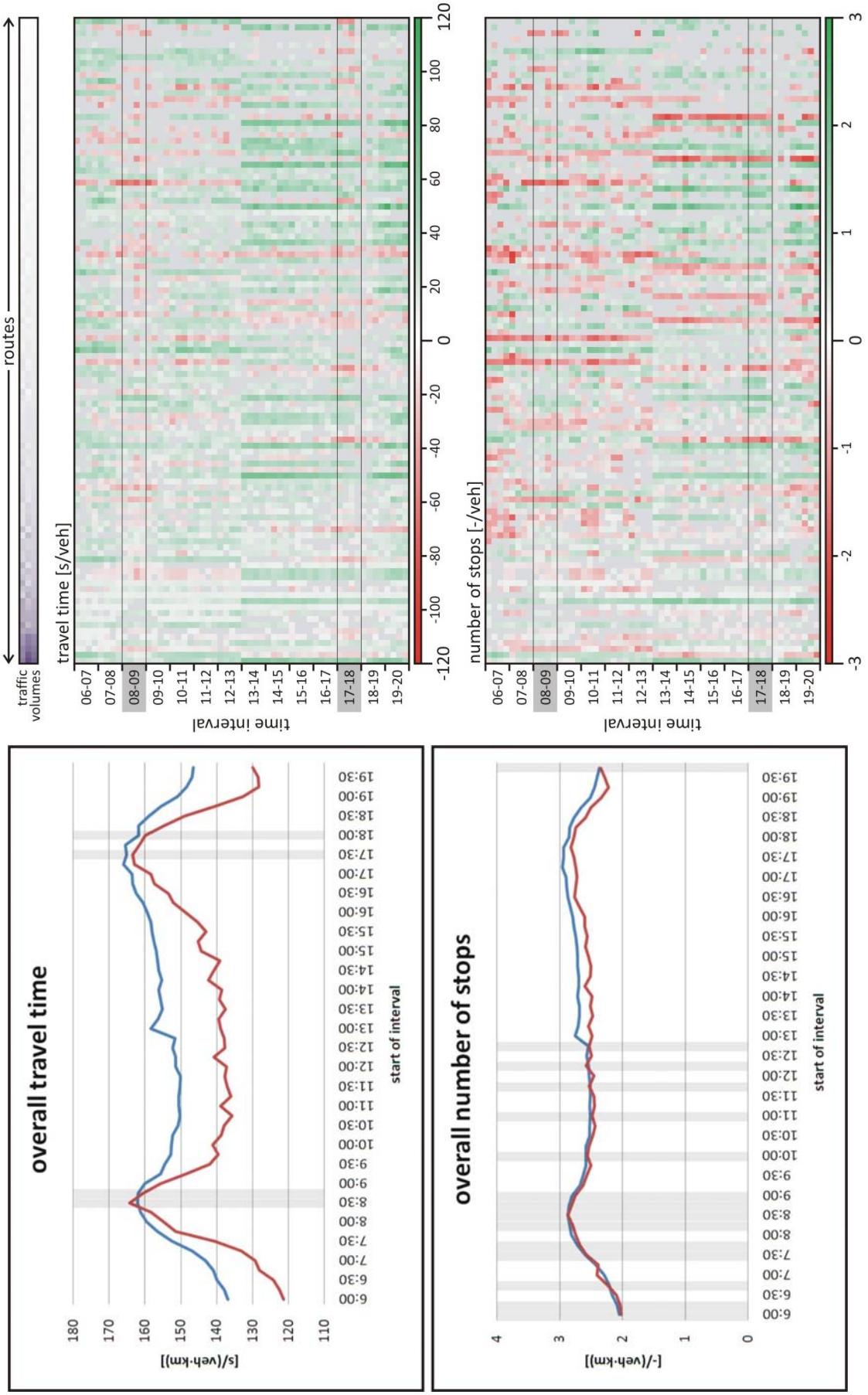


Figure D-29: T7F fixed time control
ATCS, Webster cycle lengths, estimated demand (BPR)

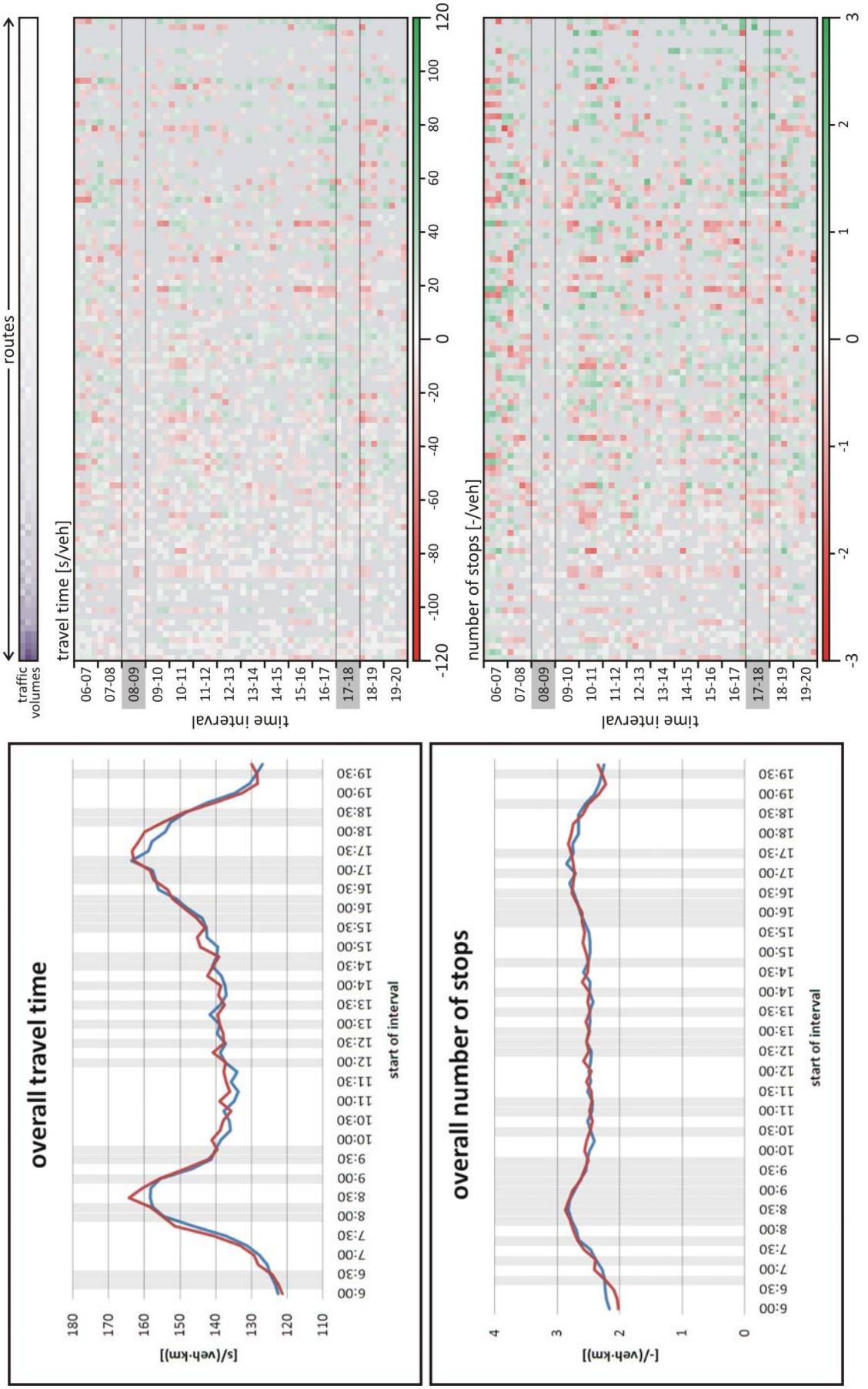


Figure D-30: ATCS, Webster cycle lengths, PGA, exact demand
ATCS, Webster cycle lengths, PGA, estimated demand (BPR)

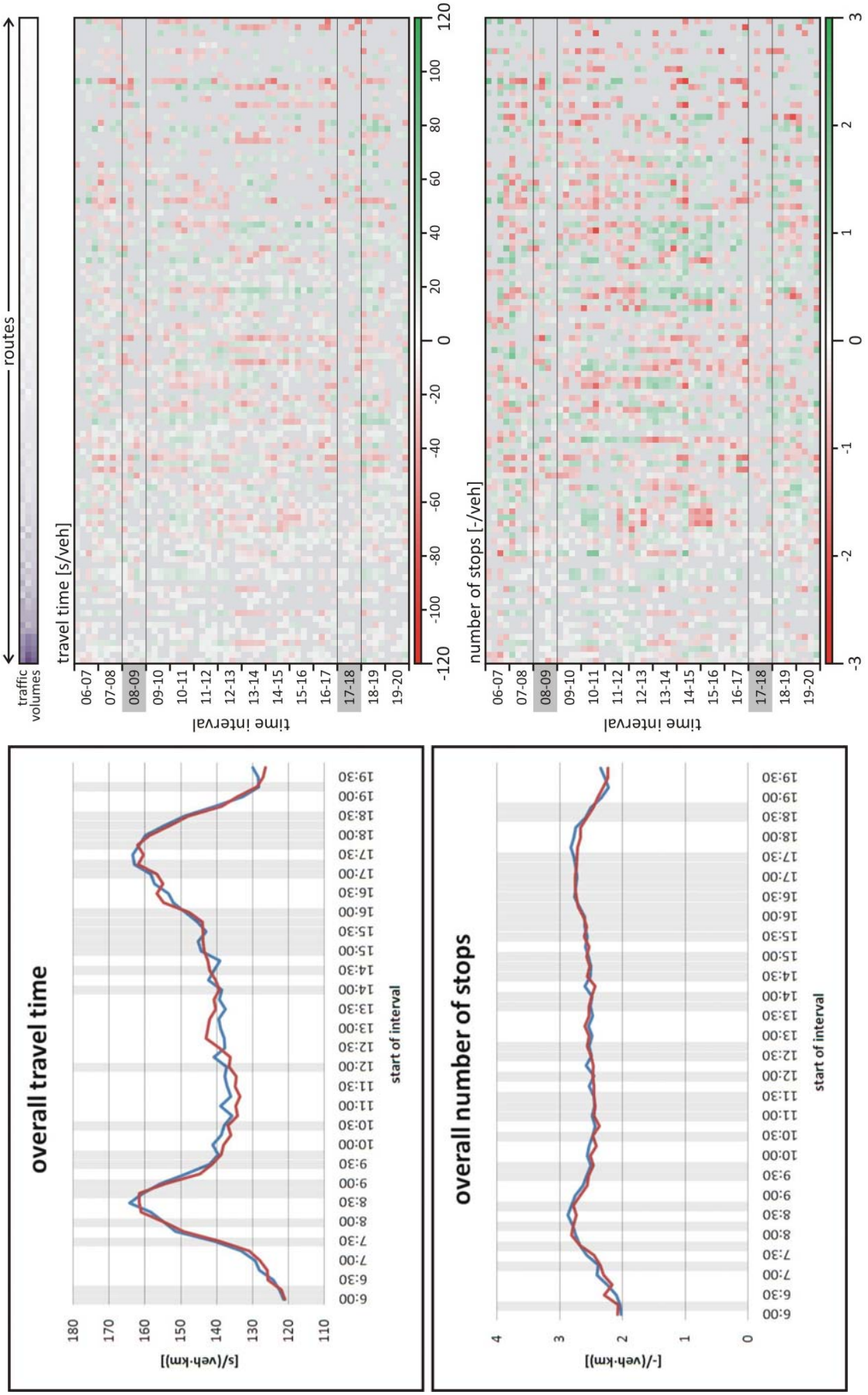


Figure D-31: ATCS, Webster cycle lengths, PGA, estimated demand (BPR)
ATCS, Webster cycle lengths, PGA, estimated demand (CTM)

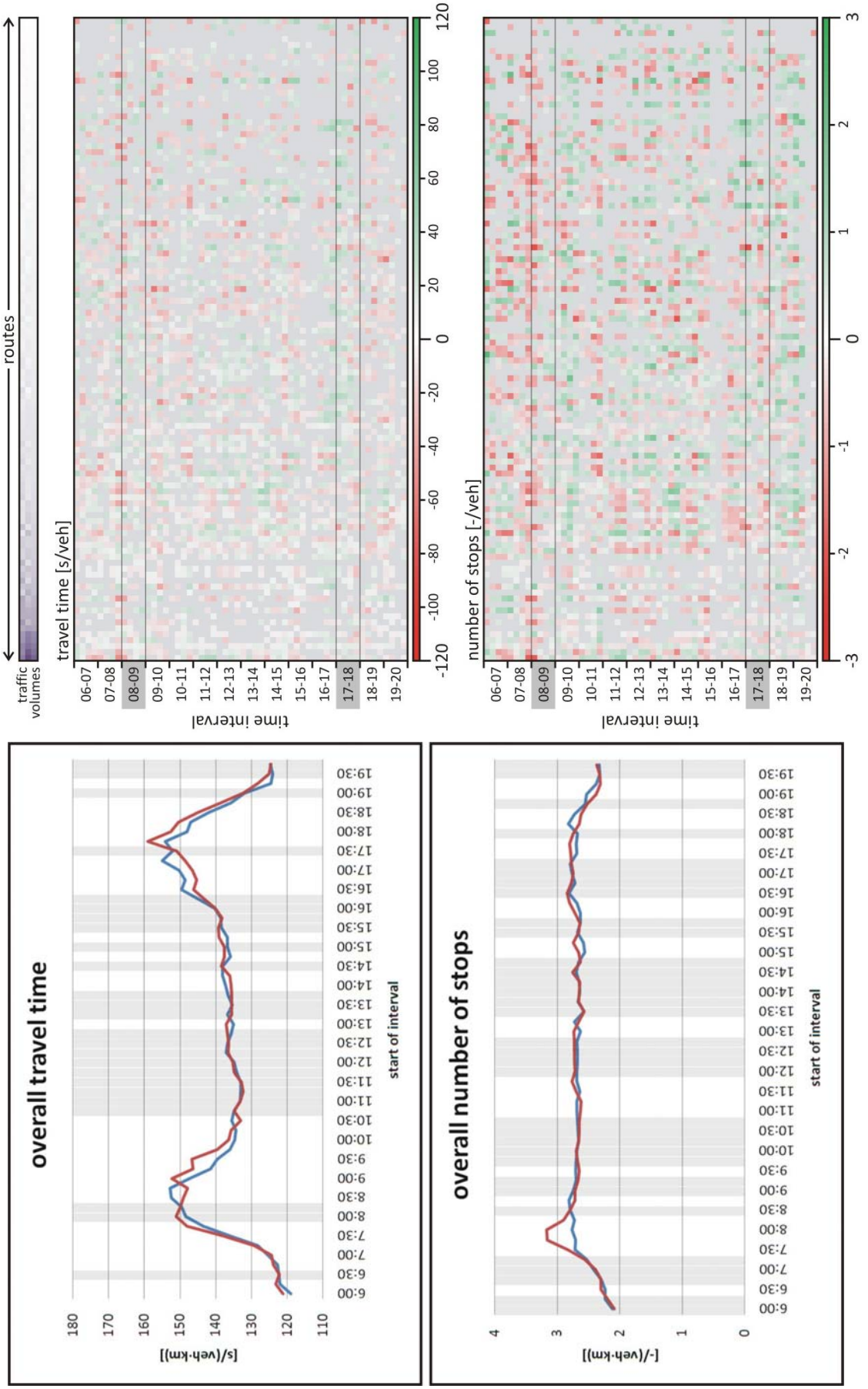


Figure D-32: ATCS, saturation based cycle lengths, PGA, estimated demand (BPR) with forecasting
ATCS, saturation based cycle lengths, PGA, estimated demand (BPR) without forecasting

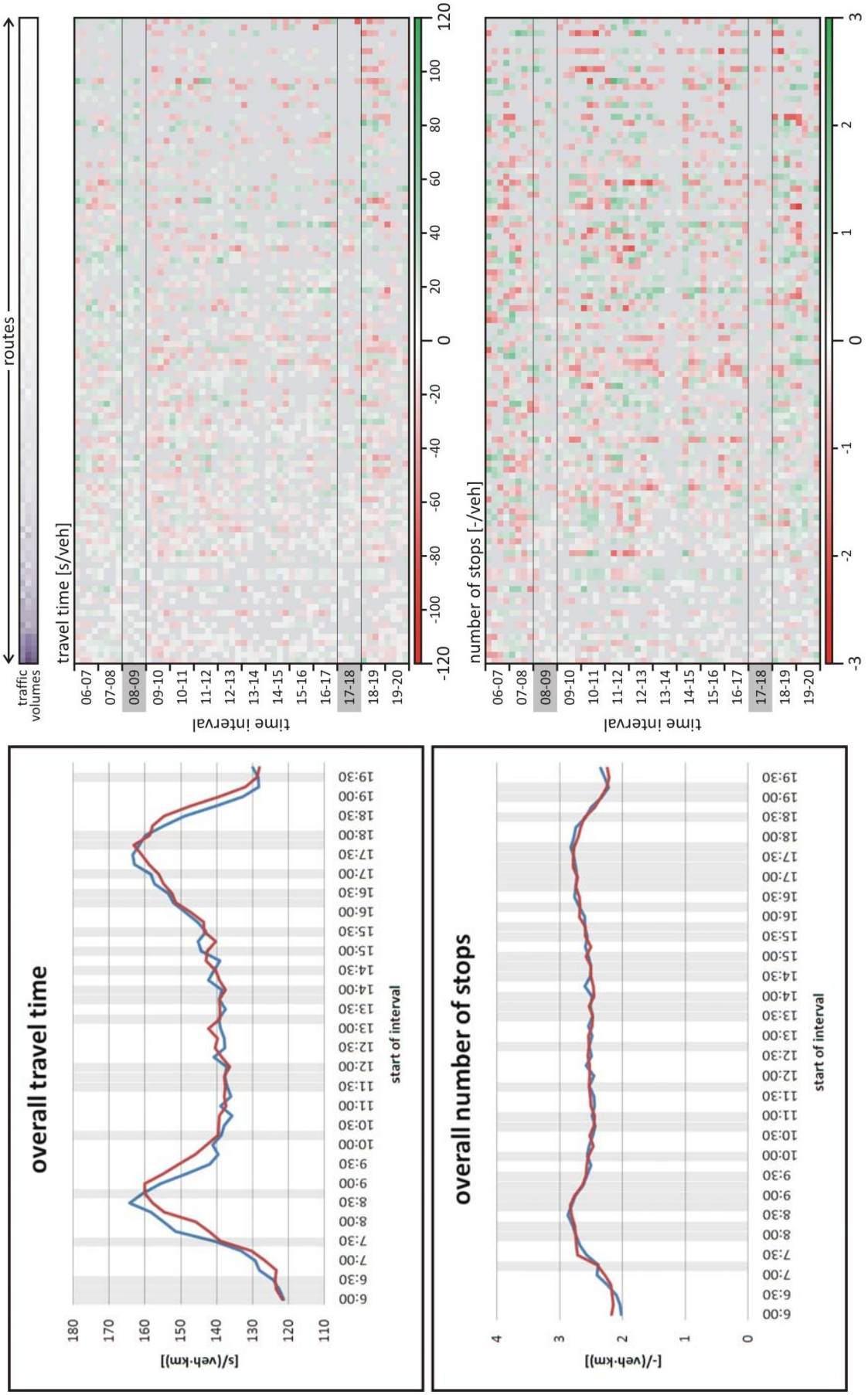


Figure D-33: ATCS, Webster cycle lengths, PGA, estimated demand (BPR) with forecasting
ATCS, Webster cycle lengths, PGA, estimated demand (BPR) without forecasting

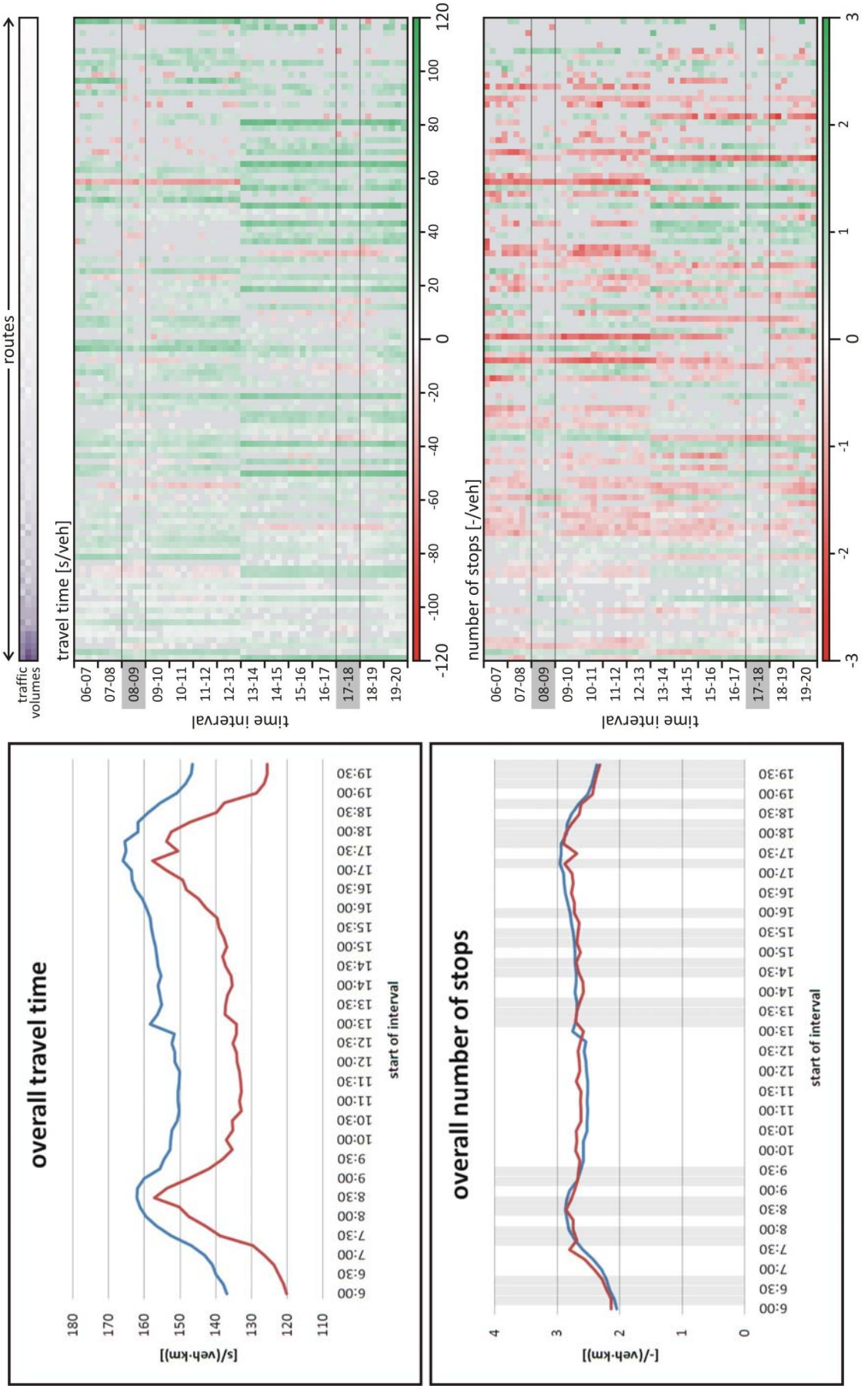


Figure D-34: T7F fixed time control
ATCS, saturation based cycle lengths, PGA, online optimization

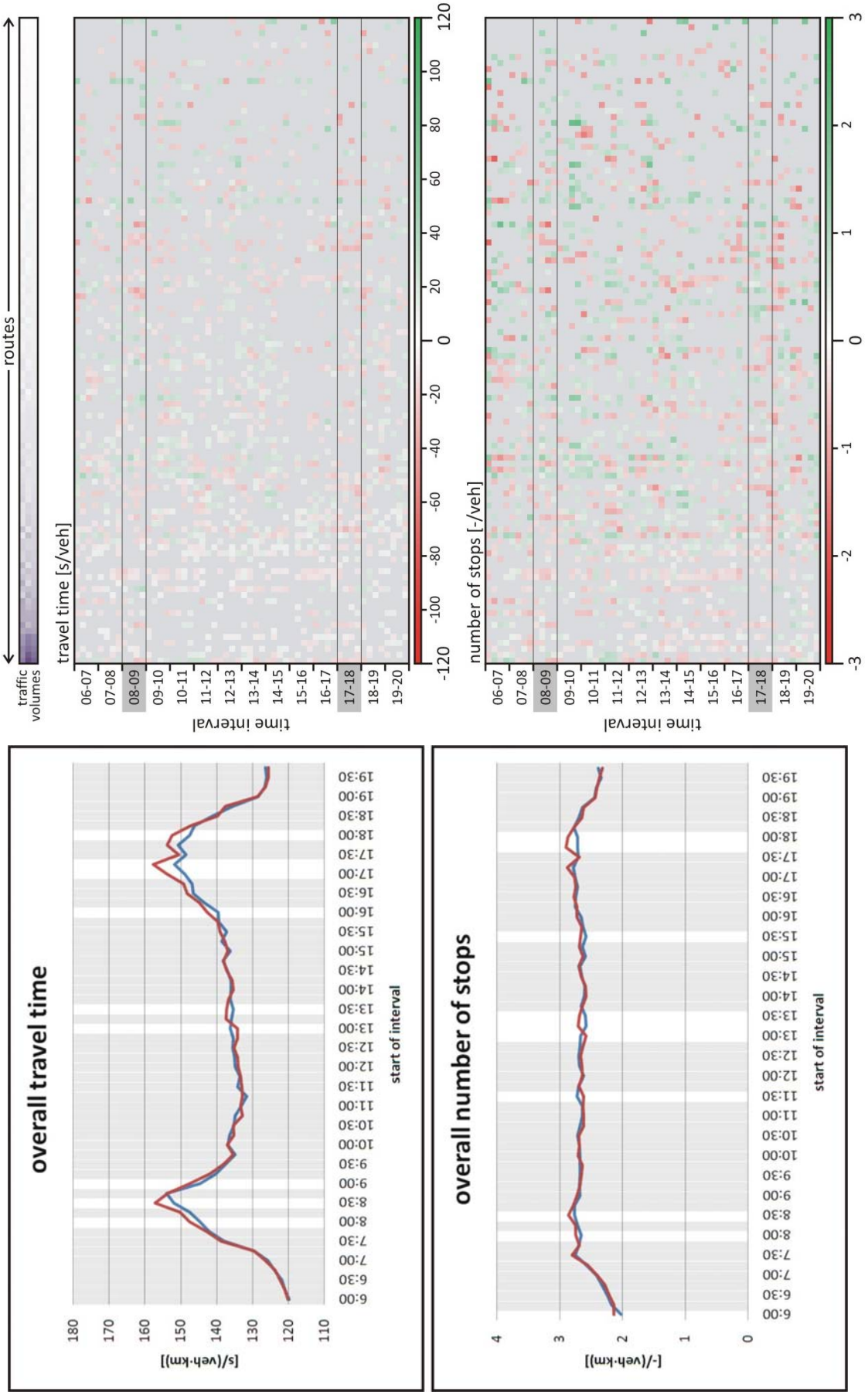


Figure D-35: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, PGA, online optimization

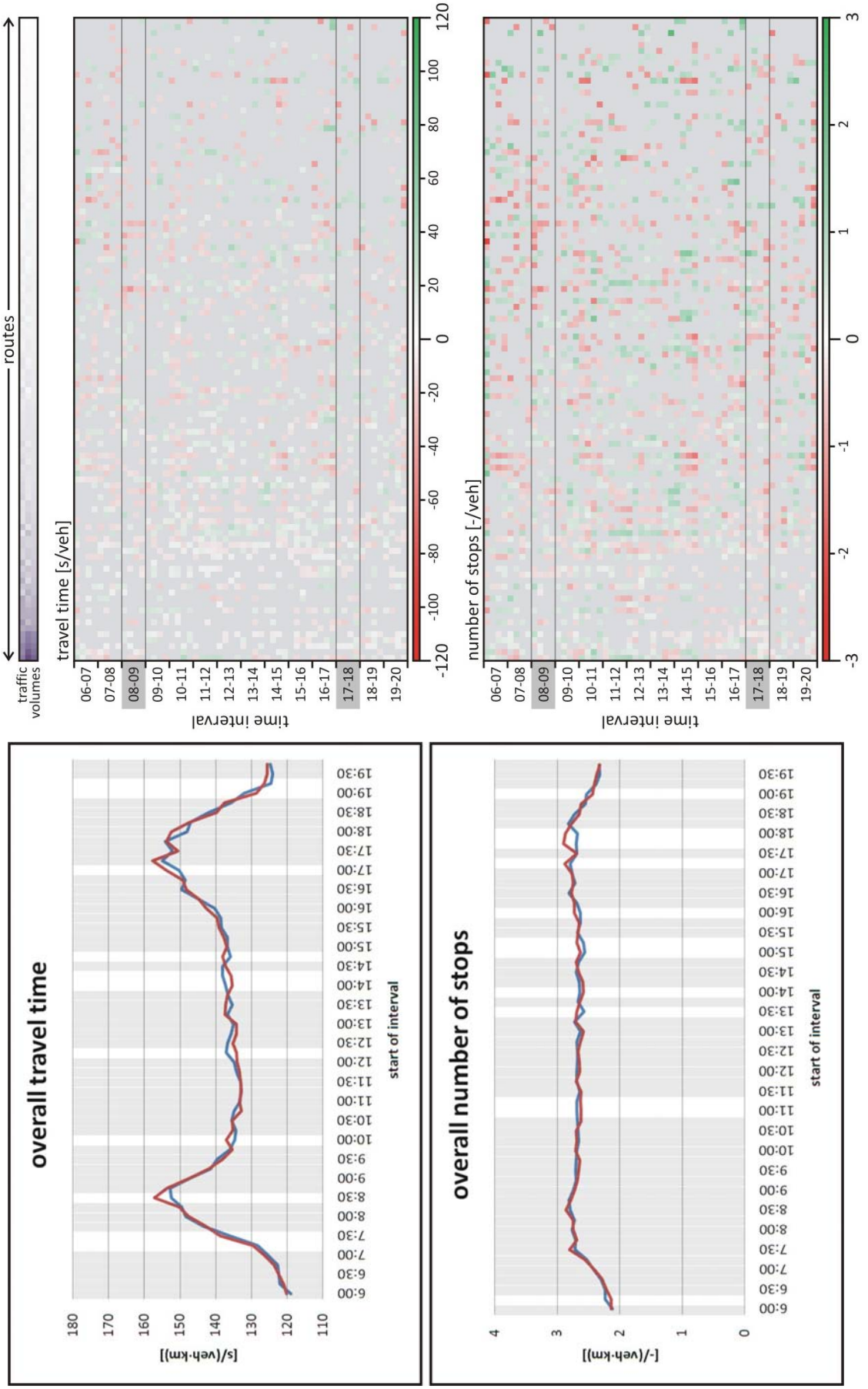


Figure D-36: ATCS, saturation based cycle lengths, PGA, estimated demand (BPR)
ATCS, saturation based cycle lengths, PGA, online optimization

D.3 Hanover Südstadt network

The following diagrams show detailed results of the Hanover Südstadt network.

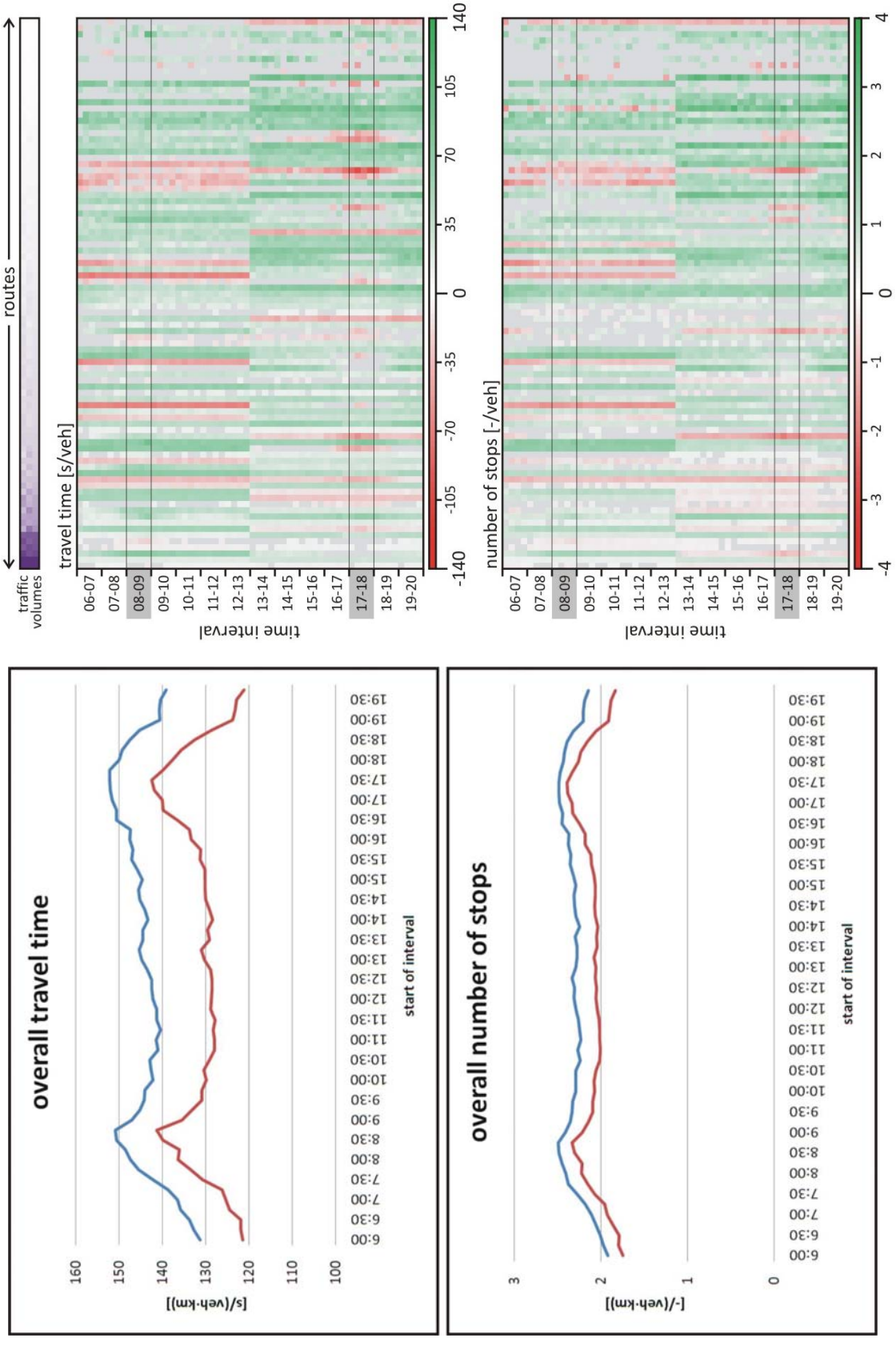


Figure D-37: original fixed time control
T7F fixed time control

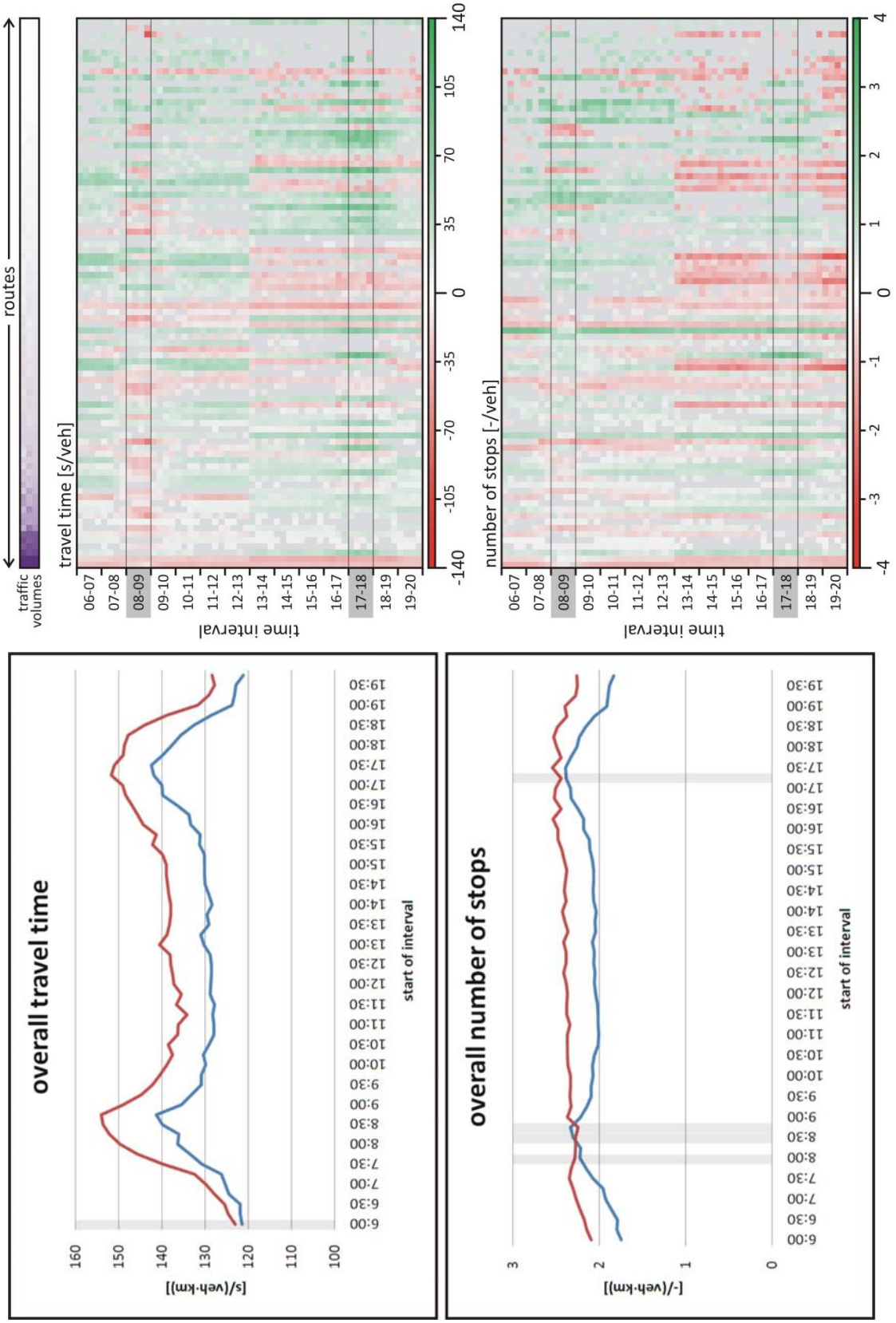


Figure D-38: T7F fixed time control
ATCS, saturation based cycle lengths, no offset optimization, exact demand

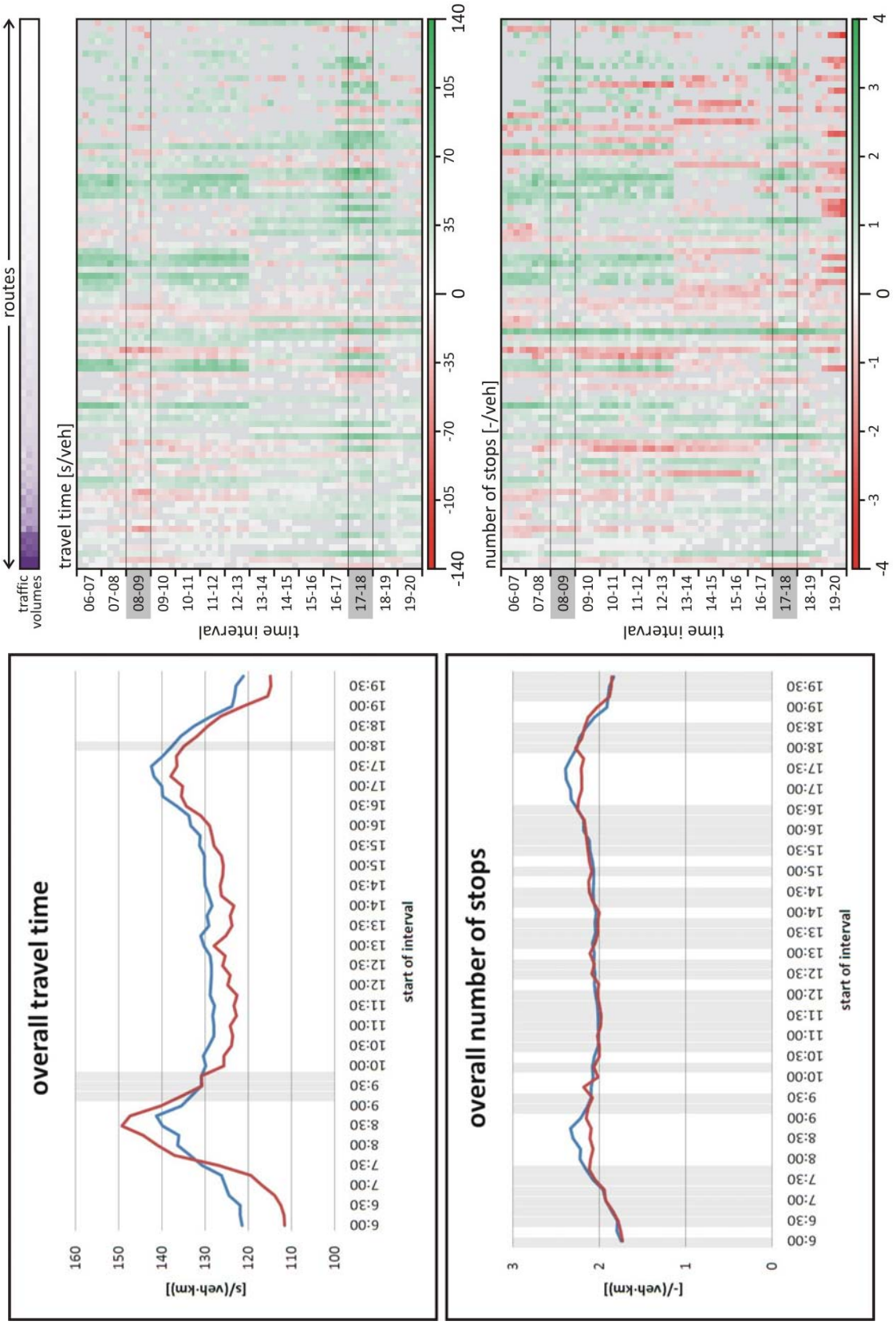


Figure D-39: T7F fixed time control
ATCS, saturation based cycle lengths, PGA, exact demand

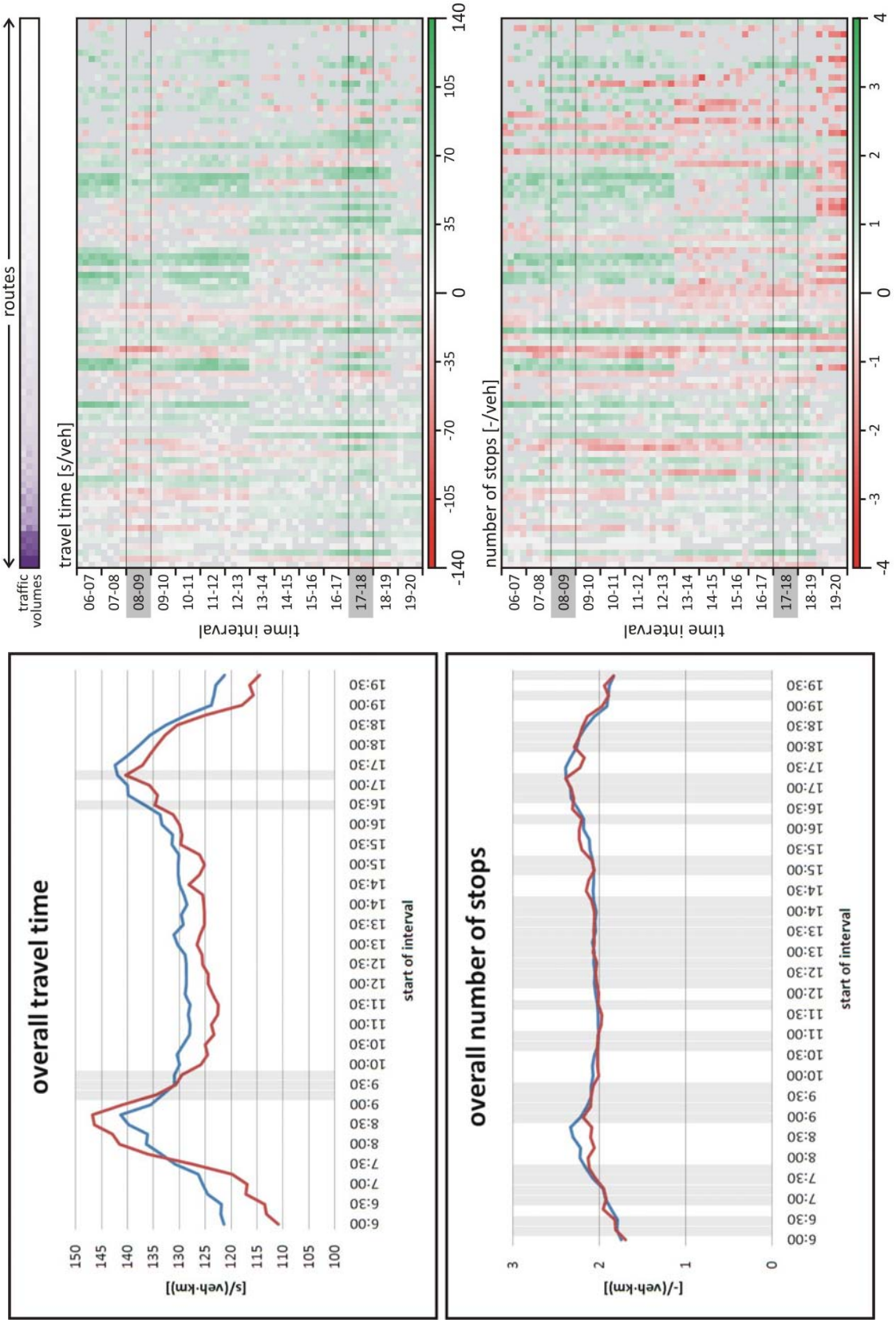


Figure D-40: T7F fixed time control
ATCS, saturation based cycle lengths, SGA, exact demand

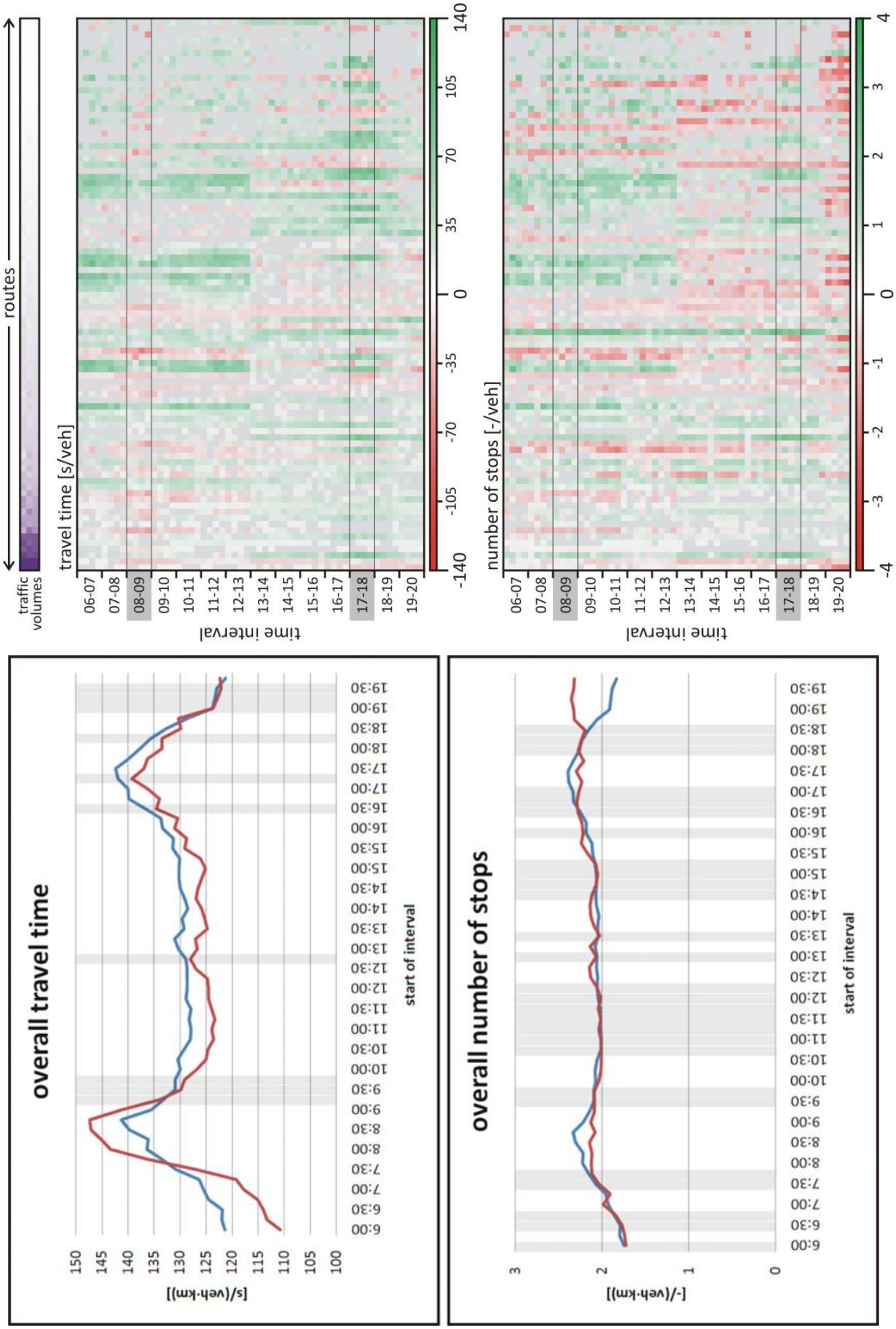


Figure D-41: T7F fixed time control
ATCS, saturation based cycle lengths, SE, exact demand

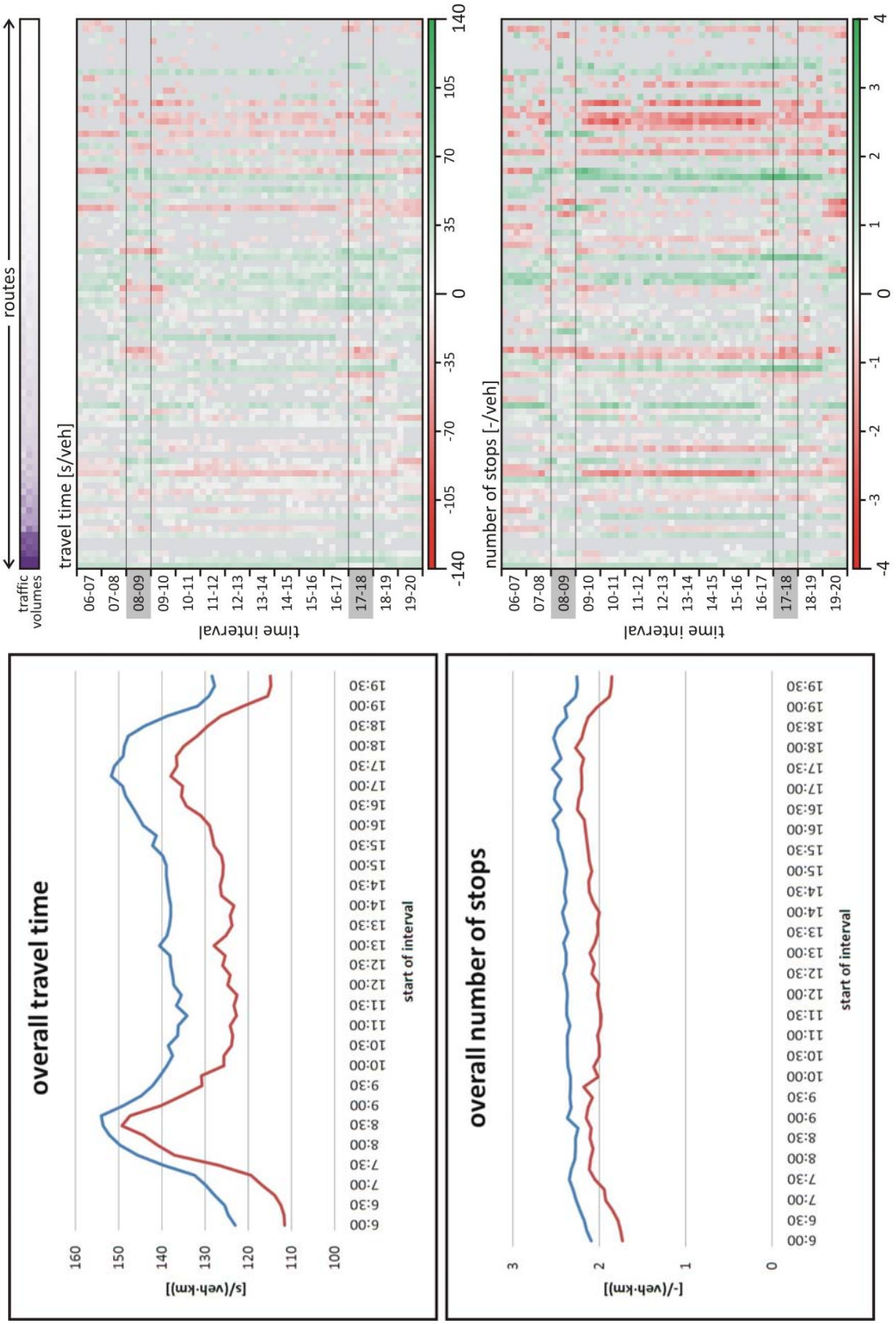


Figure D-42: ATCS, saturation based cycle lengths, no offset optimization, exact demand
ATCS, saturation based cycle lengths, PGA, exact demand

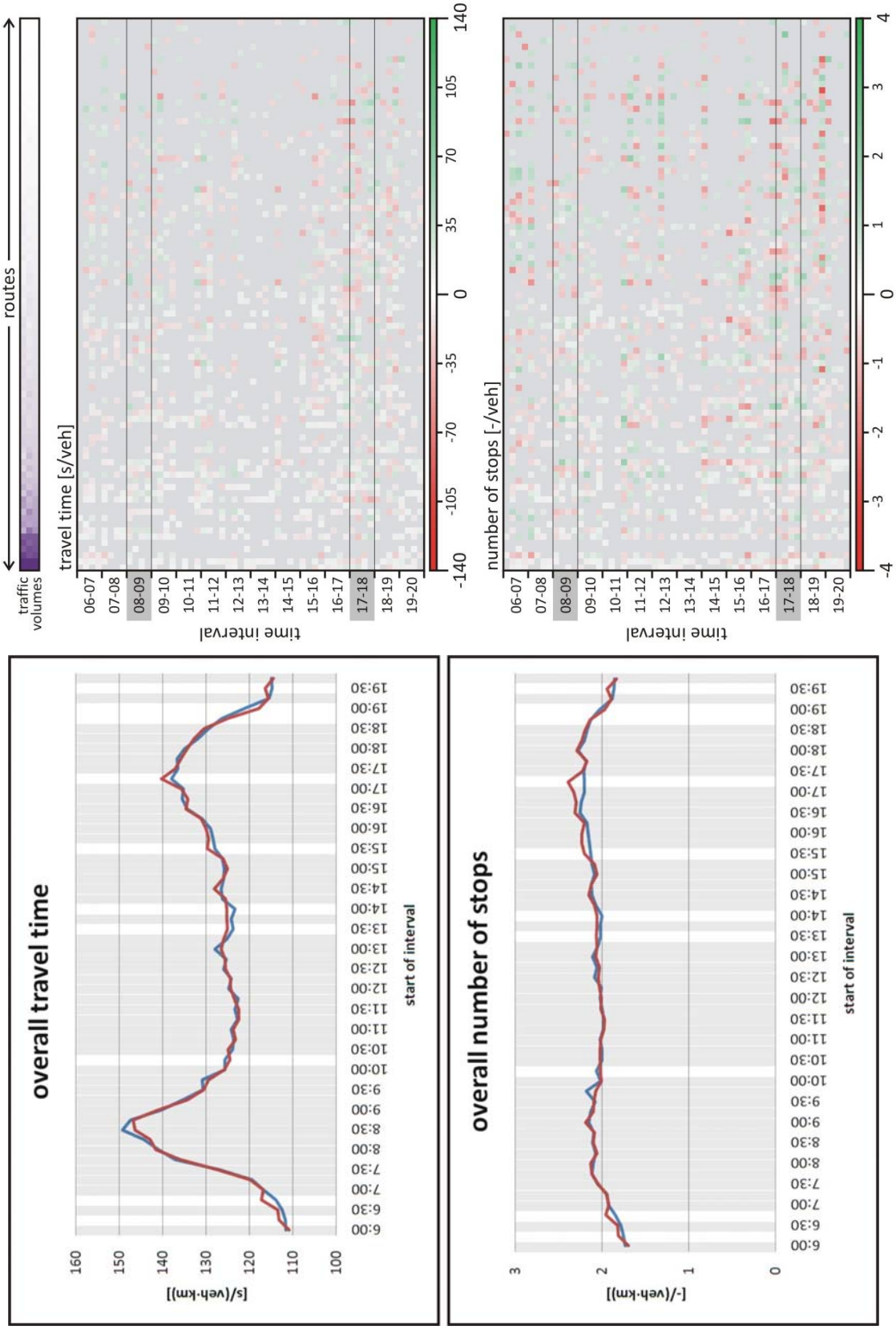


Figure D-43: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, SGA, exact demand

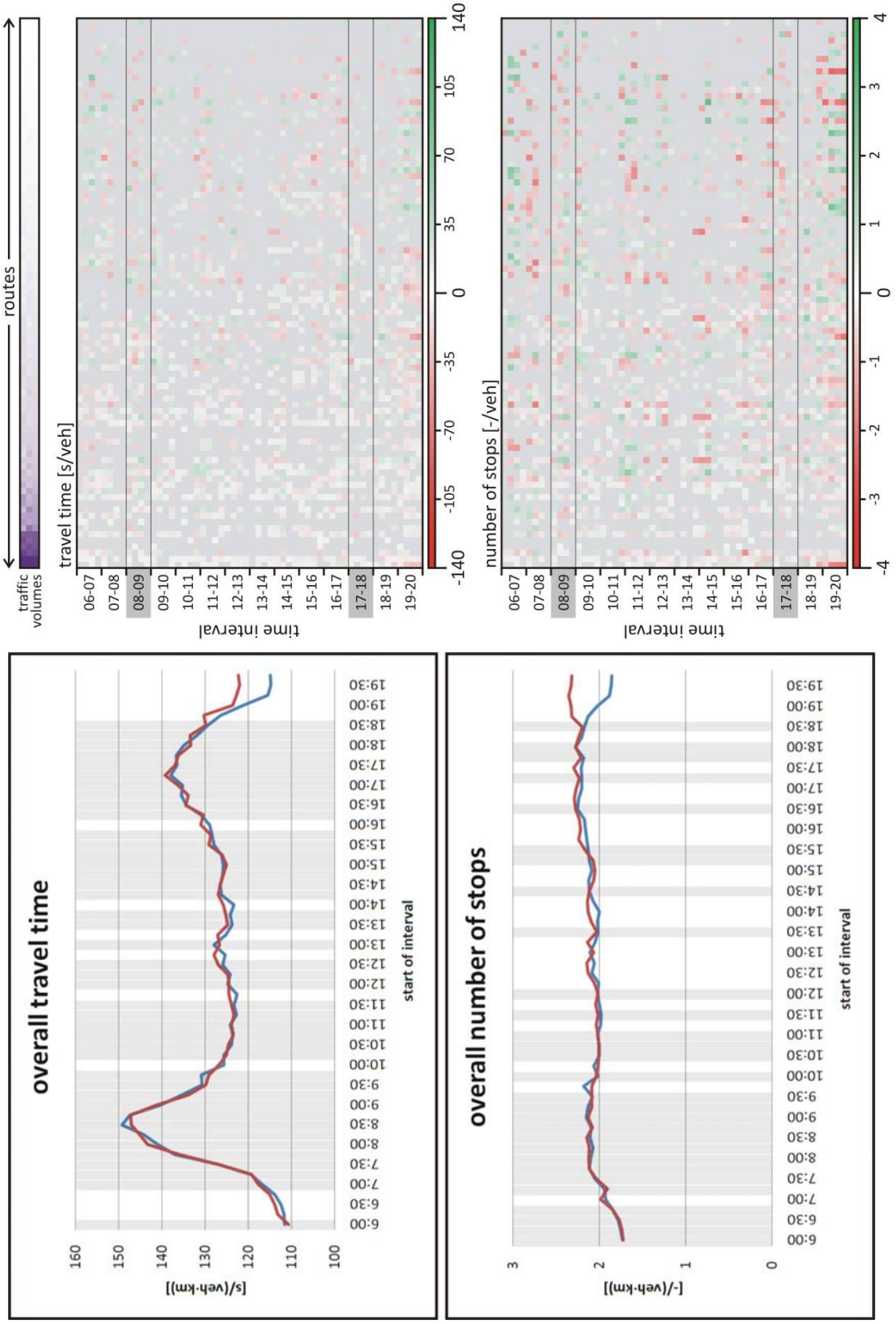


Figure D-44: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, SE, exact demand

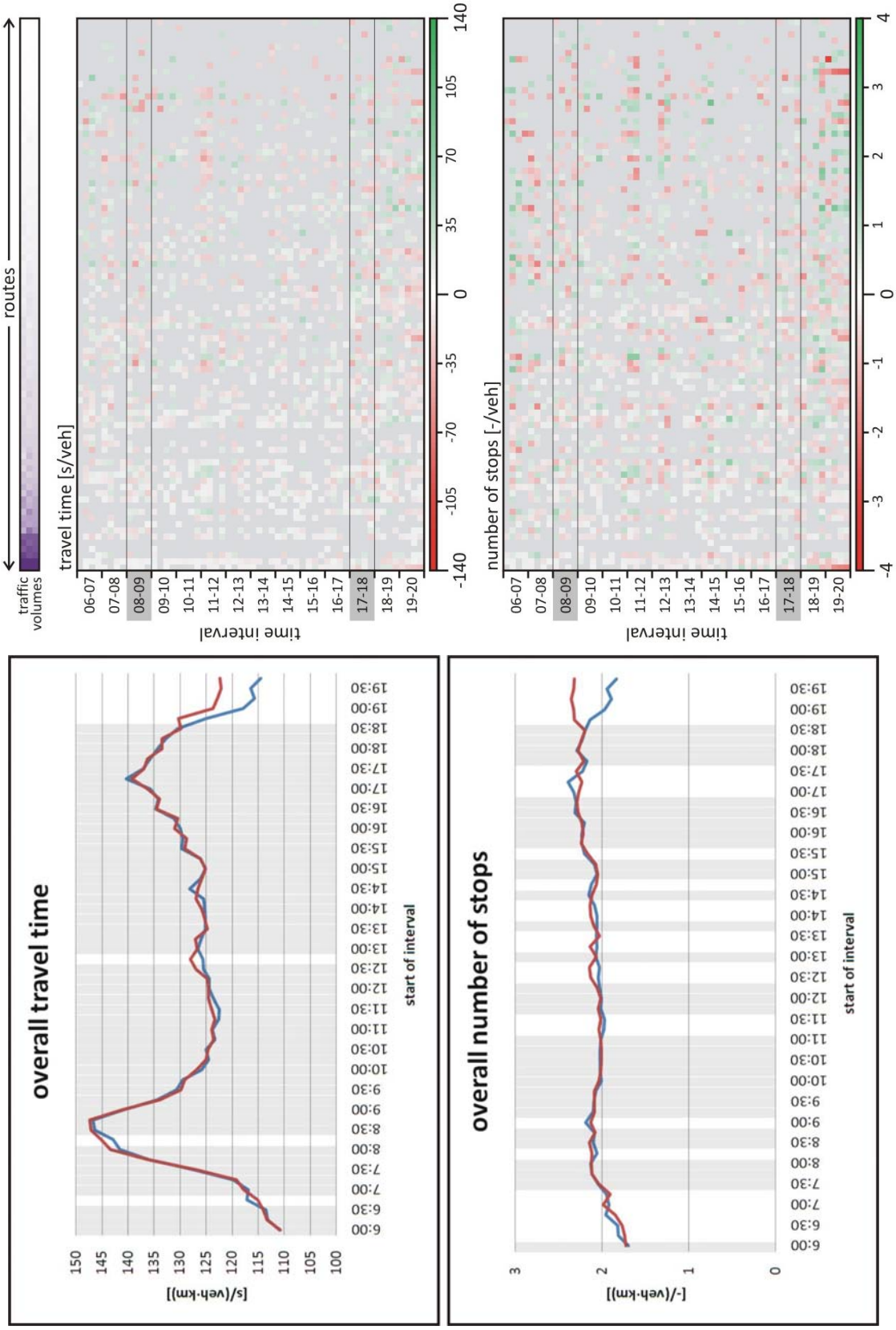


Figure D-45: ATCS, saturation based cycle lengths, SGA, exact demand
ATCS, saturation based cycle lengths, SE, exact demand

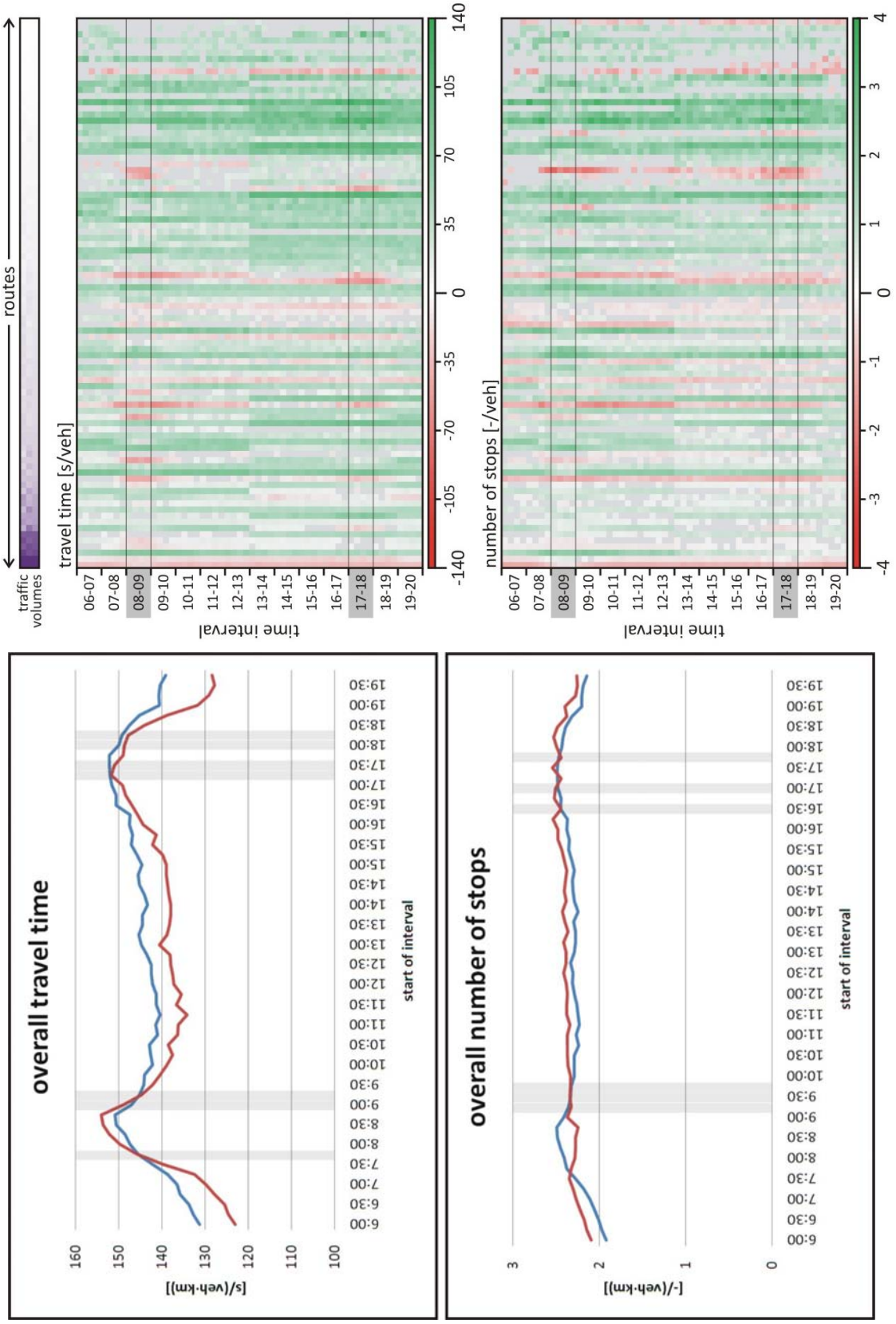


Figure D-46: original fixed time control
ATCS, saturation based cycle lengths, no offset optimization, exact demand

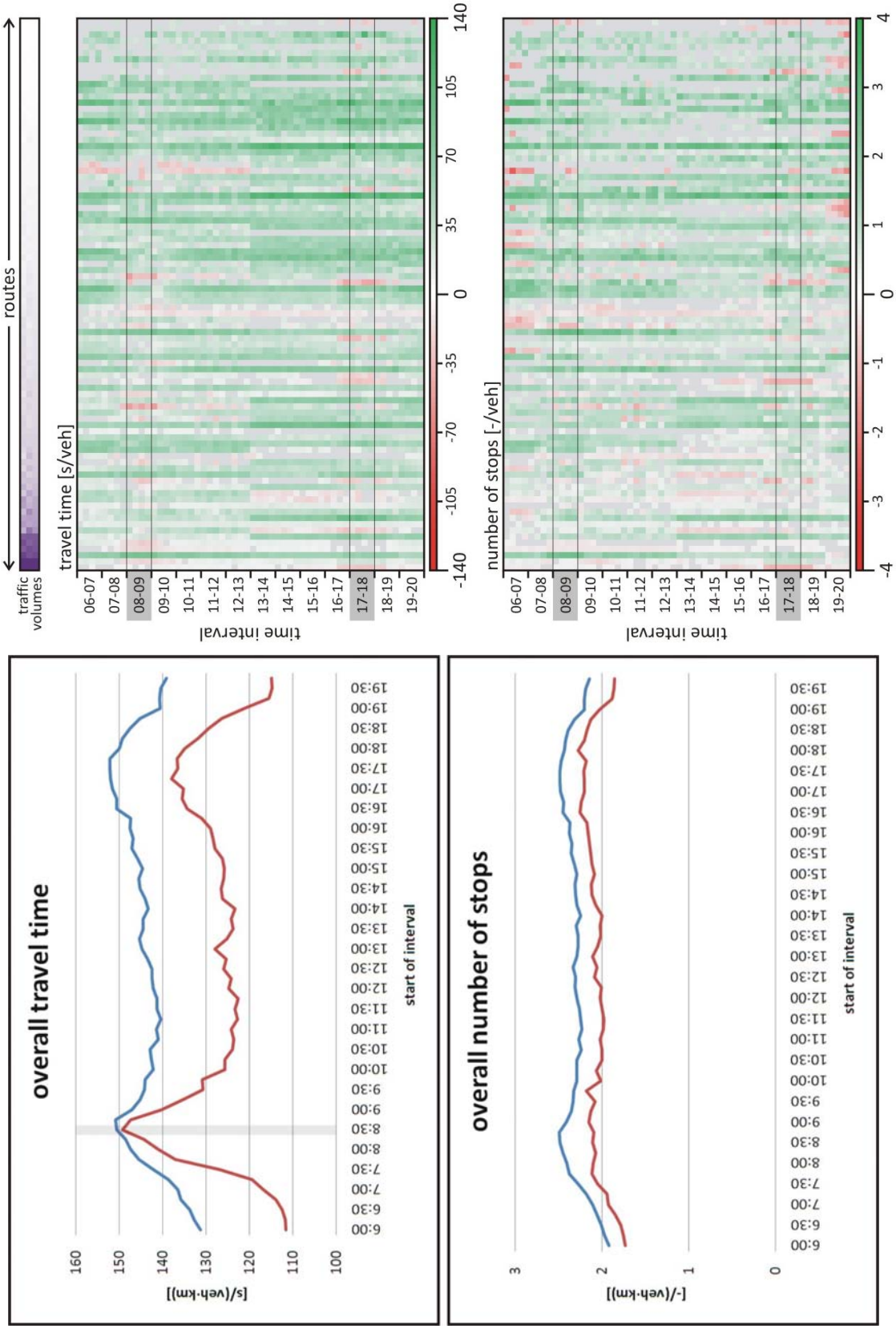


Figure D-47: original fixed time control
ATCS, saturation based cycle lengths, PGA, exact demand

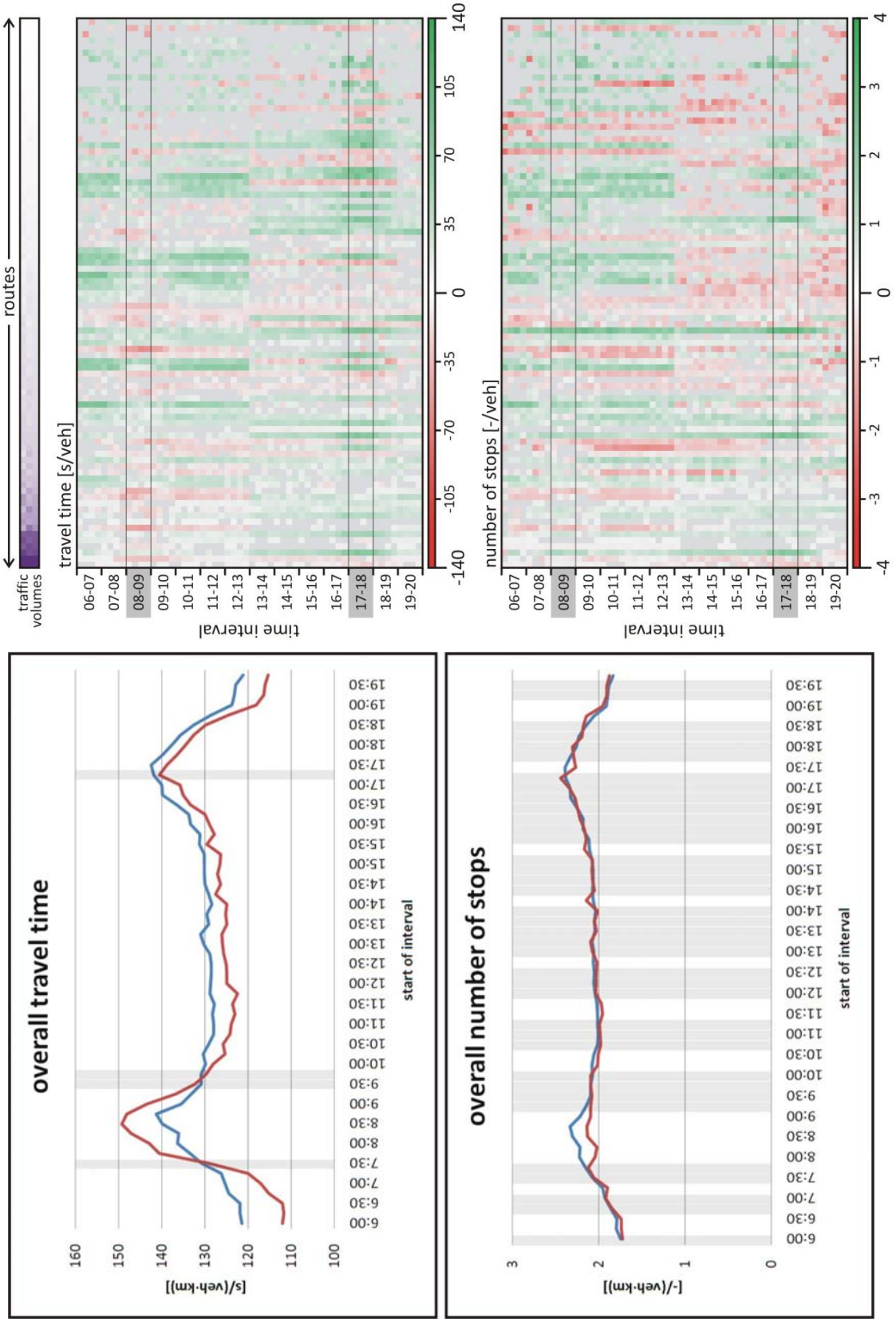


Figure D-48: T7F fixed time control
ATCS, saturation based cycle lengths, PGA, estimated demand (BPR)

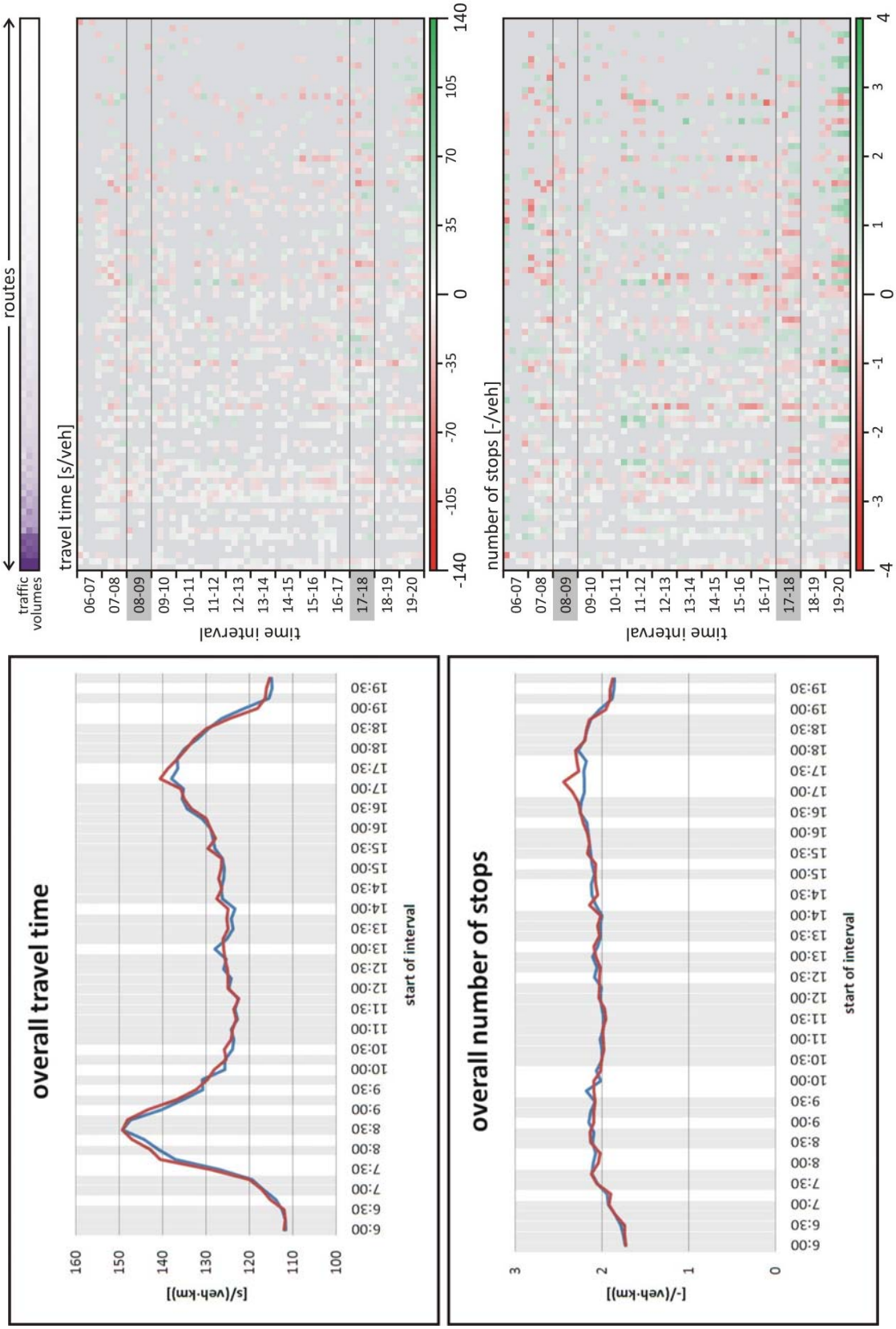


Figure D-49: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, PGA, estimated demand (BPR)

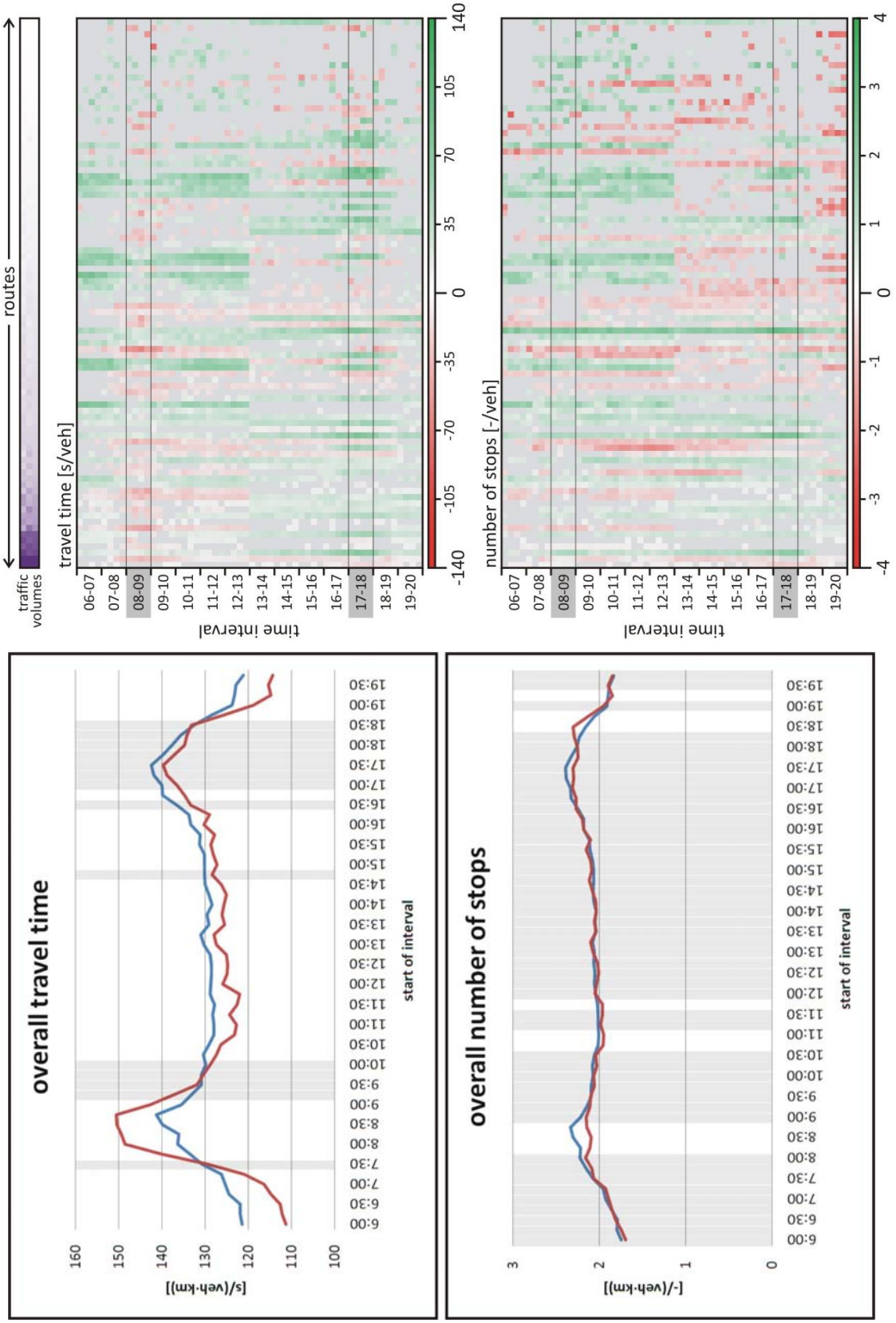


Figure D-50: T7F fixed time control
ATCS, saturation based cycle lengths, PGA, online optimization

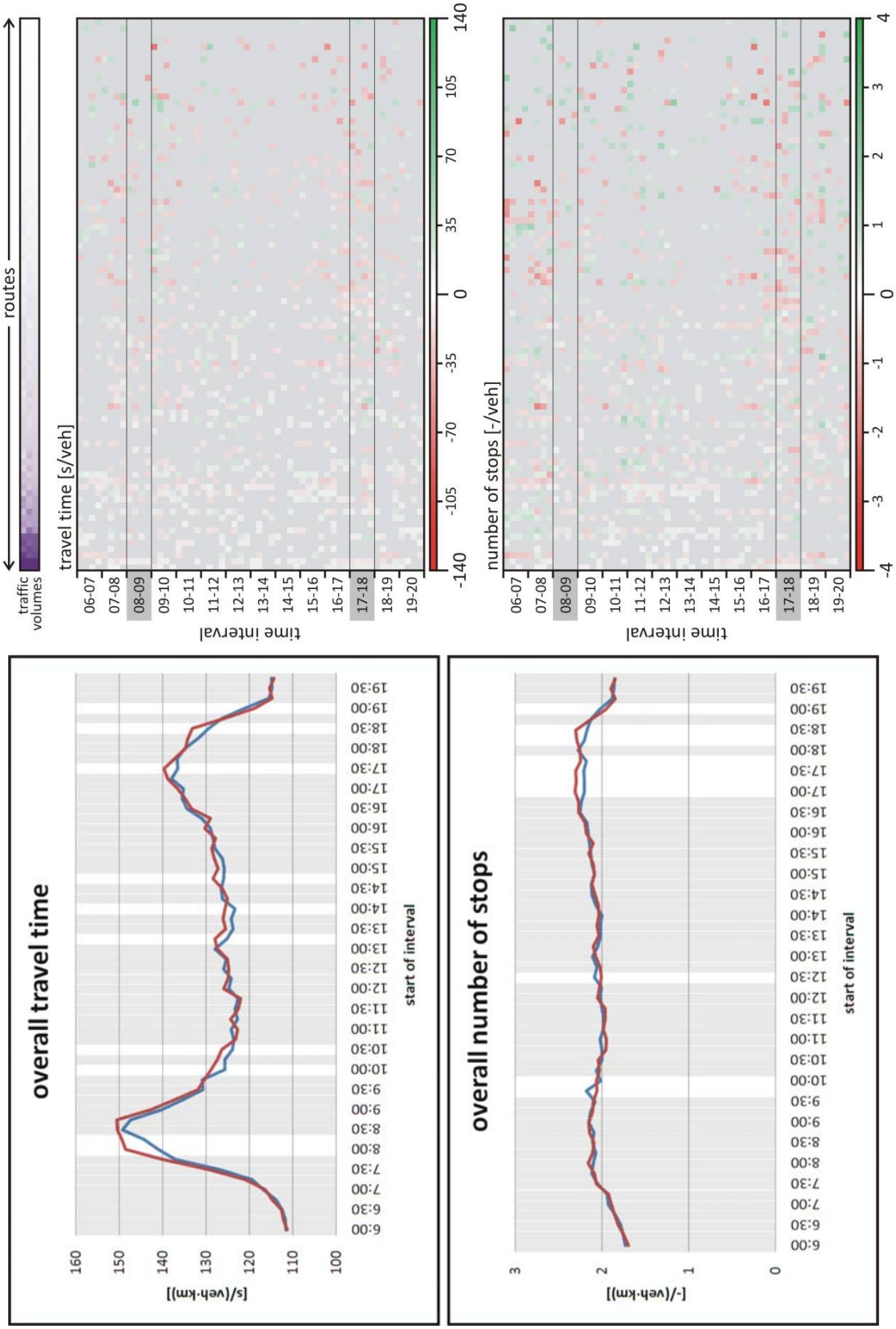


Figure D-51: ATCS, saturation based cycle lengths, PGA, exact demand
ATCS, saturation based cycle lengths, PGA, online optimization

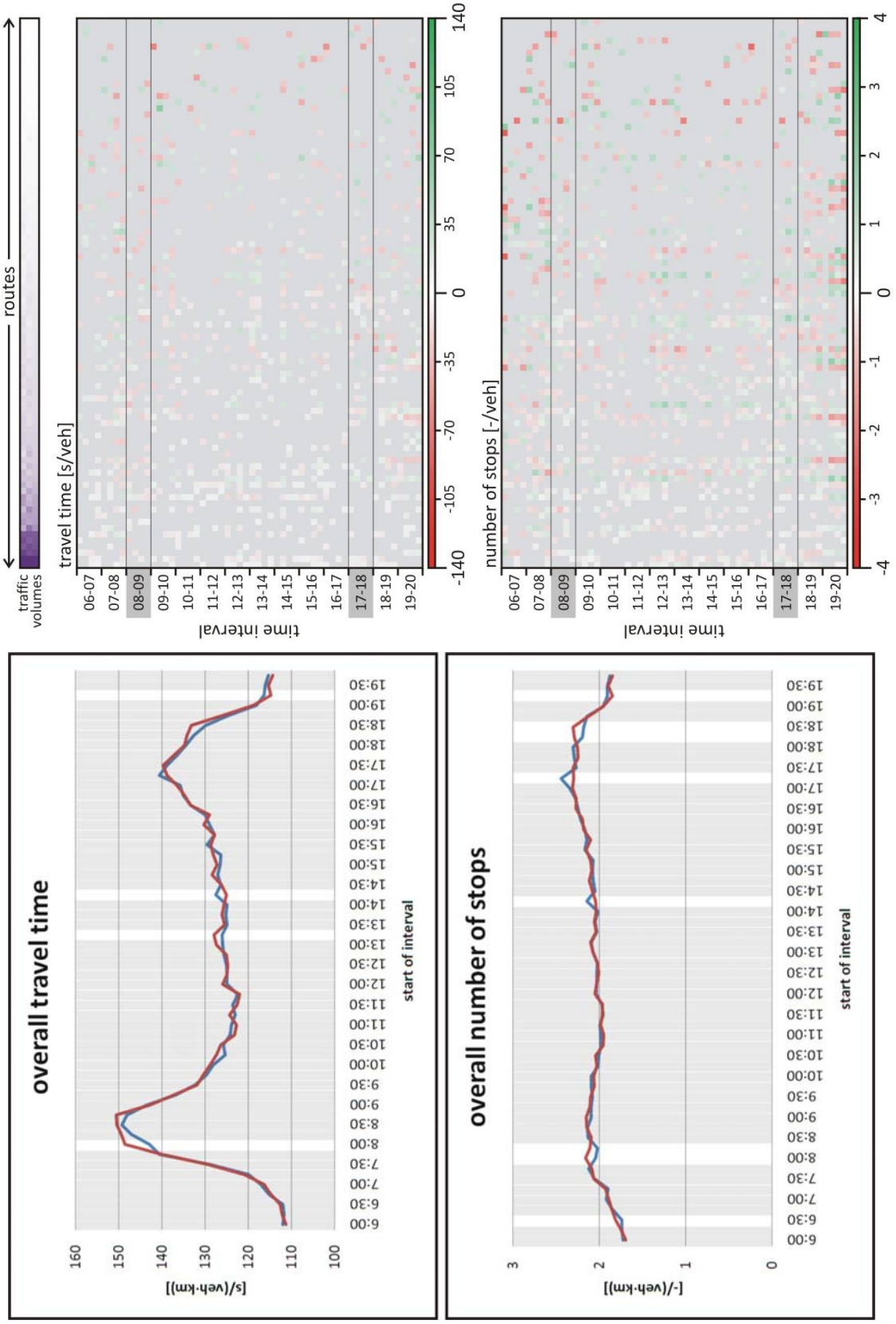
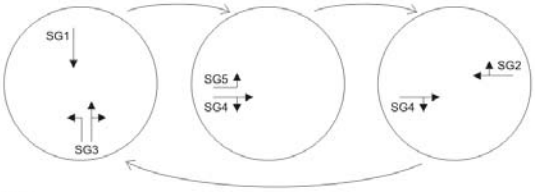


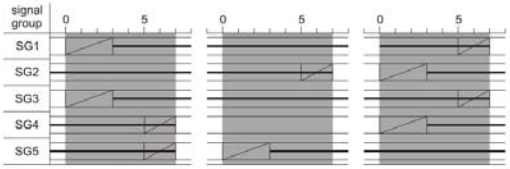
Figure D-52: ATCS, saturation based cycle lengths, PGA, estimated demand (BPR)
ATCS, saturation based cycle lengths, PGA, online optimization

E Phases and phase changes of the test networks

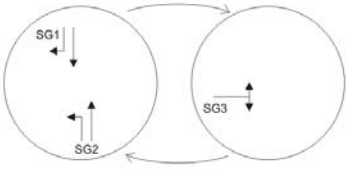
E.1 Hanover List network

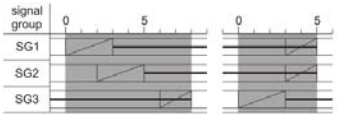
505



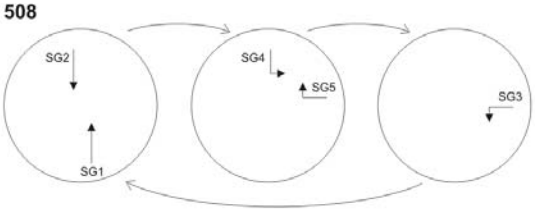


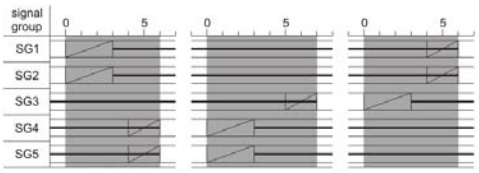
506



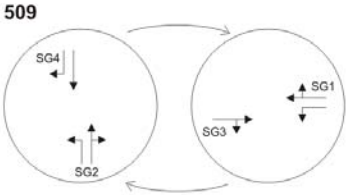


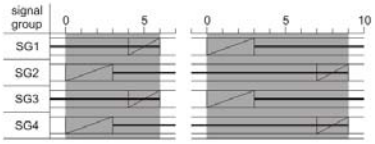
508



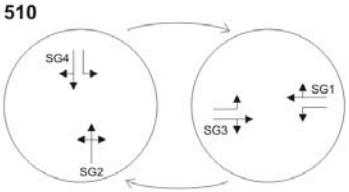


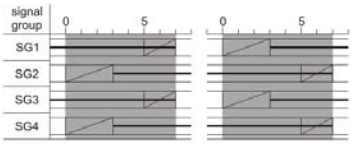
509



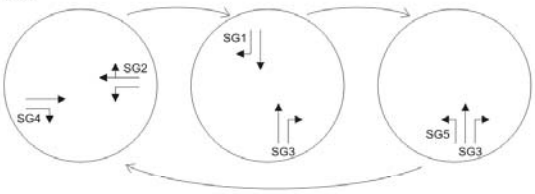


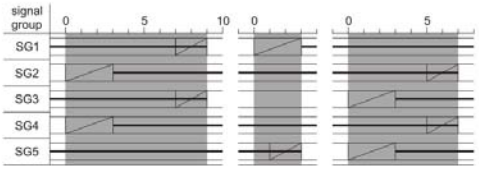
510



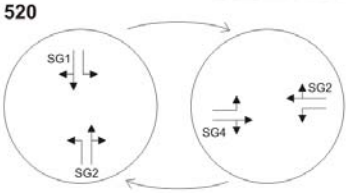


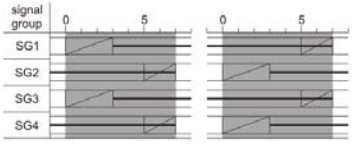
511



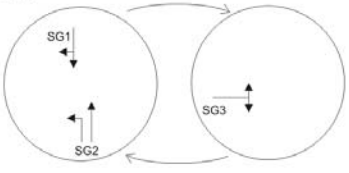


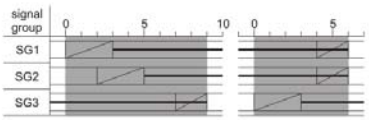
520





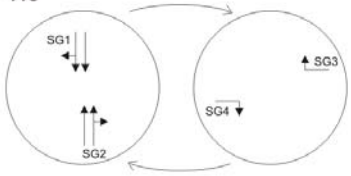
523

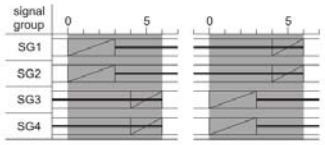




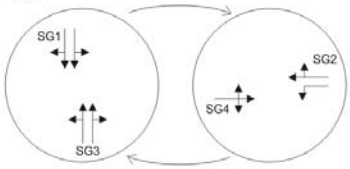
E.2 Hanover Südstadt network

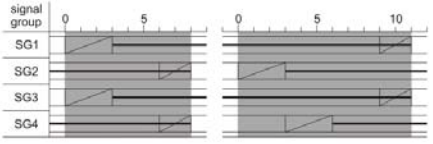
413



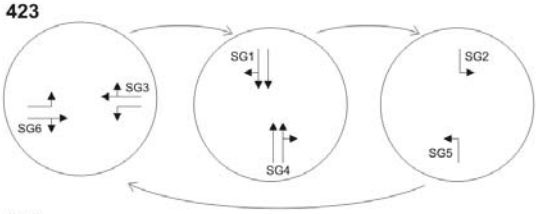


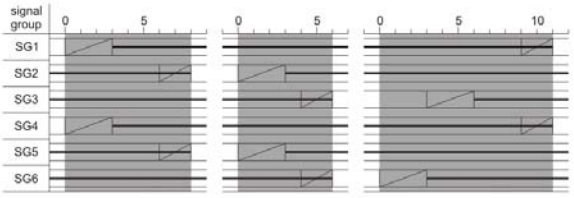
422



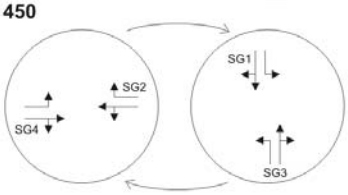


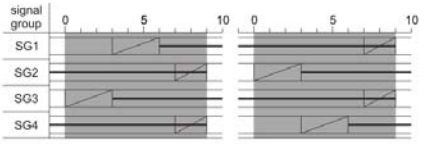
423



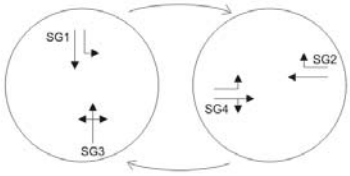


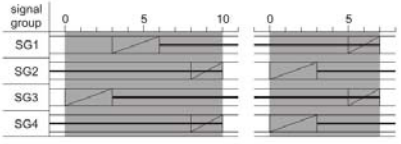
450





451





452