MEHDI KEYVAN-EKBATANI

REAL-TIME URBAN TRAFFIC CONTROL UNDER SATURATED TRAFFIC CONDITIONS

CRETE, GREECE
WINTER 2013

TECHNICAL UNIVERSITY OF CRETE

SCHOOL OF PRODUCTION ENGINEERING
AND MANAGEMENT

DYNAMIC SYSTEMS AND SIMULATIONS LABORATORY

REAL-TIME URBAN TRAFFIC CONTROL UNDER SATURATED TRAFFIC CONDITIONS

Thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

MEHDI KEYVAN-EKBATANI
Crete, Greece, December 2013

REAL-TIME URBAN TRAFFIC CONTROL UNDER SATURATED TRAFFIC CONDITIONS

Mehdi Keyvan-Ekbatani

This PhD thesis is approved by:

Advisory committee:

Markos Papageorgiou (Supervisor)
Professor, School of Production Engineering and Management,Tech. University of Crete

Ioannis Papamichail (Co-supervisor)
Assistant Professor, School of Production Engineering and Management,Tech. University of Crete

Georgios Stavroulakis (Member of advisory committee)
Professor, School of Production Engineering and Management,Tech. University of Crete

Thesis committee:

Matthew Karlaftis
Associate Professor, Dep. of Civil Engineering, National Technical University of Athens

Nikolas Geroliminis
Assistant Professor, Dep. of Civil Engineering, **É**cole **P**olytechnique **F**édérale de **L**ausanne

Ioannis Nikolos
Associate Professor, School of Production Engineering and Management,Tech. University of Crete

Argiris Delis
Associate Professor, School of Production Engineering and Management,Tech. University of Crete

To my lovely parents Zara & Asghar

ACKNOWLEDGEMENTS

Finally, after 3.5 years of living far from my family, country and friends, I am leaving Greece with a lot of new and precious experiences. During this period, I had several good and also some bad days. Maybe without the endless support of my parents, specially my mother, I could not overcome those problems and the difficult days which I had faced during my stay in Greece and I would quit my studies. My parents gave me the power to continue and finish my PhD. I would like to thank them from the bottom of my heart and hope to compensate some day. I am also grateful to my dear brother Ali, grandparents, my uncles Majid and Ehsan and to my aunt Boshra.

Without any doubt, the only reason that I decided to continue my studies in Greece was to work under supervision of one of the top class professors in the field of Transportation Engineering, Prof. Markos Papageorgiou. I feel honored of being supervised by a professor like him and I wish I could learn even much more from him. Working with him was very interesting and I really enjoyed it. Besides being very strict in the research, he is very supportive and generous in sharing his knowledge and experience with his students. Once more, I would like to thank him for his support and beneficial advices during my PhD and I wish him a successful and healthy life. I thank also Dr. Ioannis Papamichail for his constant support and assistance.

During my PhD, I had also the chance to visit École Polytechnique Fédérale de Lausanne (EPFL), Switzerland and to work under supervision of one of the well-known professors in the field of Traffic engineering, Dr. Nikolas Geroliminis. I would like to thank him for his advices and support during my stay in EPFL and also for coming to my defense session.

One of the positive aspects of living in Greece was finding very good friends whom I could really count on. I am extremely grateful to my Brazilian friend and colleague Dr. Rodrigo Castelan Carlson and his kind wife Anamaria. Rodrigo and Anamaria helped me a lot during my stay in Greece and they were beside me not only in good days but also in the bad days which I had. I had also the opportunity to work with a very good friend, Dr. Anastasios Kouvelas. I would like to thank him for his kind supports and helps. I thank also all my colleagues at Dynamic Systems and Simulation Laboratory (DSSL), specially my great friend Dr. Diamantis Manolis, Athina Tymbakianaki, Natasha Spiliopoulou, Gerasimos Loutos and Filio Malandraki. I will miss also my good friends Dr. Nikos Spanoudakis and his wife and my very good friend Christos Katrinakis.

Last but not least, I would like to thank the European Commission, particularly NEARCTIS (Network of Excellence for Advanced Road Cooperative Traffic management in the Information Society) for funding my PhD and all the professors participated in my thesis committee.

Mehdi Keyvan-Ekbatani
Greece, December 2013

SHORT BIOGRAPHY

Mehdi Keyvan-Ekbatani was born in Tehran, Iran, in 1982. He finished his high school studies in one of the top schools in Tehran with an excellent score. He received his B.sc in the field of Civil Engineering from Azad University Tehran in 2006. In 2007, he was accepted for a master program in the field of Highway and Transportation Engineering with an excellent entrance score at the Department of Civil Engineering, Sharif University of Technology, Tehran, Iran. After receiving his M.sc. in 2010, he was admitted for a PhD position at Dynamic Systems and Simulation Laboratory (DSSL), Technical University of Crete, Greece.

Since 2010, he has been a research assistant at DSSL. DSSL is among the leading institutions in Intelligent Transportation Systems (ITS); the lab was awarded with the *IEEE ITS Institutional Lead Award* in 2011. In 2012, he visited Transport Systems Laboratory (LUTS), Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland as a visiting scholar. He is the author of several journal and scientific conference papers, and has been a reviewer for several journals and conferences. His main research interests include Traffic Flow Theory, Traffic Flow Control, Public Transport Systems, Transportation Engineering, and Road Design.

Mr. Keyvan-Ekbatani was the recipient of a scholarship for his PhD from NEARCTIS (Network of Excellence for Advanced Road Cooperative Traffic management in the Information Society), funded by European Commission (2010-2013). In 2012, he received the *Best Paper Award* for the paper “Congestion Control in Urban Networks via Feedback Gating” at the European Transportation Research Arena 2012. In addition, the same paper was a finalist for the *Eltis Award Europe 2012*.

Abstract of the Thesis presented to the Technical University of Crete as a partial fulfillment of the requirements for the degree of Doctor of Philosophy.

REAL-TIME URBAN TRAFFIC CONTROL UNDER SATURATED TRAFFIC CONDITIONS

Mehdi Keyvan-Ekbatani

December/2013

Supervisor: Professor Markos Papageorgiou
Keywords: Urban traffic congestion; Signal traffic control; Operational network fundamental diagram (NFD); Sparse real-time measurement; Gating; Feedback control; Heterogeneous urban networks; Multiple-concentric gating; Time-delayed gating control

In recent decades, mitigating the traffic congestion in urban road networks has been a crucial issue for both the research and the practical operations, which calls for the development and implementation of improved traffic signal control methods and techniques. In particular, the development of efficient and practicable real-time signal control strategies under saturated traffic conditions is a major challenge, as widely used strategies like SCOOT and SCATS are deemed less efficient under saturated traffic conditions.

A practical tool, frequently employed against over-saturation of significant or sensitive links, arterials or urban network parts, is gating. The idea is to hold traffic back (via prolonged red phases at traffic signals) upstream of the links to be protected from over-saturation, whereby the level or duration of gating may depend on real-time measurements from the protected links. The method is usually employed in an ad hoc way (based on engineering judgment and manual fine-tuning) regarding the specific gating policy and quantitative details, which may lead to insufficient or unnecessarily strong gating actions.

Recently, the reproducible relationship between flow and density occurring at the network level under certain conditions (e.g. homogeneous spatial distribution of the congestion) known as macroscopic or network fundamental diagram (MFD or NFD), has gained increased popularity. Although the NFD notion is still under investigation in various aspects, it can be exploited as a fruitful basis for derivation of urban signal control approaches.

In this thesis, the notion of NFD for urban networks is exploited to improve mobility in saturated traffic conditions via application of gating measures, based on an appropriate

simple feedback control structure. Different gating control strategies (i.e. single perimeter gating control by exploiting complete and reduced NFD, multiple-concentric gating control, perimeter traffic control via remote feedback gating) have been proposed and tested on realistic simulation scenarios of two urban networks (i.e. Chania, Greece and San Francisco, USA) successfully. In the investigated examples, feedback gating is demonstrated to lead to substantial improvements of travel delays, network throughput and travel time reliability.

ΡΥΘΜΙΣΗ ΦΩΤΕΙΝΗΣ ΣΗΜΑΤΟΔΟΤΗΣΗΣ ΑΣΤΙΚΟΥ ΔΙΚΤΟΥ ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ ΥΠΟ ΚΟΡΕΣΜΕΝΕΣ ΣΥΝΘΗΚΕΣ ΚΥΚΛΟΦΟΡΙΑΣ

Mehdi Keyvan-Ekbatani

Δεκέμβριος/2013

Επιβλέπων: Καθηγητής Μάρκος Παπαγεωργίου

Λέξεις κλειδιά: Αστική κυκλοφοριακή συμφόρηση, Έλεγχος φωτεινής σηματοδότησης, Θεμελιώδες διάγραμμα του δικτύου (NFD), Αραιά μέτρηση σε πραγματικό χρόνο, Ελεγχόμενη είσοδος, Έλεγχος με ανατροφοδότηση, Ετερογενή αστικά δίκτυα, Έλεγχος εισόδου πολλαπλών ομόκεντρων περιμέτρων, Έλεγχος εισόδου με χρόνους υστέρησης

Στις τελευταίες δεκαετίες, η ανάγκη μετριασμού της κυκλοφοριακής συμφόρησης στα αστικά οδικά δίκτυα έχει γίνει ένα κρίσιμο θέμα τόσο για την έρευνα όσο και την πρακτική εφαρμογή, το οποίο καλεί για την ανάπτυξη και την εφαρμογή βελτιωμένων μεθόδων και τεχνικών ελέγχου φωτεινής σηματοδότησης. Ειδικότερα, η ανάπτυξη ικανοποιητικών και εφαρμόσιμων στρατηγικών ελέγχου φωτεινής σηματοδότησης σε πραγματικό χρόνο υπό κορεσμένες συνθήκες είναι μια μεγάλη πρόκληση, καθώς ευρέως διαδεδομένες στρατηγικές που χρησιμοποιούνται έως τώρα όπως το SCOOT και το SCATS θεωρούνται λιγότερο αποτελεσματικές υπό κορεσμένες συνθήκες.

Ένα πρακτικό εργαλείο, συχνά εφαρμόσιμο κατά του υπερκορεσμού σημαντικών ή ευαίσθητων συνδέσμων, αρτηριών ή περιοχών του αστικού δικτύου, είναι η ελεγχόμενη είσοδος (gating). Η ιδέα είναι να κρατηθεί η κυκλοφορία (μέσω παρατεταμένων κόκκινων φάσεων στους φωτεινούς σηματοδότες) ανάντη των συνδέσμων που πρέπει να προστατευτούν από υπερκορεσμό, ενώ το επίπεδο ή η διάρκεια της ελεγχόμενης εισόδου μπορεί να εξαρτάται από τις μετρήσεις σε πραγματικό χρόνο από τους προστατευόμενους συνδέσμους. Η μέθοδος αυτή εφαρμόζεται συνήθως για εξειδικευμένα και όχι γενικά προβλήματα (με βάση την εμπειρία και πειραματισμούς) ανάλογα με τη συγκεκριμένη πολιτική ελέγχου εισόδου και τις ποσοτικές λεπτομέρειες, τα οποία όμως μπορεί να οδηγήσουν σε ανεπαρκείς ή υπερβολικές δράσεις ελεγχόμενης εισόδου.

Πρόσφατα, η αναπαραγόμενη σχέση μεταξύ της ροής και της πυκνότητας ενός δικτύου υπό ορισμένες συνθήκες (π.χ. ομογενής χωρική κατανομή της συμφόρησης) γνωστό ως μακροσκοπικό θεμελιώδες διάγραμμα (MFD) ή θεμελιώδες διάγραμμα του δικτύου (NFD), έχει αποτελέσει αντικείμενο έρευνας. Παρόλο που η έννοια και ιδιότητες του NFD είναι

ακόμα υπό διερεύνηση ως προς διάφορες πτυχές, το NFD μπορεί να αξιοποιηθεί ως μια γόνιμη βάση για την παραγωγή προσεγγίσεων του ελέγχου φωτεινής σηματοδότησης σε αστικά δίκτυα.

Στην παρούσα διατριβή, η έννοια του NFD για αστικά δίκτυα χρησιμοποιείται με σκοπό τη βελτίωση της κινητικότητας σε κορεσμένες κυκλοφοριακές συνθήκες μέσω της εφαρμογής της ελεγχόμενης εισόδου, με βάση μια κατάλληλη απλή δομή ελέγχου ανάδρασης. Διάφορες στρατηγικές ελέγχου εισόδου (όπως έλεγχος εισόδου σε μια περίμετρο αξιοποιώντας πλήρες και μειωμένο NFD, έλεγχος εισόδου πολλαπλών ομόκεντρων περιμέτρων, περιμετρικός έλεγχος εισόδου μέσω απομακρυσμένων πυλών ανάδρασης) προτείνονται και διερευνώνται σε ρεαλιστικά σενάρια προσομοίωσης δύο αστικών δικτύων (Χανιά, Ελλάδα, και Σαν Φρανσίσκο, ΗΠΑ) επιτυχώς. Στα παραδείγματα που μελετήθηκαν, η ελεγχόμενη είσοδος με ανατροφοδότηση οδηγεί σε ουσιαστικές βελτιώσεις στις καθυστερήσεις κατά την διάρκεια του ταξιδίου και αύξηση της συνολικής ροής στο δίκτυο.

Contents

List of Figures

List of Tables

Chapter 1

If we knew what it was we were doing,
it would not be called research, would it?

Albert Einstein

1. Introduction

This first chapter introduces the reader to the problem under study. Section 1.1 presents the problem statement of this work and Section 1.2 the thesis goals. In Section 1.3 an outline of the document is provided. Finally, in Section 1.4 publications related to the research of this thesis are presented.

1.1. Problem Statement

Traffic congestion in urban road networks is a persisting or even increasing problem of modern society. Congestion can be reduced either by increasing road capacity (supply), or by reducing traffic demand. On the supply side, the provision of new infrastructure is usually not a feasible solution, hence it is necessary to focus on a better utilization of the existing infrastructure (e.g. via traffic management), to mitigate congestion and improve urban mobility. The field of urban traffic control (UTC) has been studied and developed in a variety of ways during the past decades. In fact, the traffic flow conditions in large-scale urban networks depend critically on the applied signal control strategies. However, as the debate regarding urban mobility and the wish for a sustainable transport system indicate, the negative effects of congested transport networks, such as excessive delays, environmental impact and reduced safety, persist or even increase; hence, introducing improved traffic signal control methods and techniques continues to be a vital issue.

In particular, the development of practicable and efficient real-time signal control strategies for urban road networks under saturated traffic conditions is a major challenge with significant scientific and practical relevance. The scientific relevance stems from the recently increased interest in the specific problem as well as recent, potentially valuable, models and insights that may contribute to improved signal control methods. The practical relevance stems from the congestion, degradation and gridlock problems encountered increasingly in modern urban road networks that could benefit highly from improved signal control under saturated traffic conditions.

Traffic signal control for urban road networks has been an area of intensive research efforts for several decades, and various algorithms and tools have been developed, proposed or implemented to increase the network traffic flow efficiency. However, most of these strategies face limitations when it comes to saturated traffic conditions that are frequently observed in modern metropolitan areas. Despite the continuous advances in the field of traffic control under saturated conditions, novel and promising developments of simple concepts in this area remains a significant objective, because some reported approaches that are based on various meta-heuristic optimization algorithms, can hardly be used in a real-time environment. In fact, a recent FHWA report (as cited in [36]) opined: “No current generally available tool is adequate for optimizing [signal] timing in congested conditions”.

A practical traffic management tool, frequently employed against over-saturation of significant or sensitive links, arteries or urban network parts, is gating or perimeter control. The idea is to hold traffic back (via prolonged red phases at traffic signals) at the perimeter of the zone (i.e. the protected network) or upstream of the links to be protected from over-saturation, whereby the level or duration of gating may be pre-fixed or may depend on real-time measurements from the protected links. The method is usually employed in an ad hoc way (based on engineering judgment and manual fine-tuning) regarding the specific gating policy and quantitative details, which may readily lead to insufficient or unnecessarily strong gating actions.

1.2. Objectives and Contributions

In view of the aforementioned difficulties, this thesis proposes and investigates a generic feedback-based gating concept which exploits the recently developed notion of network fundamental diagram (NFD) to improve mobility in saturated traffic conditions.

To this end, the following topics are addressed, which reflect the thesis' main contributions:

- A control-design model exploiting the NFD and an appropriate feedback-based gating regulator was developed for the first time in the frame of the present thesis research.
- A thorough study on NFDs derived with subsets of links in the network has been carried out.
- It is demonstrated that an efficient feedback-based gating is possible with much less real-time measurements. This is a significant achievement that opens the way for real implementations of the method due to the substantially reduced cost implied.
- Given that in large metropolitan urban networks the congestion spreads mostly heterogeneously over the network, in which case a homogenous gating strategy may not be the optimal solution, in a cooperation with Prof. Nikolas Geroliminis and his PhD student Mehmet Yildirimoglu at **École Polytechnique Fédérale de Lausanne (EPFL)**, Switzerland, a multiple concentric boundary gating strategy is introduced which implements the aforementioned feedback-based gating strategy, along with considering the heterogeneity of a large-scale urban network.
- Different methods for the flow splitting at the gated junction are proposed and applied in the microscopic simulation environment.
- When the gated link is not directly at the protected network (PN) perimeter, a travel time is needed for gated vehicles to approach the PN. Thus, a robust feedback controller, by considering this imposed time-delay on the system, is designed.
- In the case of gating remote from PN, it is shown that the feedback gating works properly with much longer time-steps.

1.3. Thesis Outline

The first introductory (current) chapter presents the motivation and the thesis goals and contributions. The outline of the rest of the document is as following:

- Chapter 2 delivers a preliminary discussion of congestion, its causes and resulting infrastructure degradation, followed by an overview of studies on urban traffic control (UTC) strategies and existing literature regarding various control algorithms. Later on, a historical overview of macroscopic fundamental diagram (MFD) and some recent studies on two-dimensional urban road networks and its application to UTC are

presented. In this chapter, the general gating scheme and a background of this concept is also discussed. In addition, some preliminaries of traffic signal operation are presented. Finally, the investigated studies in this thesis are introduced.

- In Chapter 3, at the beginning, AIMSUN simulator is briefly introduced. Moreover, different simulation scenarios implemented as test-bed in this thesis are presented.
- Chapter 4 starts with an introduction on Single-Input-Single-Output (SISO) Systems and feedback controllers. Moreover, the system modeling for feedback design and the controller design of the single and multiple perimeter control are presented. Next, feedback controller design in presence of time-delay is proposed. Finally, the gating action at the gated links is discussed in details.
- Chapter 5 demonstrates the results of different control scenarios in this thesis. The simulation results of single perimeter control (by exploiting complete and reduced operational NFD), multiple boundaries gating control and gating remote from the PN are presented, separately. At the end, the results of gating control with bigger control steps in the case of remote gating junctions are illustrated and discussed.
- Finally, Chapter 6 concludes this thesis and comments on further research.

1.4. Publications Related to Thesis

The work presented in this thesis resulted in several scientific publications listed as follows:

Journals

- Keyvan-Ekbatani, M., Kouvelas, A, Papamichail, I. & Papageorgiou, M., 2012. Exploiting the Fundamental Diagram of Urban Networks for Feedback-Based Gating, Transportation Research Part B 46 (10), 1393–1403.
- Keyvan-Ekbatani, M., Papageorgiou, M., Papamichail, I., 2013. Urban Congestion Gating Control based on Reduced Operational Network Fundamental Diagrams. Transportation Research Part C 33, 74-87.
- Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., 2013. Traffic Signal Perimeter Control with Multiple Boundaries for Large Urban Networks (in preparation)

- Keyvan-Ekbatani, M., Papageorgiou, M., 2013. Controller Design for Gating Control in Presence of Time-Delay in Urban Road Networks (in preparation).

Conferences

- Keyvan-Ekbatani, M., Kouvelas, A, Papamichail, I. & Papageorgiou, M., 2012. Congestion Control in Urban Networks via Feedback Gating, Procedia- Social and Behavioral Sciences, 48, 1599-1610.
- Keyvan-Ekbatani, M., Papageorgiou, M. & Papamichail, I., 2013. Feedback Gating Based on Sparse-Measurement Urban Network Fundamental Diagrams, Transportation Research Board 92nd Annual Meeting Washington D.C, USA.
- Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., 2013. Traffic Signal Perimeter Control with Multiple Boundaries for Large Urban Networks, IEEE ITSC 2013, Den Haag, the Netherlands.
- Keyvan-Ekbatani, M., Papageorgiou, M., Papamichail, I., 2013. Perimeter Traffic Control via Remote Feedback Gating, 16th Meeting of the EURO Working Group on Transportation, Porto, Portugal.

Workshops

- Keyvan-Ekbatani, M., Kouvelas, A, Papamichail, I. & Papageorgiou, M., Congestion control in urban networks under saturated conditions via feedback gating. 4th NEARCTIS Workshop, IFSTTAR, Lyon, France, June 10, 2011.
- Keyvan-Ekbatani, M., Papageorgiou, M., Papamichail, I., Real-time Urban Traffic Control under Saturated Conditions, NEARCTIS final event, Trinity College, Jun. 2013, Dublin, Ireland.

Chapter 2

One faces the future with one's past.

Pearl S. Buck

2. Background and Considerations

This chapter begins with an introduction of Urban Traffic Control (UTC). Then, the notion of Network Fundamental Diagram (NFD) and its existing literature is discussed in section 2.2. In Section 2.3, the control strategies which exploit the notion of NFD for UTC are presented. Right after, the definition of gating and its general scheme is introduced in section 2.4. In section 2.5, application of feedback controller for gating control strategy is discussed briefly. In section 2.6, some preliminaries of traffic signal operation are presented. Finally, the studied gating strategies in this thesis are introduced in section 2.7.

2.1. Urban Traffic Control (UTC)

Traffic congestion is a growing problem in most urban areas across the world. In recent years, the problem has often been tackled by management of existing capacity rather than the traditional concept of more road building. This requires efficient traffic management tools and has led to widespread implementation of advanced traffic control systems. The objective of UTC has traditionally been to implement signal timings that minimize the total vehicular delay in the network.UTC systems constitute a scientific field with long-lasting and extensive research and development activities. Many methodologies have been proposed so far, but there is still space for new developments, particularly for saturated traffic conditions. SCOOT (Split Cycle Offset Optimization Technique) [30], which is an adaptive system that responds automatically to fluctuations in traffic flow, is the most common UTC system used in the

UK. In fact, widely used strategies like SCOOT and SCATS [39], although applicable to large-scale networks, are deemed less efficient under saturated traffic conditions. On the other hand, more advanced traffic-responsive strategies like OPAC [23], PRODYN [18], and RHODES [44] use optimization algorithms with exponential increase of complexity, which do not permit a straightforward central network-wide application. Thus, most available strategies face limitations when it comes to saturated traffic conditions that are frequently observed in modern metropolitan areas. [9] proposed a dynamic method to control an oversaturated traffic signal network by utilizing a bang-bang like model for the oversaturated intersections. A noteworthy and practicable attempt to address saturated traffic conditions was the more recently developed signal control strategy TUC [15], see also [3]. Aboudolas et al. [3] formulated the problem of network-wide signal control as a quadratic-programming problem that aims at minimizing and balancing the link queues so as to minimize the risk of queue spillback. Furthermore, a number of research approaches have been proposed, that employ various computationally expensive numerical solution algorithms, including genetic algorithms ([4]; [38]), multi-extended linear complementary programming [14], mixed-integer linear programming ([37]; [6]) and ant colony optimisation [51]; however, in view of the high computational requirements, the network-wide implementation of these optimization-based approaches might face some difficulties in terms of real-time feasibility.

2.2. Network Fundamental Diagram (NFD)

In the past 40 years, various theories were proposed to describe vehicular traffic on an aggregated level [26]. Recently, it was found that the notion of a fundamental diagram (e.g. in the form of a flow-density curve, see Figure 2-1) can be applied (under certain conditions) to two-dimensional urban road networks as well. In the two-dimensional fundamental diagram (i.e. summation of weighted flow versus average density), as the accumulation or density (k) is increasing (the network operates with free flow condition, shown by green), the traffic flow increases up to the yellow region where capacity of the network is reached (see Figure 2-1). Exceeding the range with critical density (i.e. yellow zone), the network starts degrading and enters the over-saturated region indicated by the red rectangle. The aim of the traffic experts is to maintain the overall traffic state, by applying various traffic management tools (e.g. traffic signal optimization, gating, route guidance and etc.) in the yellow range and to avoid spillover and gridlock creation (shown by dark red in which the flow is close to zero

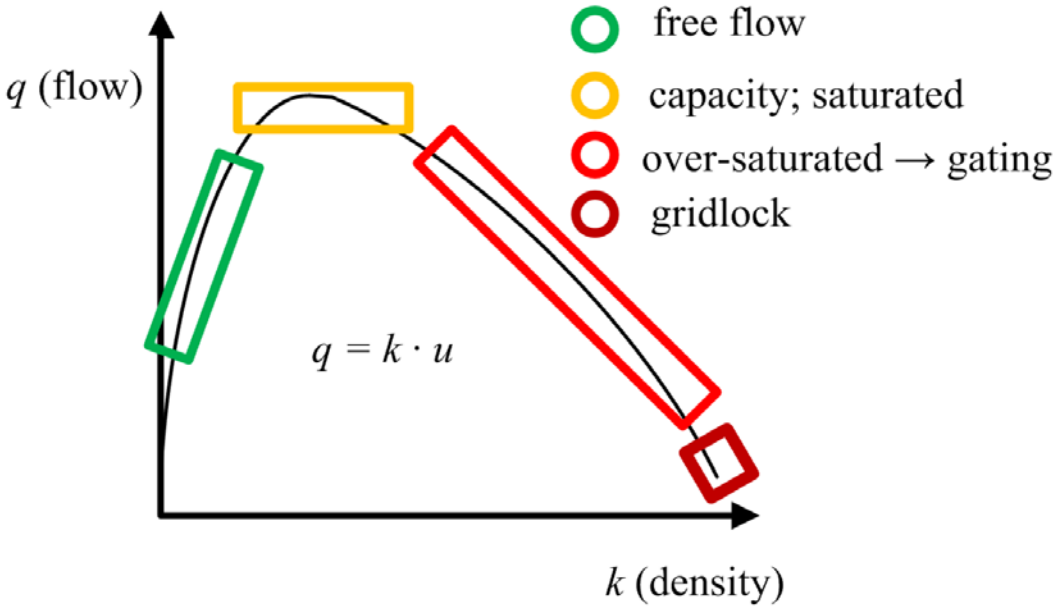

Figure 2-1 Network Fundamental Diagram (network total weighted flow vs. average density)

in Figure 2-1). In this thesis, gating control is implemented as an efficient traffic management tool to keep the traffic state in the yellow region.

In fact, a fundamental-diagram-like shape of measurement points was first presented in [21]; re-initiated later by Mahmassani et al. [42] and Daganzo [11], but also observed in a field evaluation study by Dinopoulou et al. [16], see Fig. 6 therein. The concept is sometimes called MFD (macroscopic fundamental diagram), but since the ordinary fundamental diagram (for highways) is also macroscopic, we prefer to call it NFD (network fundamental diagram) for better distinction. Recent studies on this notion can be found in [24] for simulation-based experiments; [26] for real-data based investigations; [12], [19] and [29] for analytical treatments.

The principle requirement for the well-defined NFD is the homogeneity of the area-wide traffic conditions, which is not universally expected in the real world. Buisson and Ladier [8] further investigated the NFD shape in heterogeneous environments. Based on the analyses, carried out using a real data-set from Toulouse in France, they figured out that network types and unusual events, such as incidents, have a strong impact on the NFD shape. In order to further clarify the necessary condition for well-defined NFD, Mazloumian et al. [43] have identified that the spatial distribution of link densities is the key factor for defining the NFD shape. The findings suggest that NFD can be applied for unevenly congested network if the

network can be partitioned in homogeneous zones. Based on these finding, Ji and Geroliminis [32] investigated the methodology of the network partitioning into compact shape zones, where well-defined NFD was expected and perimeter control can be applied based on the NFD concept.

Although the exact NFD curve may depend on the origin-destination demand [31], it may be quite stable from day to day, particularly if the traffic load is homogeneously distributed in network links [27]. In simulated environments, where different signal control strategies are tested, this homogeneity condition may call for activation of a dynamic traffic assignment device to reduce possible transient phenomena, such as a hysteresis between network filling and emptying data ([3]; [27]). Mahmassani and Peeta [41] compared network performance under user equilibrium (UE) and system optimal (SO) dynamic assignments. They investigated a hysteresis pattern, in the average network density and average network speed relationship by using simulation data for a network that included both freeways and urban street arterials. They found that for the same value of the average network density, the average network speed is lower during the recovery period compared to the loading period. Recent studies ([53], [25]) found that if traffic is distributed heterogeneously in a network, for the same value of average network density, the average network flow is higher during the loading period compared to the recovery period.

A network fundamental diagram may be an *ideal* NFD, if based on *exact* knowledge of the displayed quantities (this is practically only possible in analytic or simulation-based studies) for *all* links $z \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of all network links; or an *operational* NFD, if based on available (more or less accurate) measurements and estimates at a subset $\mathbb{M}$ of all links, i.e. $\mathbb{M} \subseteq \mathbb{Z}$; an operational NFD is called *complete*, if the measurements cover all network links, i.e. if $\mathbb{M} = \mathbb{Z}$.

The operational NFD's y-axis reflects the Total Travelled Distance (*TTD* in veh·km per h), while the x-axis reflects the Total Time Spent (*TTS* in veh·h per h) by all vehicles in the PN. *TTD* and *TTS* are obtained from the loop measurements via the following equations:

$$TTS(k) = \sum_{z \in \mathbb{M}} \frac{T \cdot \hat{N}_z(k)}{T} = \sum_{z \in \mathbb{M}} \hat{N}_z(k) = \hat{N}(k) \tag{2-1}$$

$$TTD(k) = \sum_{z \in \mathbb{M}} \frac{T \cdot q_z(k) \cdot L_z}{T} = \sum_{z \in \mathbb{M}} q_z(k) \cdot L_z \tag{2-2}$$

where z is the link where a measurement is collected; $\mathbb{M}$ is the set of measurement links, here $\mathbb{M} = \mathbb{Z}$; k = 0, 1, 2,… is a discrete time index reflecting corresponding cycles; T is the cycle time; q_z is the measured flow in the link z during cycle k; L_z is the length of link z; and $\hat{N}_z(k)$ is the estimated number of vehicles in link z during cycle k, which is derived from measured occupancy measurements via the following equation

$$\hat{N}_z(k) = L_z \cdot \frac{\mu_z}{100\lambda} \cdot o_z(k-1) \tag{2-3}$$

where o_z is the measured time-occupancy (in %) in link z during cycle k; μ_z is the number of lanes of link z; and λ is the average vehicle length. Eq. (2-3) is reasonably accurate, particularly if the detector is located around the middle of the link [46]. According to the derivations in eq. (2-1) and (2-2), *TTS* equals the number of vehicles in all PN links equipped with detectors; while *TTD* is a length-weighted sum of the corresponding network link flows.

2.3. Exploiting the NFD for UTC

The NFD concept for urban road networks has been an issue of intensive investigations recently; indeed, the conditions under which it appears, the stability of its shape under different O-D patterns or at different peak periods or days-of-the-week, the impact of different signal control strategies, the possible hysteresis between the network filling and emptying phases, are still under the loop of ongoing analytical or empirical investigations and research. Nevertheless, based on what is known or observed in data, it is not too early for the NFD concept to be considered as a basis for the derivation of traffic control strategies.

Daganzo [11] used the NFD concept to propose a control rule that maximizes the network outflow; however, as discussed later, that rule cannot be directly employed for practical use in urban networks. Based on the NFD, Zhang et al. [61] developed a Bang-Bang control strategy to optimize the accumulated vehicular number. Other works ([28]; [55]) pursued a model-predictive control (MPC) approach. However, MPC calls for sufficiently accurate model and external disturbance predictions, which may be a serious impediment for practicable control. In fact, Geroliminis et al. [28] tested the MPC concept only on the basis of the same simple model used within the optimal control problem; while Strating [55] used detailed microscopic simulation, but reported a failure to produce sensible control results. Lia et al. [35] introduced a fixed-time signal timing perimeter control by exploiting the NFD,

albeit without adaptation to the prevailing real-time traffic conditions. Aboudolas and Geroliminis [2] applied a feedback approach for application to multiple sub-networks with separate individual NFDs in a heterogeneous urban network. However, the boundaries of sub-networks are static and congestion propagation might violate the homogeneity of regions. Zheng et al. [62] introduced a three-dimensional NFD relating the accumulation of cars and buses with the total circulating flow in the network. They addressed that this finding may be applied for a perimeter flow control concept which maximizes the network capacity of vehicles or the passenger capacity in bi-modal urban networks. However, the concept of 3D-NFD might not be applicable in the case of real data. The interactions between the different bus lines and the surface traffic, increased dwell times due to various reasons (e.g. high number of passengers in some bus-stops) and generally additional imposed delay on the system may be such that a well-shaped NFD would not be possible to derive. Aboudolas et al. [1] proposed a proportional controller for a perimeter control strategy by exploiting the 3D-NFD concept. This control strategy might be beneficial by combining the 3D-NFD with idea of reduced NFD (i.e. few real-time measurements) which will be discussed in Chapter 5, thoroughly.

2.4. Gating Control

A practical tool, frequently employed against over-saturation of significant or sensitive links, arteries or urban network parts, is gating ([60]; [7]; [40]). The idea is to hold traffic back (via prolonged red phases at traffic signals) upstream of the links to be protected from over-saturation, whereby the level or duration of gating may depend on real-time measurements from the protected links. The method is usually employed in an ad hoc way (based on engineering judgment and manual fine-tuning) regarding the specific gating policy and quantitative details, which may readily lead to insufficient or unnecessarily strong gating actions.

The objective of this thesis is to mitigate urban traffic congestion via gating, by exploiting the notion of the network fundamental diagram (NFD) for an urban network part that needs to be protected from the detrimental effects of over-saturation.

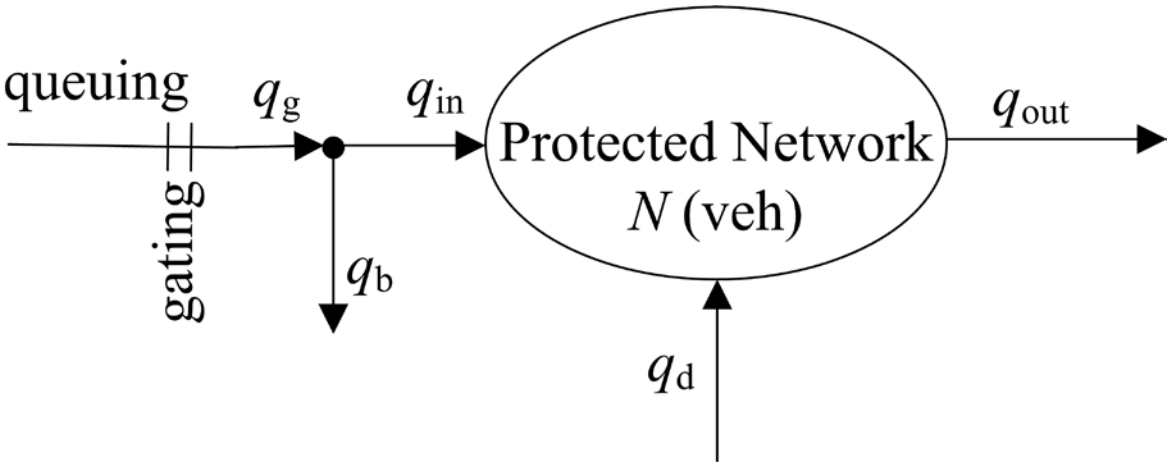

Figure 2-2 General scheme of the protected network and gating

2.4.1. General Scheme

To gate the traffic flow (usually during the peak periods) in an urban network, the area to be protected from possible congestion and the locations where gating queues will be created, must be defined. The general scheme of gating, including the protected network (PN), is sketched in Figure 2-2 To implement gating, the usual traffic lights settings must be modified at (one or more) upstream junctions, which may be located more or less closely to the problematic area. In Figure 2-2, the double line indicates the gating location, upstream of which vehicle queues may grow temporarily faster than without gating; q_g is the gated flow, a part of which (q_b) may not be bound for the protected network (PN); while q_{in} is the part of the gated flow that enters the protected network; q_d represents other (non-gated or internal) inflows to the PN (disturbances); finally q_{out} and N stand for the PN exit flow (both internal and external) and the number of vehicles included in the PN, respectively.

If N is allowed to grow beyond certain limits (i.e. the yellow zone (capacity) in Figure 2-1) the PN exit flow q_{out} decreases (entering the red zone in NFD shown in Figure 2-1) due to link queue spillovers and gridlock. To avoid this PN degradation, gating should reduce the PN inflow q_{in} appropriately, so as to maximize the PN throughput (by maintaining the traffic state in the capacity level shown by yellow in Figure 2-1). This may incur some temporary vehicle delays in the queues of the gated junctions, which, however, may be eventually offset (at least for the q_{in} portion of the gated flow) thanks to the higher PN exit flow enabled by gating; on the other hand, the flow q_b will experience gating delays without any direct

reward; these delays will be generally smaller if the gating junction is closer (or attached) to PN, due to accordingly smaller (or zero) flows q_b. Overall, gating will be beneficial if the saved delays in the PN are higher than the unnecessary delays incurred to the q_b portion of the gated flow. In some situations, e.g. when major problems in PN causes congestion to spread rapidly to adjoining areas, the use of gating could provide even higher benefits to the overall network.

2.5. Applying Feedback Controller for Gating Purpose

Field implementations call for simple and efficient systems that would expedite their application. Thus, in this PhD thesis, a simple feedback controller has been applied for the gating strategy by exploiting the notion of NFD and has been tested in different scenarios (will be discussed in Chapter 3) for various studies (see section 2.7), successfully. As demonstrated in Figure 2-3, the real-time measurement is fed into the controller and by defining a pre-specified reference value $TT\hat{S}$ (i.e. a value inside the range of the critical range of NFD, see the yellow zone in Figure 2-1) for the controller, the *TTS* or *N* in PN is kept close to the $TT\hat{S}$. More specifically, the feedback controller tries to regulate the number of vehicles in PN close to the reference value, according to the real-time traffic state in PN (*TTS*), by ordering flow q_g at the gated links. Further details regarding the implemented model, controller type and the designing procedure of the controller will be discussed in chapter 4.

2.6. Basics in Traffic Signal Operations

Signalized intersections play a critical role in the safe and efficient movement of vehicular and pedestrian traffic. The objective of traffic signal timing is to assign the right-of-way to alternating traffic movements in such a manner to minimize the average delay to any group of vehicles or pedestrians and reduce the probability of crash producing conflicts. This section describes the controller settings that define the operation of the signal when it is part of a signal system. The settings addressed include cycle length, offset, and phase sequence [52].

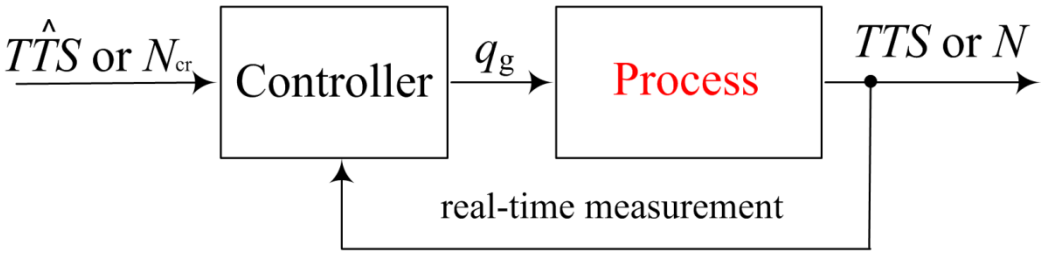

Figure 2-3 Feedback-based gating scheme

2.6.1. Phase

A phase is the part of the cycle assigned to a fixed set of traffic movements. As demonstrated in Figure 2-3, an intersection may have simply two phases or more. Phase and stage are the American and the British term utilized in traffic engineering, respectively.

2.6.2. Cycle Length

Cycle length is defined as the total time to complete one sequence of signalization to all movements at an intersection (see Figure 2-4). For an intersection with coordinated-actuated control, the cycle length is most easily measured as the time between two successive terminations of a given coordinated phase. For an intersection with pre-timed control, the cycle length is measured as the time between two successive starts (or terminations) of any given phase. Signals that are part of a signal system typically have the same system cycle length. However, some lower-volume signalized intersections may operate at a cycle length that is one-half the system value. The optimum cycle length for a given signal system will depend on its traffic volume, speed, intersection capacity, intersection phase sequence, and segment length. Analytic techniques that consider all of these factors typically reveal that the optimal cycle length for minor arterial streets and grid networks is in the range of 60 to 120 s. This range increases to 90 to 150 s for major arterial streets.

2.6.3. Offset

The offset for a signalized intersection is defined as the time difference between the intersection reference point and that of the system master. The intersection reference point is typically specified to occur at the planned start (or end) of the green interval for the first coordinated phase.

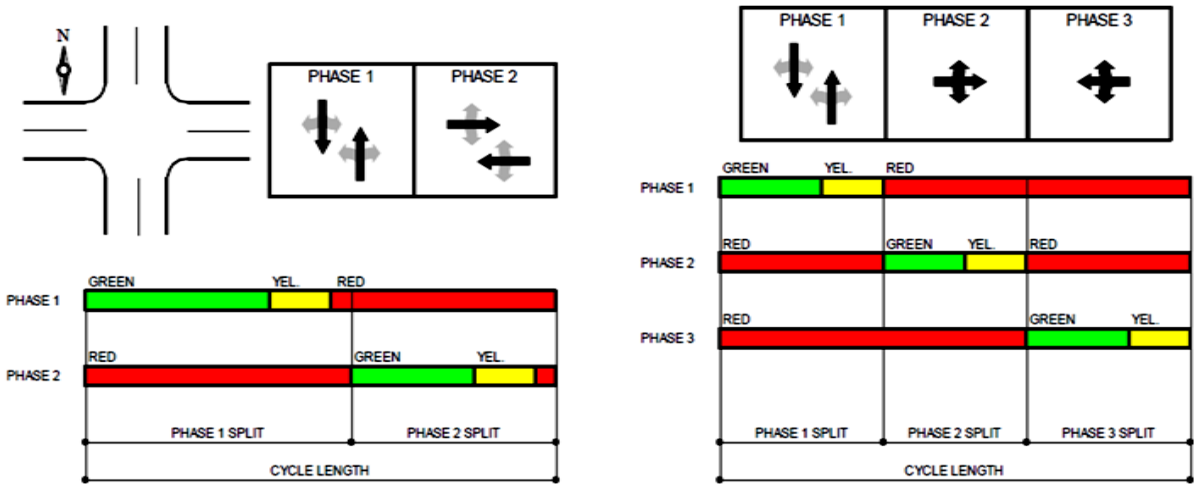

Figure 2-4 Two-phase (left) and three-phase (right) signal operation

The objective in selecting an offset for an under-saturated intersection is to provide a timely green indication for the platoon associated with the progressed movement. If there are queued vehicles in the lanes serving the progressed movement at the onset of green, then the offset should be set such that the green interval starts before the progressed movement arrives by an amount that is sufficient for the initial queue to clear before the platoon arrives.

The objective in selecting an offset for an over-saturated intersection is to minimize the adverse effect of bay overflow and spillback. In fact, other signal timing and capacity improvements are more effective than offset at achieving this goal. Regardless, an offset that facilitates throughput (i.e., progression away from the over-saturated intersection) should be sought when over-saturations are present.

2.6.4. Under-Saturated vs. Over-Saturated Conditions

The determination of which saturation category best describes a given traffic period is based on whether a queue is present at the end of the phase (i.e., an overflow queue). A signal phase that has a recurring overflow queue during the traffic period is referred to as oversaturated; otherwise, it is referred to as under-saturated. Similarly, an intersection that has a recurring overflow queue for all signal phases during the traffic period is referred to as over-saturated; otherwise, it is referred to as under-saturated.

2.6.5. Yellow Change and Red Clearance Interval

The yellow change interval is intended to alert a driver to the impending presentation of a red indication. This interval should have a duration in the range of 3 to 6 s.
The red clearance interval is optional. If not used, its value is 0 s. Non-zero values are used to allow a brief period of time to elapse following the yellow indication and during which the signal heads associated with the ending phase and all conflicting phases display a red indication.

2.7. Different Gating strategies Studied in this Dissertation

As addressed earlier, this thesis exploits the notion of the NFD for efficient feedback-based gating strategies (or gating scenarios). The proposed gating scenarios are studied in the following order in this thesis:

- *Study 1*: Single boundary gating control by exploiting *complete* operational NFD of PN (i.e. all measurements in PN).
- *Study 2*: Single perimeter gating control by applying *reduced* operational NFD of PN (i.e. few measurements in PN).
- *Study 3*: Multiple concentric boundaries gating control; the simulation-based results of this gating scenario are compared with the single-boundary gating results. Moreover a new flow distribution algorithm is implemented at the gated links.
- *Study 4*: Gating control remote form the PN (introducing feedback controller by considering time-delay term).
- *Study 5*: Remote gating control of PN (by increasing control step without considering time-delay term in the action of the controller).

The simulation scenarios which each of the aforementioned gating strategies are studies on, the methodological details and the results of these studies are presented in Chapters 3, 4 and 5, respectively.

Chapter 3

The question is not what you look at, but what you see.

Henry David Thoreau, writer

3. Study Environment

In this chapter, the microscopic simulator AIMSUN used for the simulation experiments is presented in section 3.1. Right after, a short description of the studied urban road network of the city of Chania, in Greece and San Francisco, USA, is provided. Finally, the simulation *scenarios* implemented in this thesis along with the modeling characteristics and the simulation features applied are introduced.

3.1. Microscopic Simulator AIMSUN

This section provides a short introduction to the commercial microscopic traffic simulator AIMSUN (TSS, 2008). AIMSUN is an extensible environment that offers, in a single application, all the tools needed by transportation professionals. The reader is referenced to [56] for details. In general, microscopic modeling of traffic flow considers the movement of individual vehicles in dependence of the movement of adjacent vehicles.

Most micro-simulation models use various algorithms and driver behavior models to simulate the movement of individual vehicles within a network. Each vehicle that enters the network is assigned a vehicle type (auto, truck, bus, etc.) and corresponding vehicle performance characteristics (acceleration, deceleration, speed, and turning characteristics). It is also assigned one of ten driver characteristics (ranging from aggressive to cautious), giving each vehicle a unique and realistic performance profile that it maintains while traveling through

the network. The position and speed of each vehicle on the network is updated once per simulation step based on its own performance and driver characteristics, the actions of vehicles around it, roadway properties, and traffic control devices. Thus, the interaction of vehicle to vehicle, vehicle to road, and vehicle to control devices are modeled accurately for each simulation. Default vehicle and driver characteristics can also be modified to better reflect actual traffic conditions for a given scenario.

Once a vehicle is assigned performance and driver characteristics, its movement through the network is determined by three primary algorithms: (a) car following algorithm, (b) lane changing algorithm and (c) gap acceptance algorithm. There are other algorithms which influence vehicle behavior too, such as those which govern queue discharge and traffic signal control, but these three are perhaps the most important and are common to all traffic simulation models.

AIMSUN (Advanced Interactive Micro-Simulation for Urban and non-urban Networks) is a full function microscopic simulation tool with a broad range of simulation capabilities. It can simulate surface street networks, motorways, interchanges, weaving sections, pre-timed and actuated signals, stop controlled intersections, and roundabouts. It also includes features about full trip distribution capabilities, dynamic traffic assignment, real-time vehicle guidance, and 3-D animations. AIMSUN is used in conjunction with the traffic network graphical editor (TEDI) and is part of the Generic Environment for Traffic Analysis and Modeling (GETRAM) open simulation environment.

Vehicle attributes such as length, width, maximum speed, and normal and maximum acceleration are assigned when a vehicle enters the network. Users can select from a wide variety of vehicle types, and within each type there will be some variation in these parameters based on statistical distributions. Within the vehicle stream there is variation of driver performance characteristics, such as desired minimum headway, turning speed and speed acceptance (obedience to the speed limit). AIMSUN establishes mean driver performance values and varies driver behavior for each vehicle about the mean (within specified minimum and maximum values).

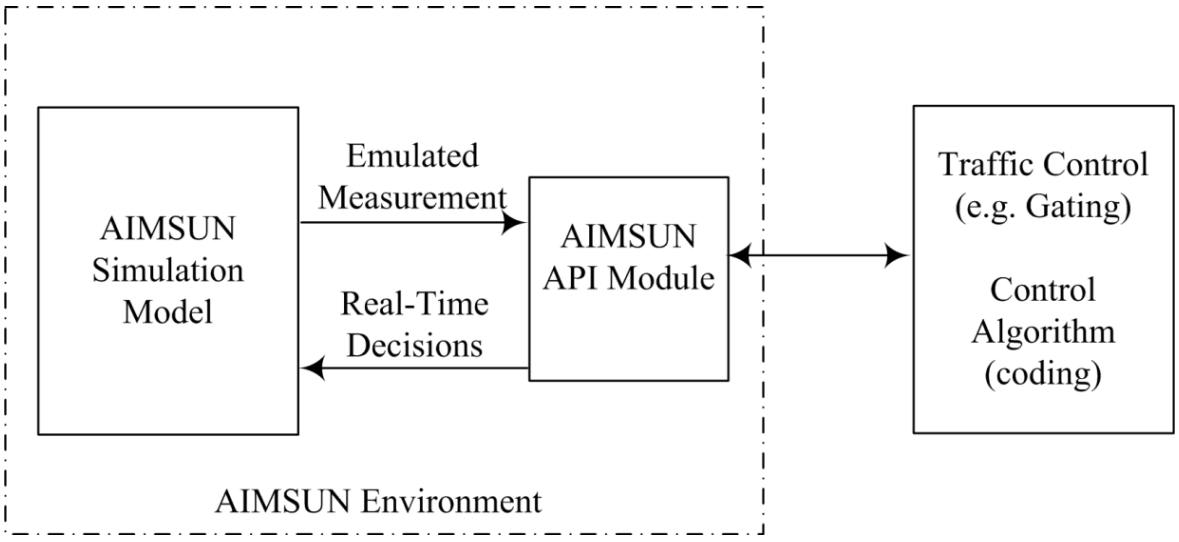

Figure 3-1 Schema of AIMSUN API module

3.1.1. Introduction to AIMSUN API

The AIMSUN API (Application Programming Interface) module extends the functions of AIMSUN environment including user defined applications which can exchange information and/or modify its state dynamically, with the AIMSUN module.

The AIMSUN API module is placed, in the functional point of view, between the AIMSUN simulation model and the external application (e.g. Traffic control strategies) defined by the user (see Figure 3-1). Considering that, there are two types of communication processes: on one side there is a communication process between the AIMSUN and the AIMSUN API module and on the other side between the AIMSUN API module and the external application. The communication process between AIMSUN API module and the AIMSUN simulation model is provided by the AIMSUN environment, but the communication between the AIMSUN API module and the External Application has to be implemented by the user, depending on the requirements of the external application.

In this thesis, as demonstrated in Figure 3-1, the API is applied for implementation of the control algorithm and emulation of real-time control system where AIMSUN delivers emulated measurements to the API and receives real-time decisions accordingly.

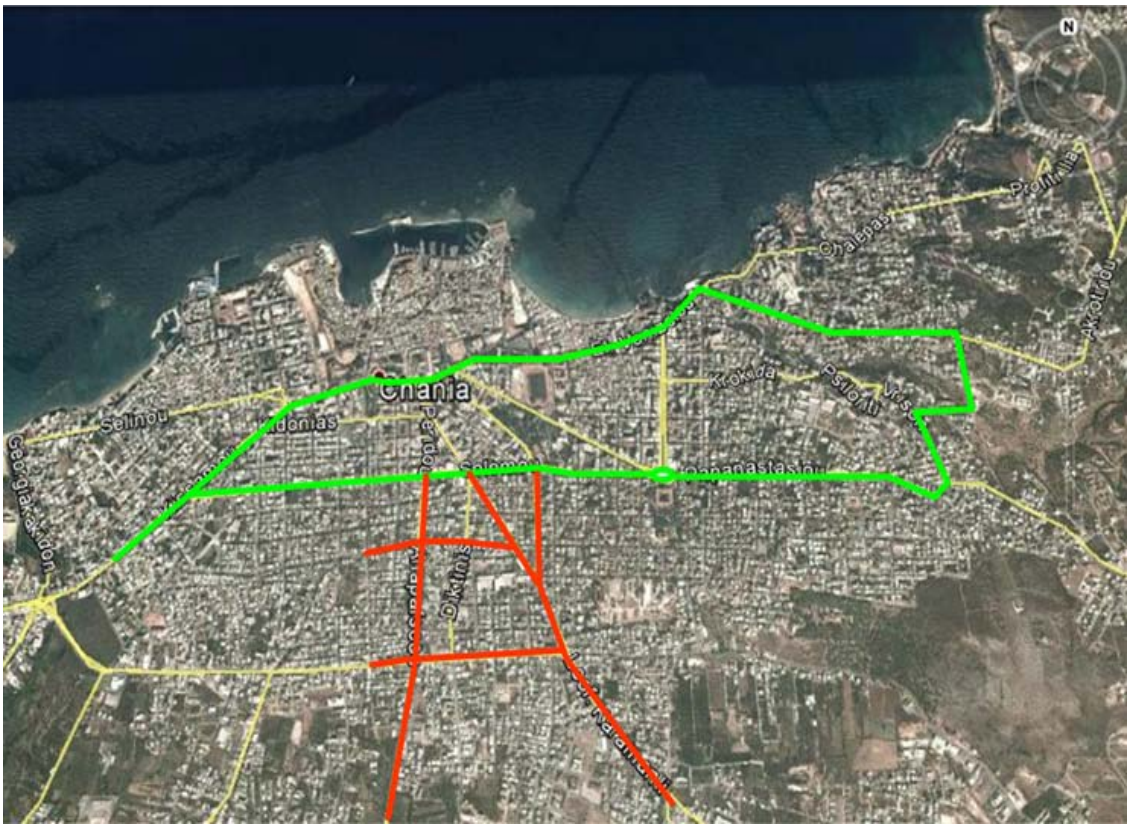

Figure 3-2 Satellite view of Chania urban network and the territory modeled in AIMSUN (shown by green and green + orange)

3.2. Implemented Urban Networks in this Thesis

The proposed gating control strategies in this PhD dissertation (introduced in Chapter 2) are studied on the following networks modeled in AIMSUN:

- A grater part of the Chania, Greece, urban road network (i.e. within the green line shown in Figure 3-2.
- An extended version of the above network which includes also the links shown by the orange line in Figure 3-2.
- The congested part of the downtown San Francisco, USA, indicated by the green line in Figure 3-3.

3.3. Urban Network of Chania, Greece

Chania, located at the north-western part of Crete, Greece, is the capital of the prefecture of Chania and covers 12.5km^2. Chania is the second biggest prefecture of Crete in size, population and development. It has a population of over 60.000 residents and is built over the

ruins of the ancient city of Kydonia. Figure 3-2 exhibits a satellite view of the trial Chania urban road network which consists of 27 and 29 signalized junctions for the green-colored and (green + orange)-colored networks, respectively. Many links in that network consist of only one lane, which means that unexpected events (accidents, double-parking, etc.) may block the link and therefore deteriorate the traffic conditions, even if their duration is only a few minutes.

Moreover, congestion problems are not limited in the streets with the unexpected events but they are propagated to many other streets and may sometimes lead to partial gridlock situations. Thus, the implemented control strategy should be able to deal with those problems. During the morning and evening hours there is a frequent bus service in almost every part of the network. Pedestrian movements are not a problem in the network and there is no reason for a special treatment. Public transport priority is not a subject in Chania, so it is not implemented in the experiments.

Congestion problems are encountered every day especially in the central and the northern part of the network for about one to two hours in the morning and evening. In most traffic arterials of the city there is heavy congestion 19:00–21:30 on Tuesday, Thursday and Friday evening, because of the shopping center. Another reason for the congestion is the high frequency of buses, which embark and disembark people at stops frequently blocking one direction of the street. Other reasons are the lower capacity due to illegal parking on the main streets and the high usage of vehicles by the residents of the city.

Heavy congestion problems are emerging on the entire network during the rainy days when there is an excessive inner and outer demand, usually a demand that cannot be sustained by the network’s infrastructure. Heavy congestion problems are emerging, also, during the tourist summer season. The heavily loaded urban network of the city is further loaded by additional private cars and motorbikes, rented cars and bikes, tourist buses, and by the increased movement of taxis. The aforementioned problems are encountered every day except Sunday and some rare occasions such as off-days. The city of Chania suffers from traffic congestion, lack of parking supply and traffic-generated pollution of the natural and built environment.

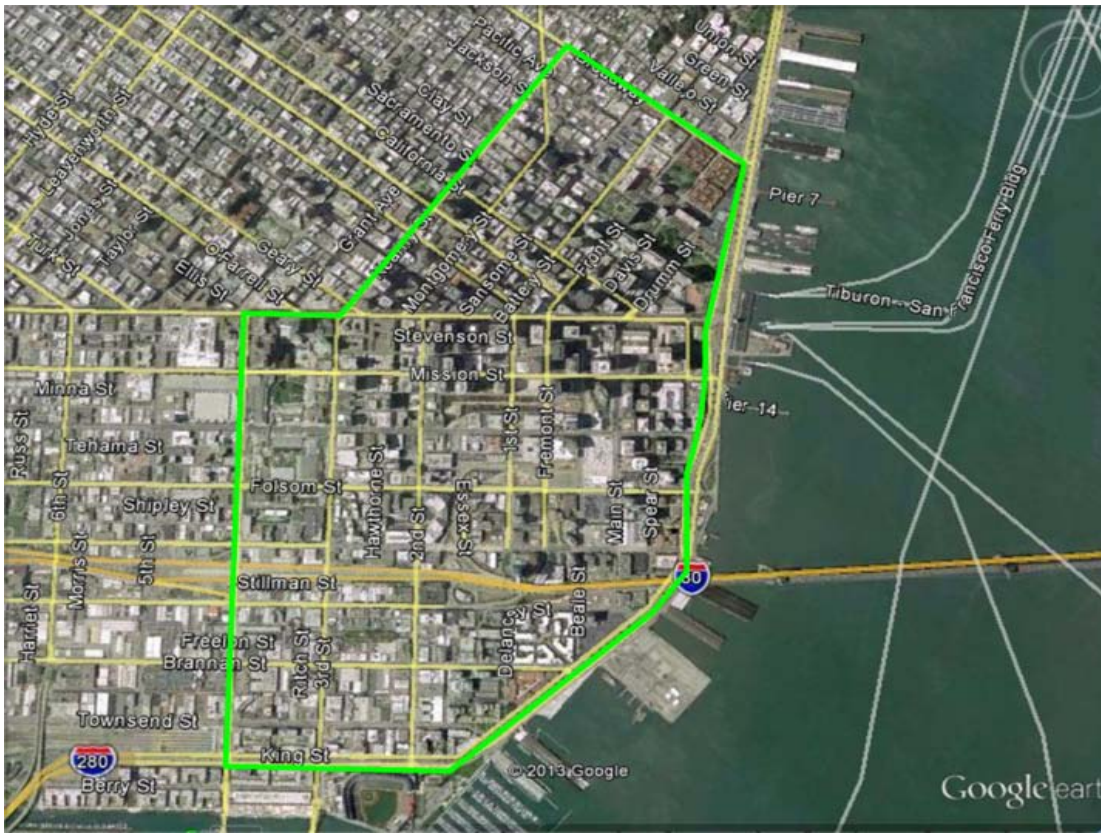

Figure 3-3 Satellite view of San Francisco urban network and the territory modeled in AIMSUN (shown by green)

3.4. Urban Network of San Francisco, USA

San Francisco, officially the city and county of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area. In Figure 3-3, a satellite view of the downtown of San Francisco and the under study district, marked by green line is shown.

The only consolidated city-county in California, San Francisco, encompasses a land area of about 46.9 square miles (121 km^2) on the northern end of the San Francisco Peninsula, giving it a density of about 17,620 people per square mile (6,803 people per km^2). It is the most densely settled large city in the state of California and the second-most densely populated major city in the United States after New York City. San Francisco is the fourth most populous city in California, after Los Angeles, San Diego and San Jose, and the 14th most populous city in the United States with an estimated population of 825,863 in 2012. The city is also the financial and cultural hub of the larger San Jose-San Francisco-Oakland metropolitan area, with a population of 8.4 million.

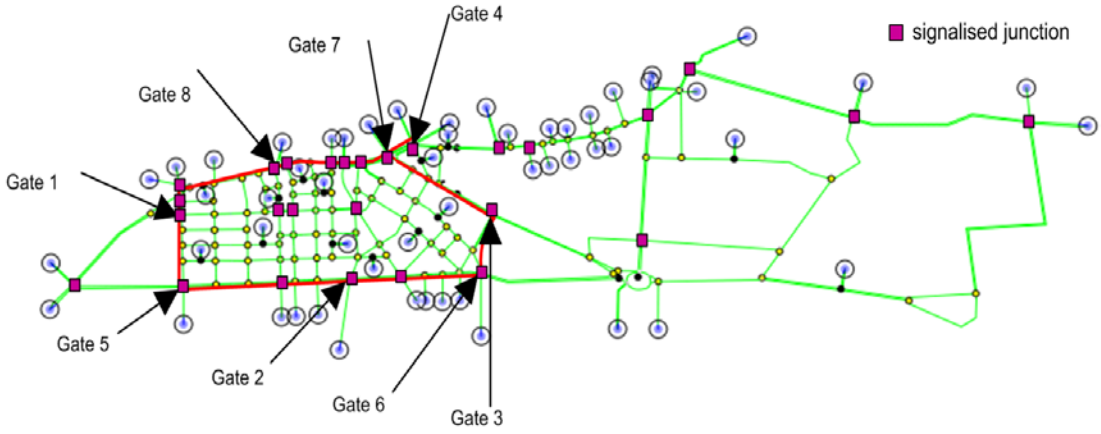

Figure 3-4 Chania urban network modelled in AIMSUN with gating position at the border of PN

The city suffers from congestion during the peak period, despite of implementation of various control strategies, due to the high number of trip attractions and productions in the downtown. Thus, employing this network (modeled in a simulator) sounds appealing for investigating novel traffic control strategies like gating.

3.5. Applied AIMSUN Scenarios

Four different simulation scenarios are used for investigating the different gating strategies introduced in Chapter 2:

- *Scenario 1*: The region indicated by green in Figure 3-2, is modeled in AIMSUN (see Figure 3-4). The central business district (CBD) of the Chania urban road network, where the congestion usually starts during the peak period, is considered as the protected network. Eight gating links are specified exactly at the border of the protected network, according to Figure 3-4. The PN is separated from the rest of the network by the red border in Figure 3-4. The total network includes 27 signalized junctions (indicated in Figure 3-4 by dark pink squares); while the protected network (PN) includes 19 signalized junctions, consists of 165 links and is separated from the rest of the network by the red border in Figure 3-4. Eight gated junctions are indicated with corresponding big arrows. In the middle of every link inside the red border line, a loop detector has been installed, and the related measurements are collected at every cycle (in this case 90 s). The gating links have been chosen to provide sufficient space for vehicle queuing, so that further upstream junctions are not significantly

obstructed. As indicated with small circled links in Figure 3-4, multiple origins and destinations are introduced at the network boundaries, but also at internal network locations, including the PN area. These origins and destinations (O-D) account for various corresponding in- and outflows, including on-street and off-street parking arrivals and departures, that may partially affect the PN area. The introduced O-D flows are realistic (based on real measurements) but not exact (particularly with regard to the used O-D rates).

When running the simulation with an O/D Matrix, a route choice system distributes the vehicles over the different paths from each origin to each destination. The distribution per path can depend on the cost of each link that composes the path. The cost of each link depends on the cost function associated to it. The default cost function in AIMSUN is travel time (selected also in this dissertation), but other cost functions for specific situations like tolls or, for instance, cost functions that take into account other factors like the psychological weight of low speeds can be selected. A time step of 30s is chosen as the route choice period in this study. This means that the route choice distribution is updated every 30s while AIMSUN is running. In this work, the tool's embedded real-time dynamic traffic assignment (routing) option is activated, as this is deemed to lead to a more realistic distribution of the demand within the network. AIMSUN is embedded with several different route choice models. Logit type models are among the most popular route choice models. Based on discrete choice theory, Logit models assign a probability to each alternative path between each origin-destination. In this study, C-Logit model [5] is chosen which is able to take the network topology into account and allows for alternative routes with little overlapping. In particular, if gating measures create long queues and delays at the gated links, alternative routes (if available) may be selected by the drivers towards their respective destinations; clearly, this reflects the medium-term routing behavior of drivers to any introduced gating measures. Note also that this diversion may jeopardize to some extent the intended gating impact if drivers divert and enter the PN via non-gated links; therefore, the choice of gating links should also consider the availability and potential attractiveness of alternative routes that bypass the gating location. For *study 1,* this simulation scenario is implemented.

- *Scenario 2*: This scenario is a modified version of *Scenario 1*. More specifically, in this scenario two types of streets have been specified for the PN in AIMSUN. The main (typically signal-controlled) streets feature a maximum speed of 50 km/h while the secondary (typically non-controlled) streets have a maximum speed of 25 km/h within AIMSUN. This accounts, among others, for a more realistic dynamic routing of vehicles. Note that this street distinction is not present in *scenario 1*. In addition, the O-D demand in the present scenario was reduced, compared to the *scenario 1*, to account for the aggravated congestion resulting due to the introduced street classes; in fact, the resulting highest level of congestion (without gating) is lower in this scenario compared to the previous one. It should be noted that the utilized 4-hour O-D demand scenarios in scenarios create simulated traffic conditions that are roughly similar to the real traffic conditions, but without any claim or attempt for quantitatively accurate reproduction of the real conditions. This scenario is used as the test-bed for *study 2*. It should be noted that in both *scenario 1* and *scenario 2*, one PN is defined and is prevented from oversaturation by a single-boundary gating strategy.

- *Scenario 3*: A part of the downtown San Francisco urban road network (shown previously in Figure 3-3 by green line) is modeled in the AIMSUN microscopic simulation environment, according to Figure 3-5. In a large-scale urban network like Figure 3-5, due the significant existing heterogeneous traffic condition, considering a single-boundary gating strategy (as for *scenario 1* and *scenario 2*) may not lead to an optimal gating action. Thus, partitioning the network and defining multiple PNs might improve the gating performance over a single-boundary gating strategy. To investigate the gating strategy introduced in *study 3* (see section 2.7), this simulation scenario is applied. In this scenario, two concentric PNs (PN_1 shown by red bold line in Figure 3-6 and PN_2 shown by dark pink line in Figure 3-5) are specified as the regions which should be protected by gating control. The region where the congestion usually starts during the peak period is considered as the first protected network. PN_1 consists of 119 links and 34 signalized junctions while 9 gated junctions are indicated by big black arrows. The dark pink line, together with left and right edges of the red bold rectangle (at the border of the entire network), in Figure 3-5 indicate the bigger PN (PN_2). PN_2 includes 317 links and 15 outer gating positions (all 11 intersections on the pink border and 4th, 5th, 6th and 7th intersection on the red border) out of the

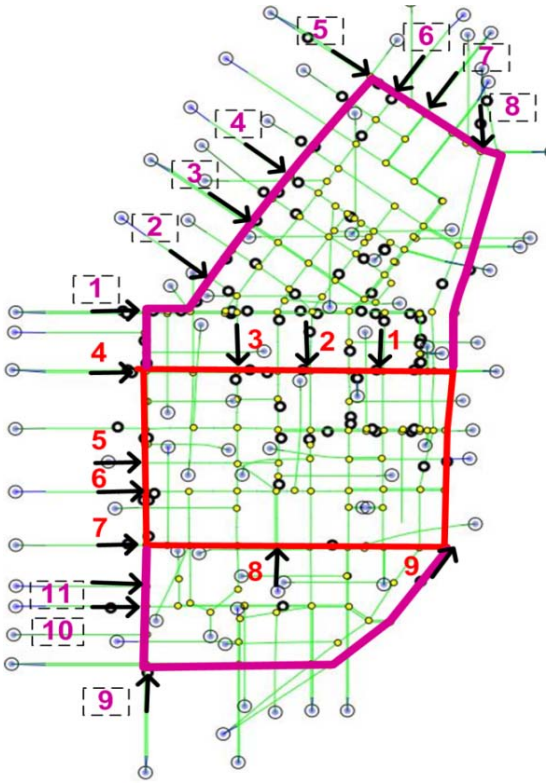

Figure 3-5 San Francisco urban network modeled in AIMSUN

82 controlled junctions. The approach presented in this part of the study includes two feedback controllers; the first one is associated with PN_1 and regulates the inflows through 9 intersections on the red border; the second one is related to PN2 and controls the entering flows of 11 intersections on the pink border ($PN_2 \backslash PN_1$). An emulated mid-block loop detector is installed in every link inside the PN_2, and the relevant measurements are collected every 60 s, which is the shortest cycle of the traffic lights inside the whole network. The gating positions are chosen so as to provide enough queuing space and have minimum impact on the traffic flow of further upstream junctions.

For a realistic distribution of the demand in the network, the tool's embedded real-time dynamic traffic assignment (DTA) option is activated during the AIMSUN simulation runs. A route choice period of 300s is considered for the simulations in this work. The utilized 5-hour O-D demand scenario creates simulated traffic conditions that are comparable to the real traffic conditions, without claiming any quantitatively accurate reproduction of the real conditions.

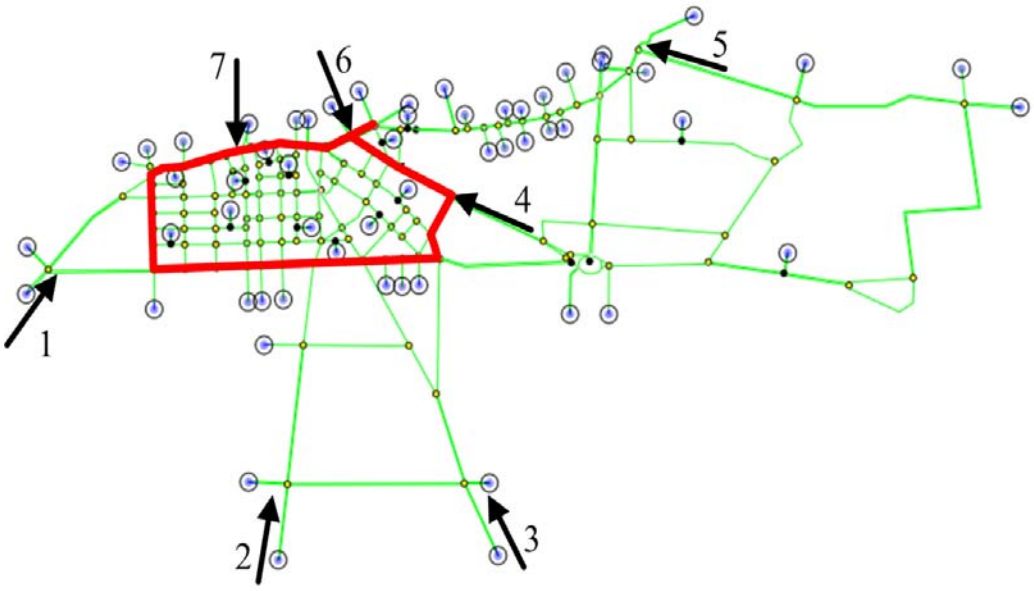

Figure 3-6 Extended Chania urban network modelled in AIMSUN with gating position further upstream of PN

- *Scenario 4*: In this scenario, the simulated Chania urban network is extended at the southern part of the network (see the orange lines in Figure 3-2), as shown in Figure 3-6. The region where the major peak-period congestion is typically encountered is again considered as the PN (specified by red bold line in Figure 3-6). Seven gated links are indicated by arrows in Figure 3-6, three of them located at the border of the protected network (i.e. gated junctions 4, 6 and 7); and the four other located further upstream of the PN (i.e. junctions 1, 2, 3 and 5). It should be noted that this scenario includes all simulation features defined in *scenario 2*. The introduced O-D flows are realistic (based on real measurements) but not identical with O-D flows in *scenario 1* and *scenario 2*. This scenario is used as a test-bed for *study 4* and *study 5*.

Chapter 4

It is the theory that decides what can be observed.

Albert Einstein

4. Modeling/Controller Design

Gating may be enabled via very simple, but highly efficient and robust feedback regulators that are well-known in Control Engineering. The regulators are strictly based on real-time measurements, without any need for online model or demand predictions. On the other hand, for a proper choice of the feedback structure (among several offered in classical feedback theory), it is essential to know the basic dynamics of the process under control, and this task is indeed rendered quite simple and easy when using the notion of the NFD.

In this chapter, a brief introduction to automatic system control, Single-Input-Single-Output (SISO) Systems, modeling and a procedure to design a proper PI feedback regulator for gating purpose are presented. Moreover, the gating action at the gated junctions is explained in details.

4.1. Introduction

Control systems are an integral part of everyday life in today's society. They control our appliances, our entertainment centers, our cars, and our office environments; they control our industrial processes and our transportation systems; they control our exploration of land, sea, air, and space.

In most modern engineering systems, there is a need to control the evolution with time of one or more of the system variables. Controllers are required to ensure satisfactory transient and

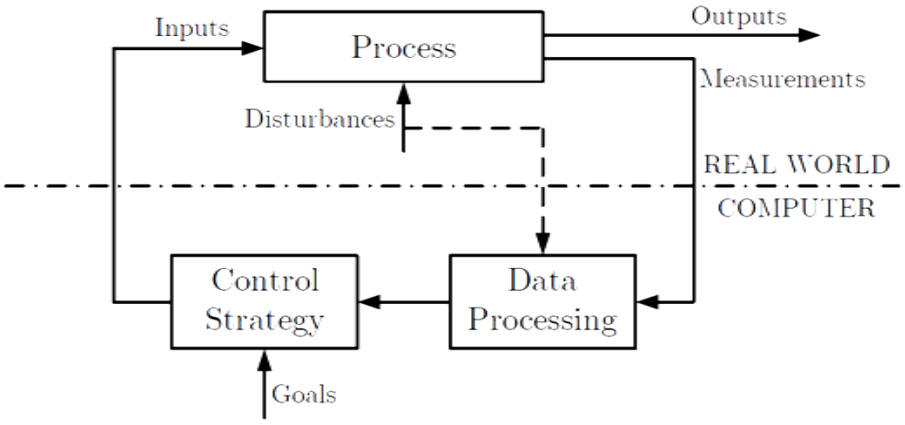

Figure 4-1 Basic elements of an automatic control system

steady-state behavior for these engineering systems. To guarantee satisfactory performance in the presence of disturbances and model uncertainty, most controllers in use today employ some form of negative feedback. A sensor is needed to measure the controlled variable and compare its behavior to a reference signal. Control action is based on an error signal defined as the difference between the reference and the actual values.

The controller that manipulates the error signal to determine the desired control action has classically been an analog system, which includes electrical, fluid, pneumatic, or mechanical components. These systems all have analog inputs and outputs (i.e. their input and output signals are defined over a continuous time interval and have values that are defined over a continuous range of amplitudes). In the past few decades, analog controllers have often been replaced by digital controllers whose inputs and outputs are defined at discrete time instances. In most engineering applications, it is necessary to control a physical system or plant (e.g. traffic system) so that it behaves according to given design specifications. Typically, the plant is analog, the control is piecewise constant, and the control action is updated periodically (e.g. every traffic cycle etc.). This arrangement results in an overall system that is conveniently described by a discrete-time model. A discrete-time system processes a given input sequence $x[k]$ to generate an output sequence $y[k]$ (where k is integer) with more desirable properties.

Automatic Control comprises those theoretical methods and practical procedures that enable the development of technical systems capable of accomplishing autonomously certain pre-specified tasks. Figure 4-1 illustrates the basic elements of an automatic control system. The

process (e.g. traffic flow in an urban network) includes all technical or physical phenomena that should be influenced according to specific goals.

Input values (e.g. the traffic lights, the variable message signs, etc.) may be selected from an admissible control region. The disturbance values cannot be manipulated but may possibly be directly measurable via appropriate devices (e.g. traffic demand), or may be estimated or predicted via appropriate algorithms (e.g. traffic demand, origin-destination pattern, etc.). The process outputs are the quantities chosen to represent the behavioral aspects of interest (e.g. the outputs of urban traffic may be the total travel time, the queue lengths, etc.). The data processing block in Figure 4-1 comprises the estimation and/or prediction tasks, based on real-time measurements of internal process quantities or disturbances.

The task of the control strategy is to specify in real time the process inputs, based on available measurements/estimations/predictions, so as to achieve the pre-specified goals regarding the process outputs, despite the influence of various disturbances.

The control system, shown in Figure 4-1, is characterized by a closed-loop structure; whereby the calculation of inputs is effectuated on the basis of measurements of process-internal quantities, which, by their turn, are influenced by the inputs.

4.2. Regulators

Automatic Control theory offers a number of methods and theoretical results for designing a regulator in a systematic and efficient way. A necessary condition for application of the Automatic Control theory to a particular process control problem is the availability of a mathematical model capable of describing the basic process behavior [17]. In fact, the model to be used for regulator design (the design model), may be quite simple if it includes the major aspects of the process behavior and if the designed regulator is sufficiently robust. Most regulators resulting from application of Automatic Control methods are very simple, as they consist of one single equation, but their efficiency and reliability are usually much higher than those of human regulators. It is important to note, that when designing a regulator, the mathematical process model is only used off-line, i.e. the online application does not include any model equations.

In control engineering, a SISO system is a simple control system with one input and one output. SISO systems are typically less complex than Multiple-Input-Multiple-Output

systems (MIMO). MIMO systems have too many interactions for most of us to trace through them quickly, thoroughly and effectively in our heads. The design methods for linear MIMO regulators are more difficult and advanced. Such methods are the Linear-Quadratic (LQ) optimization, pole assignment methods, decentralized control, hierarchical control, etc. Particular attention should be paid to the robustness properties of the designed regulators, via recently developed powerful methods and tools. Further methods for particular classes of regulators are available within Automatic Control theory, like, for example, nonlinear regulators (for nonlinear processes) and adaptive regulators, whereby the regulator parameters are adjusted automatically in real-time by suitable mechanisms, in order to account for process uncertainties or for time-varying process behavior.

Frequency domain techniques for analysis and controller design dominate SISO control system theory. Bode plot, Nyquist stability criterion, Nichols plot, and root locus are the usual tools for SISO system analysis [17]. To name just two more popular, controllers can be designed through the polynomial design, root locus design methods. Often SISO controllers will be Proportional-Integral (PI), Proportional-Integral-Derivative (PID) or lead-lag. In a SISO control system (e.g. PI feedback control), the control goal is to maintain the process value close to a pre-specified value (i.e. reference or set value).

4.3. Feedback Control

The development of feedback control methodologies has had a tremendous impact in many different fields of the engineering. Besides, nowadays the availability of control system components at a lower cost has favored the increase of the applications of the feedback principle (for example in consumer electronics products). As addressed earlier, feedback control is actually essential to keep the process variable close to the desired value in spite of disturbances and variations of the process dynamics.

The typical feedback control system is presented in Figure 4-2. Actuator is the device that can influence the controlled variable of the process [20]. In this thesis, traffic lights and *TTS* in PN are the actuator and the control variable, respectively. The control variable (*TTS*) is indirectly influenced by the traffic lights. In other words, traffic lights influence the inflow to PN and consequently the *TTS* (estimated number of vehicles in PN) is also affected by the traffic lights. The sensor (i.e. detectors in traffic control) measures the control variable, where

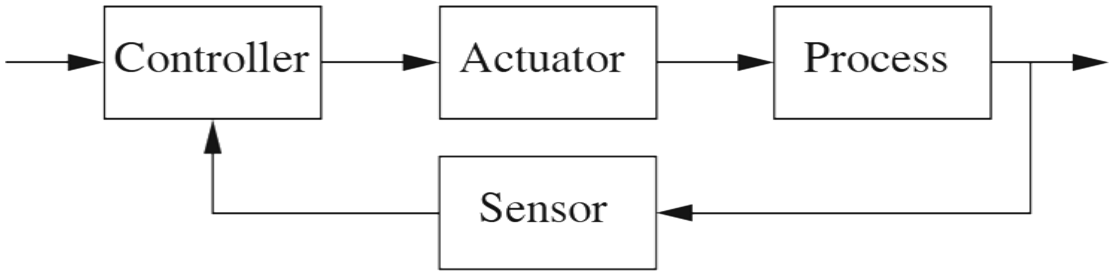

Figure 4-2 Typical components of a feedback control loop

in the case of urban traffic control, since *TTS* is not measureable (estimated from eq. 2-1) the occupancy is measured by the detectors.

Obviously, the overall control system performance depends on the proper choice of each component. From the purposes of controller design, the actuator and sensor dynamics are often neglected (although the saturation limits of the actuator have to be taken into account). The block diagram of a feedback control loop is displayed in Figure 4-3, where P is the process, C is the controller, r is the reference signal (or set-value), $e = r - y_n$ is the control error, y_n is $y+n$, u is the manipulated (control) variable, y is the process (controlled) variable, d is a load disturbance signal and n is a measurement noise signal.

4.4. On–Off or Bang-Bang Controller

One of the most adopted (and one of the simplest) controllers is undoubtedly the On–Off controller, where the control variable can assume just two values, u_{max} and u_{min}, depending on the control error sign. Formally, the control law is defined as follows:

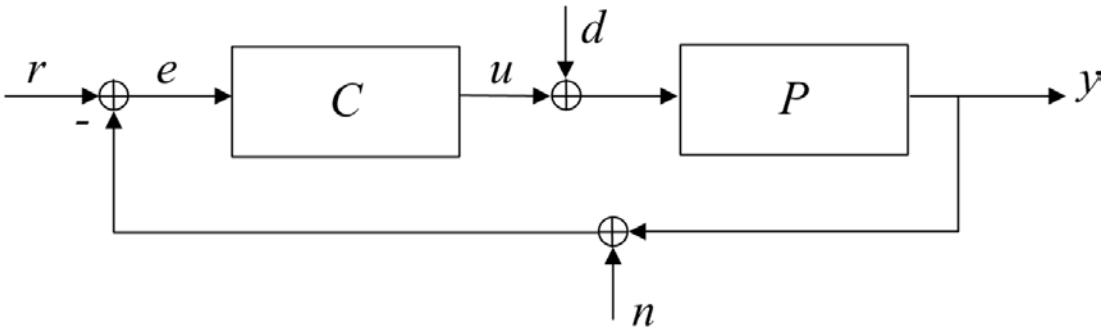

Figure 4-3 Schematic block diagram of a feedback control loop

$$u = \begin{cases} u_{\max} & \text{if} \quad e \geq 0 \\ u_{\min} & \text{if} \quad e < 0 \end{cases} \tag{4-1}$$

i.e., the control variable is set to its maximum value when the control error is positive and to its minimum value when the control error is negative. Generally, $u_{\min} = 0$ (off) is selected and the controller is usually implemented by means of a relay. The main disadvantage of the On–Off controller is that a persistent oscillation of the process variable (around the set-point value) occurs.

Actually, in practical cases, the On–Off controller characteristic is modified by inserting a dead zone (this results in a three-state controller) or hysteresis in order to cope with measurement noise and to limit the wear and tear of the actuating device. The typical controller functions are shown in Figure 4-4. Because of its remarkable simplicity (there are no parameters to adjust), the On–Off controller is indeed suitable for adoption when no tight performance is required, since it is very cost-effective in these cases.

4.5. System Modeling for Feedback Gating Control Design

The developed model and feedback controller are summarized in Figure 4-5. The model input is the gated flow q_g (see Figure 2-2); the model output is the PN's *TTS;* while the main external disturbance is the uncontrolled PN inflow q_d. The model is first developed in a continuous-time environment for convenience. To start with, we have in the general case

$$q_{in}(t) = \beta \cdot q_g \cdot (t - \tau) \tag{4-2}$$

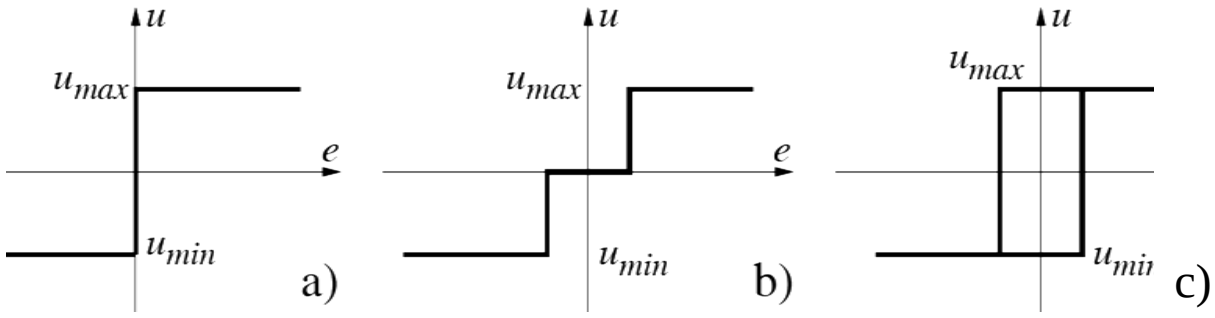

Figure 4-4 Typical On–Off controller characteristics. a) ideal; b) modified with a dead zone; c) modified with hysteresis

where β is the portion of gated flow (q_g) that enters the PN; t is the time argument; τ is the travel time needed for gated vehicles to approach the PN (when the gating link is not directly at the PN boundary). The conservation equation for vehicles in the PN (see Figure 2-2) reads:

$$\dot{N}(t) = q_{in}(t) + q_d(t) - q_{out}(t). \tag{4-3}$$

As in the discrete-time case eq. (2-1), we have also for the ideal values $TTS_{id}(t) = N(t)$ (where N is the real number of vehicles within PN), but *TTS* in Figure 4-5 denotes the operational value, which differs from the ideal value in two respects: firstly, detectors may not be available in each and every PN link, hence the operational *TTS* will be smaller by some factor $A \le 1$; secondly, the occupancy measurement and, most importantly, the estimation eq. (2-3) may not be exact, hence we introduce a measurement/estimation error ε_1; which finally yields

$$TTS(t) = A \cdot N(t) + \varepsilon_1(t) \tag{4-4}$$

From this operational $TTS(t)$, we may derive, using the operational NFD, the corresponding (operational) *TTD*, i.e.

$$TTD(t) = F\left[TTS(t)\right] + \varepsilon_2(t) \tag{4-5}$$

where $F(.)$ is a nonlinear best-fit function of the operational NFD's measurement points, and ε_2 denotes the corresponding fitting error (due to NFD scatter). Since *TTD* in eq. (4-5) and Figure 4-5 is the operational quantity, the ideal TTD_{id} (considering all PN links, not just the ones equipped with detectors) will be bigger, i.e.

$$TTD_{id}(t) \cdot B = TTD(t) \tag{4-6}$$

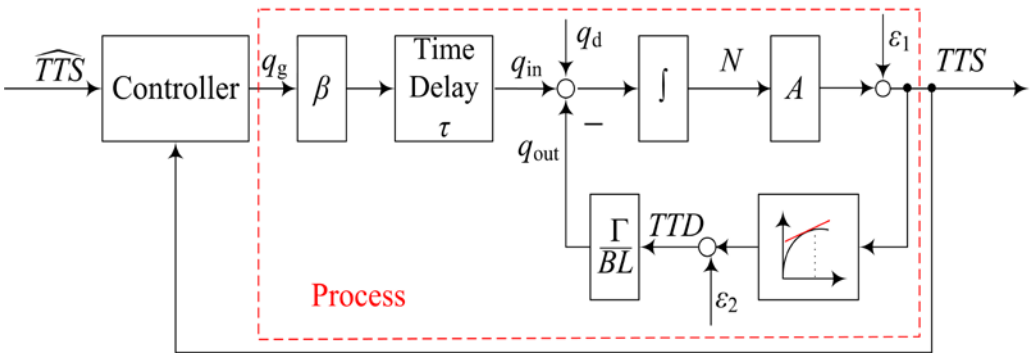

Figure 4-5 Block diagram of the system and the feedback controller

where $B \leq 1$ is the flow-analogous factor of A earlier.

To proceed, we will now introduce the modeling assumption that the PN outflow q_{out} is proportional to TTD_{id}, i.e.

$$q_{\text{out}}(t) = \frac{\Gamma}{L} TTD_{\text{id}}(t) \tag{4-7}$$

where Γ is a sort of network exit rate, $0 \leq \Gamma \leq 1$, and L is the average PN link length. Replacing eq. (4-6) in eq. (4-7), we complete the process model derivation according to Figure 4-5. The overall model (from q_{g} to *TTS*) is obtained by replacing eq. (4-2), eq. (4-4)-(4-7) into (4-3) and turns out to be a time-delayed nonlinear first-order system. Its portion from q_{in} to *TTS* (i.e. without the time delay) reads

$$\frac{d}{dt}(TTS(t)) = \left(q_{\text{in}} + q_{\text{d}} - \frac{\Gamma}{BL} F\left[TTS(t)\right] \right) \cdot A + \varepsilon(t)$$

Where ε may be derived from the previous errors ε_1 and ε_2.

This model may be linearized around an optimal steady state that is within the aforementioned maximum throughput region (see the yellow region in Figure 2-1) of the NFD. The introduction of a desired steady state is quite usual in Control Engineering to enable the derivation of a linearized model and subsequent linear feedback control design. In fact, it is the goal of the feedback regulator, to be eventually derived, to maintain the system state around this steady state; thus, if successful, the steady-state and the linearized system dynamics assumptions are actually imposed via the intended control action. Denoting steady-state variables with bars, we have

$$\overline{q}_{\text{in}} + \overline{q}_{\text{d}} = \overline{q}_{\text{out}} \tag{4-8}$$

$$\overline{q}_{\text{out}} = \frac{\Gamma}{BL} \overline{TTD} \tag{4-9}$$

while $\overline{\varepsilon}_1$ and $\overline{\varepsilon}_2$ are set equal to zero. With the notation $\Delta x = x - \overline{x}$ used analogously for all variables, the linearization yields

$$\frac{d}{dt}(\Delta TTS) = \left(\Delta q_{\text{in}} + \Delta q_{\text{d}} - \frac{\Gamma \overline{F}'}{BL} \Delta TTS \right) \cdot A + \varepsilon \tag{4-10}$$

where $\bar{F}'$ is the slope of the NFD at the optimal set-point $\hat{TTS}$, i.e. $\overline{TTS} = \hat{TTS}$. This set-point should be selected within the optimal *TTS*-range (i.e. yellow region in Figure 2-1) of the NFD, for the Chania PN, for maximum efficiency. Note that $\bar{F}'$ may be virtually equal to zero if the set-point is optimal; nevertheless, we will assume $\bar{F}' > 0$ here, in order to enable proper linearized modeling. This assumption has no impact on the employed regulator (4-25), whose operation is only governed by the regulation error $\hat{TTS} - TTS$.

The continuous-time state eq. (4-10) of the protected network (using the conservation equation and the NFD) may be directly translated in discrete time by use of standard formulas [54] as follows (the ε is dropped and added again in the eq. (4-24))

assuming $C = \Gamma \bar{F}' / BL$, we have $\Delta \dot{TTS}(t) + C \cdot A \cdot \Delta TTS(t) = \left(\Delta q_{\text{in}}(t) + \Delta q_{\text{d}}\right) \cdot A$ (4-11)

Applying the Laplace transform and assuming zero initial conditions

$$\left(s + C \cdot A\right) \Delta TTS(s) = \left(\Delta q_{\text{in}}(s) + \Delta q_{\text{d}}(s)\right) \cdot A. \tag{4-12}$$

Because of the superposition property we have

$$\frac{\Delta TTS_{\text{in}}}{\Delta q_{\text{in}}} = G_{\text{in}}(s) = \frac{A}{s + C \cdot A}; \qquad \frac{\Delta TTS_{\text{d}}}{\Delta q_{\text{d}}} = G_{\text{d}}(s) = \frac{A}{s + C \cdot A} \tag{4-13}$$

$$\Delta TTS = \Delta TTS_{\text{in}} + \Delta TTS_{\text{d}} \tag{4-14}$$

The discrete transfer function of a continuous system proceeded by zero order holder (ZOH) is

$$G(z) = \left(1 - z^{-1}\right) \cdot Z\left\{\frac{G(s)}{s}\right\} \tag{4-15}$$

then we have

$$G_{\text{in}}(z) = G_{\text{d}}(z) = \frac{z-1}{z} Z\left\{\frac{A}{(s + C \cdot A)s}\right\} \quad \text{if} \quad a = C \cdot A \tag{4-16}$$

from the table of transforms

$$Z\left\{\frac{1}{s(s+a)}\right\} = \frac{z\left(1 - e^{-aT}\right)}{a(z-1)\left(z - e^{-aT}\right)}; \text{therefore} \tag{4-17}$$

$$G_{\text{in}}(z) = G_{\text{d}}(z) = \frac{z-1}{z} \cdot A \cdot \frac{z\left(1 - e^{-C \cdot A \cdot T}\right)}{C \cdot A(z-1)\left(z - e^{-C \cdot A \cdot T}\right)} = \frac{1 - e^{-C \cdot A \cdot T}}{C\left(z - e^{-C \cdot A \cdot T}\right)}; \tag{4-18}$$

$$\mu = e^{-C\cdot A\cdot T} \Rightarrow G_{\text{in}}(z) = G_{\text{d}}(z) = \frac{1-\mu}{C}\cdot\frac{1}{z-\mu} \tag{4-19}$$

$$\Delta TTS(z) = \frac{1-\mu}{C}\cdot\frac{1}{z-\mu}\Delta q_{\text{in}}(z) + \frac{1-\mu}{C}\cdot\frac{1}{z-\mu}\Delta q_{\text{d}}(z) \tag{4-20}$$

$$= \frac{1-\mu}{C}\cdot\frac{1}{z-\mu}\left[\Delta q_{\text{d}}(z) + \Delta q_{\text{in}}(z)\right] \tag{4-21}$$

$$z\cdot\Delta TTS(z) - \mu\cdot\Delta TTS(z) = \frac{1-\mu}{C}\left[\Delta q_{\text{d}}(z) + \Delta q_{\text{in}}(z)\right]. \tag{4-22}$$

By applying the inverse z-transfer $Z\left\{f(t-nT)\right\} \rightleftarrows z^{-n}F(z)$

$$\Delta TTS(k+1) - \mu\cdot\Delta TTS(k) = \frac{1-\mu}{C}\left[\Delta q_{\text{d}}(k) + \Delta q_{\text{in}}(k)\right] \tag{4-23}$$

By rearranging

$$\Delta TTS(k+1) = \mu\cdot\Delta TTS(k) + \zeta\cdot\left[\Delta q_{\text{in}}(k) + \Delta q_{\text{d}}(k)\right] + \varepsilon(k) \tag{4-24}$$

where $\zeta = (1-\mu)BL/\Gamma\bar{F}'$. It is trivial to include in these models the time delay, by replacing q_{in} from eq. (4-2).The derived simple model includes a number of parameters that have clear physical meaning; nevertheless, the precise value of some of these parameters may be difficult to obtain in practice, particularly if the PN is a sizeable network (as in the Chania and San Francisco example). However, the main reason for developing the gating model is to deduce the basic structure of the underlying dynamics, which is essential for a proper choice of the regulator structure.

4.6. PI Feedback Controller Design

Field implementations call for simple and efficient systems that would expedite their application. The gating control problem is to regulate the *TTS* (number of vehicles) in the PN, so as to maintain it around a selected set-point $\hat{TTS}$, via appropriate manipulation of the gated flow q_{g} in real time; thus, we have a Single-Input-Single-Output (SISO) control problem with q_{g} as the control input and *TTS* as the control output. To avoid congestion-caused degradation (i.e. a *TTD* decrease), the targeted set-point $\hat{TTS}$ should be selected within the range of critical values, where *TTD*, and hence the PN throughput, are maximized.

To this end, given the derived model structure in the previous section, the following standard proportional-integral-type (PI) feedback controller is well suitable

$$q_{\mathrm{g}}(k) = q_{\mathrm{g}}(k-1) - K_{\mathrm{P}}\left[TTS(k) - TTS(k-1)\right] + K_{\mathrm{I}}\left[T\hat{T}S - TTS(k)\right] \quad (4\text{-}25)$$

where K_{P} and K_{I} are the (non-negative) proportional and integral gains, respectively. Good regulator gain values may be found with appropriate Control Engineering methods or manual fine-tuning; model parameter estimation (e.g. of μ and ζ in eq. (4-24)), by use of real q_g versus *TTS* measurements, may be useful in this endeavor; in any case, feedback regulators are quite robust to moderate parameter value changes.

If gating is applied at multiple links, the flow calculated by the (unique) regulator (4-25) must be split among the gated links according to some pre-specified policy. As long as the feedback-ordered total inflow is roughly followed via the gating traffic signal actions, the performance of the control (in terms of delay reduction) is not expected to depend significantly on the inflow splitting policy, except perhaps for special cases of network topology or demand patterns. What the splitting significantly affects, is the resulting queuing and delays at the individual gated links. For example, one may envisage the application of delay-balancing or queue-balancing splitting policies as in [50]. This thesis is mainly concerned with the overall gating control design and its application on different scenarios (see section 3.5); while splitting and queue management issues, which may include a variety of policies or wishes by the responsible traffic authorities and may incur corresponding requirements on detector equipment and overall cost, are left for more detailed future investigations. Thus, the splitting of the total ordered inflow in this thesis is simply conducted in proportion to the respective saturation flows of the gated links which is discussed in section 4.7 in details.

The flow calculated by the regulator (4-25) must be constrained by pre-specified minimum and maximum values to account for physical or operational constraints. For the lower bound, $q_{\min}$, one may choose the flow corresponding to the minimum-green settings of the gated links (as in thesis), or higher, e.g. if some gated links need to be protected from over-spilling. The upper bound has two components, a constant and a variable one, similarly to ALINEA ramp metering [58], and it is decided in real time which of the two is to be applied at each control step; the constant upper bound may be specified according to the maximum-green settings of the gated links (as in this thesis), or lower, e.g. if some downstream links need to

be protected from over-spilling; the variable upper bound aims at activating the regulator more promptly under certain circumstances, see [58] for further details on the reasoning and method. It should be noted that, in this thesis, upper and lower flow bounds are actually specified also for every individual gated link. If the regulator flow distribution is found to violate some of these individual bounds, then the surplus flows are re-distributed among the rest of the gated links.

The necessary and sufficient condition on the (non-negative) regulator parameters K_{P} and K_{I} for closed-loop stability of the linearized system eq. (4-24) under regulator eq. (4-25) can be easily established by use of the Jury-Blanchard criterion [54] to be $K_{\mathrm{I}} > 0$ and $2K_{\mathrm{P}} + K_{\mathrm{I}} < 2(\mu+1)/\zeta$; this means that the system can be stabilized even if the P-term in eq. (4-25) is dropped, i.e. for $K_{\mathrm{P}} = 0$. On the other hand, the rigorous proof of stability for the nonlinear system is more cumbersome and is left for a more control-oriented investigation; roughly speaking, if *TTS* is higher than the set-point $\hat{TTS}$, then the last term on the right of the regulator eq. (4-25) will continuously reduce the ordered flow q_{g}, such that *TTS* approaches its set-point; however, this action reaches its limit when q_{g} reaches its lower bound $q_{\min}$ mentioned earlier; then, if the sum $q_{\min} + q_{\mathrm{d}}$ of controlled and uncontrolled inflows happens to be higher than the *TTS*-dependent outflow q_{out}, the network cannot recover from over-saturation as eq. (4-3) indicates. This circumstance sets according limits to the level of the inflow q_{d}, that is left uncontrolled, in conjunction with the lower admissible bound q_{g} for the controlled inflow.

It is interesting at this point to consider the optimal rule of [11] for saturated network control. Translated in the present notation and context, that rule suggests: At each k, set $q_{\mathrm{in}}(k) = 0$ if $TTS(k) > \hat{TTS}$; else set $q_{\mathrm{in}}(k)$ as high as possible, subject to the constraint $TTS(k+1) \le \hat{TTS}$. The first part of this rule may be readily implemented in practice, using of course a positive lower bound $q_{\min}$ for $q_{\mathrm{in}}(k)$ instead of zero for obvious reasons; but the literal implementation of the second part of the rule would call for an exact model and uncontrolled inflow q_{d} information to guarantee that $TTS(k+1)$ will not exceed $\hat{TTS}$, which is not practicable. One way to render the rule practicable, is to apply an upper bound to $q_{\mathrm{in}}(k)$ when $TTS(k) \le \hat{TTS}$, i.e., overall,

$$q_{\text{in}}(k) = \begin{cases} q_{\min} & \text{if } TTS(k) > \hat{TTS} \\ q_{\max} & \text{else} \end{cases} \quad (4\text{-}26)$$

This is a bang-bang regulator (like the one deployed in electric irons) which is equivalent to the regulator eq. (4-25) in terms of set-point and real-time data requirements. Such a bang-bang regulator would incur a stationary oscillation of $TTS(k)$ around $\hat{TTS}$, but, given the relatively wide range of throughput-maximizing *TTS* values (in practice and also in simulation, see simulation results NFDs, Figure 5-1, Figure 5-7, Figure 5-14 and Figure 5-20 in chapter 5), the oscillation may not really affect the resulting efficiency. However, the implied frequent switching of the gated link green phase between a minimum and a maximum value may not be desirable with the drivers and the road authorities. In contrast to eq. (4-26), the regulator eq. (4-25) offers a smooth control behavior and $TTS(k) = \hat{TTS}$ under stationary conditions, as can be easily deduced from eq. (4-25).

Gating could be activated only within specific time windows (e.g. at the peak periods) or if some real-time measurement-based conditions are satisfied. After distributing the regulator-ordered flow to the gated links, the individual sub-flows must be converted to appropriate green times by modifying the usual traffic signal settings in the corresponding junctions. In this thesis, this was done simply by modifying the duration of the signal stages where gated inflows are involved; while more elaborated procedures involving changes of the stage structure, e.g. so as to reduce delays for PN exiting flows, are considered which will be discussed in section 4.7.

It should be noted that, under any signal implementation policy and conditions, the total implemented PN inflow may be different than the flow ordered by the regulator for a number of reasons, including limited accuracy of signal specification, low demand, over-spilling downstream link or flow constraints; however, the regulator is largely robust to these potential occurrences thanks to its feedback structure, as it will be demonstrated in the next section.

4.6.1. Regulator's Gain Values (Without Time Delay)

The gating control strategy, implemented in *study 1*, *study 2* and *study 3*, is applied at the border of PN. Thus, the ordered flow by the regulator enters the PN promptly and there is not a significant delay between the action and the influence of the controller. This means that the

time-delay term τ in eq. (4-2) (see Figure 4-5 also), is considered zero, hence the K_{P} and K_{I} values may be specified for the controlled system eq. (4-24), eq. (4-25) to exhibit a time-optimal dead-beat regulator behavior, i.e. so as to reach the set-point within one single step, see also [48] for a similar feedback control design by use of the system's z-transforms. By rearranging eq. (4-25) and adding $T\hat{T}S$ and subtracting

$$q_{\mathrm{in}}(k) = q_{\mathrm{in}}(k-1) - K_{\mathrm{P}}\left[T\hat{T}S - TTS(k-1)\right] + K_{\mathrm{P}}\left[T\hat{T}S - TTS(k)\right] + K_{\mathrm{I}}\left[T\hat{T}S - TTS(k)\right] \quad (4\text{-}27)$$

Shifting the time

$$q_{\mathrm{in}}(k+1) - q_{\mathrm{in}}(k) = -K_{\mathrm{P}} \cdot \Delta TTS(k) + K_{\mathrm{P}} \cdot \Delta TTS(k+1) + K_{\mathrm{I}} \cdot \Delta TTS(k+1) \quad (4\text{-}28)$$

by applying the z-transform

$$zq_{\mathrm{in}} - q_{\mathrm{in}} = -K_{\mathrm{P}} \cdot \Delta TTS + (K_{\mathrm{P}} + K_{\mathrm{I}}) \cdot z \cdot \Delta TTS \Rightarrow \underbrace{\frac{\Delta q_{\mathrm{in}}}{\Delta TTS}}_{C \ \text{controller}} = \frac{(K_{\mathrm{P}} + K_{\mathrm{I}}) \cdot z - K_{\mathrm{P}}}{z-1} \quad (4\text{-}29)$$

By closing the loop with C (controller) from (4-29) and G (process transfer function) from (4-13)

$$H = \frac{GC}{1+GC} \quad (4\text{-}30)$$

$$= \frac{\left[(K_{\mathrm{P}} + K_{\mathrm{I}}) \cdot z - K_{\mathrm{P}}\right](1-\mu)}{(z-1)(z-\mu)C + \left[(K_{\mathrm{P}} + K_{\mathrm{I}}) \cdot z - K_{\mathrm{P}}\right](1-\mu)} \quad (4\text{-}31)$$

To have dead-beat $H = \dfrac{GC}{1+GC} = \dfrac{1}{z}$ (4-32)

In discrete-time control theory, dead-beat control problem consists of finding what input signal must be applied to system in order to bring the output to the steady state in the smallest number of time steps. For an N-th order linear system, it can be shown that this minimum number of steps will be at most N (depending on the initial condition), provided that the system is null controllable (that it can be brought to state zero by some input).

The solution is to apply feedback so that all poles of the closed-loop transfer function are at the origin of the z-plane. A closed-loop transfer function which has all poles of the transfer

function at the origin is sometime called a dead-beat transfer function. For non-linear systems dead-beat control is an open research problem.

The dead-beat response has the following characteristics:

1. Zero steady state error
2. Minimum rise time
3. Minimum setting time
4. Less than 2% overshoot/undershoot

According to eq. (4-32), a dead-beat behavior is established for set-point step-changes with $K_P = \mu/\zeta$ and $K_I = (1-\mu)/\zeta$; and for disturbance q_d step-changes with the same K_P but $K_I = 1/\zeta$. Thus, a least-squares parameter estimation is first conducted (discussed in the following sub-section in details) for μ and ζ in eq. (4-24), using time-series of (q_{in}, *TTS*)-measurements within and around the critical *TTS*-range. Once the parameters μ and ζ have been specified, the regulator parameters for (set-point) dead-beat behavior can be calculated.

4.6.1.1. Model Identification

To derive the proper model for the control purpose, model identification procedure is carried out. In control engineering, the field of system identification uses statistical methods to build mathematical models of dynamical systems from measured data. System identification also includes the optimal design of experiments for efficiently generating informative data for fitting such models as well as model reduction.

In the general case, from the eq. (4-24) we have

$$\Delta TTS(k+1) = \mu \cdot \Delta TTS(k) + \zeta \cdot \Delta q_{in}(k) + \gamma \cdot \Delta d(k) \tag{4-33}$$

Where $d = q_d$ is the disturbance inflow. Values for $TTS(k)$ and $q_{in}(k)$, k=1,…, K, may be assumed to be available, since they are needed for the operation of the feedback regulators. Since the desired set-point $\hat{TTS}$ is also given, we may obtain corresponding values for $\Delta\hat{TTS}(k) = TTS(k) - \hat{TTS}$, k=1,…, K, on the other hand, the exact corresponding steady-state values $\bar{q}_{in}$ and $\bar{d}$ are not known. A reasonable assumption is to calculate these values as average, i.e.

$$\bar{q}_{\text{in}} = \frac{1}{K}\sum_{k=1}^{K} q_{\text{in}}(k) \tag{4-34}$$

$$\bar{d} = \frac{1}{K}\sum_{k=1}^{K} d(k) \tag{4-35}$$

Using $\bar{q}_{\text{in}}$ from eq. (4-34) and from the available $q_{\text{in}}(k)$, we may derive values for $\Delta q_{\text{in}}(k)$, k=1,…, K. However, there are no measurements available for the uncontrolled inflows $d(k)$. Hence, as an approximation, we may use a constant term in place of $\gamma \cdot \Delta d(k)$ in eq. (4-33) (see eq. 4-36); this constant term could be taken equal to the average of $\gamma \cdot \Delta d(k)$; which, in view of eq. (4-35), turns out to be zero (see eq. (4-37) and (4-38) for details).

$$\gamma \cdot \Delta d(k) \to \frac{\gamma}{K}\sum_{k=1}^{K} \Delta d(k) = \Omega \tag{4-36}$$

$$\Delta d(k) = d(k) - \bar{d} = d(k) - \frac{1}{K}\sum_{k=1}^{K} d(k) \Rightarrow \tag{4-37}$$

$$\frac{\gamma}{K}\sum \Delta d(k) = \frac{\gamma}{K}\sum d(k) - \frac{\gamma}{K}\sum d(k) = 0 \tag{4-38}$$

In conclusion, the least-squares parameter estimation procedure is carried out to specify appropriate values for μ and ζ for the system

$$\Delta TTS(k+1) = \mu \cdot \Delta TTS(k) + \zeta \cdot \Delta q_{\text{in}}(k) \tag{4-39}$$

based on available data for $TTS(k)$ and $q_{\text{in}}(k)$ as explained above.

Field data (or simulation-based data as in this thesis) is often accompanied by noise. Even though all control parameters (independent variables) remain constant, the resultant outcomes (dependent variables) vary. A process of quantitatively estimating the trend of the outcomes, also known as regression or curve fitting, therefore becomes necessary.

The curve fitting process fits equations of approximating curves to the raw field data. Nevertheless, for a given set of data, the fitting curves of a given type are generally *NOT unique*. Thus, a curve with a minimal deviation from all data points is desired. This *best-fitting curve* can be obtained by the method of least squares.

Generally, the method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (least square error) from a given set

of data. Suppose that the data points are (x_1, y_1, z_1), (x_2, y_2, z_2), …, (x_n, y_n, z_n) where *x* and *y* are the independent variables and *z* is the dependent variable. The fitting curve $f(x, y)$ has the deviation (error) *e* from each data point, i.e. $e_1 = z_1 - f(x_1, y_1)$, $e_2 = z_2 - f(x_2, y_2)$, …, $e_n = z_n - f(x_n, y_n)$. According to the method of least squares, the best fitting curve has the property that:

$$\Pi = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[z_i - f(x_i, y_i) \right]^2 = \sum_{i=1}^{n} \left[z_i - (R \cdot x_i + S \cdot y_i) \right]^2 \rightarrow \underset{R,S}{\text{Min}} \qquad (4\text{-}40)$$

To derive *R* and *S* (i.e. the model parameters) we have

$$\frac{\partial \Pi}{\partial R} = 2 \sum_{i=1}^{n} x_i \left[z_i - (R \cdot x_i + S \cdot y_i) \right] = 0 \qquad (4\text{-}41)$$

$$\frac{\partial \Pi}{\partial S} = 2 \sum_{i-1}^{n} y_i \left[z_i - (R \cdot x_i + S \cdot y_i) \right] = 0 \qquad (4\text{-}42)$$

In this thesis, to estimate μ and ζ for (4-39), we have $z = \Delta TTS(k+1)$, $x = \Delta TTS(k)$, $y = \Delta q_{\text{in}}(k)$, $R = \mu$ and $S = \zeta$.

In the case of presence of time-delay (τ) in the system, $z = \Delta TTS(k+1)$, $x = \Delta TTS(k)$ and $y = \Delta q_{\text{in}}(k-m)$ where $m \triangleq$ time-delay, and the parameter estimation problem (for μ and ζ) should be solved for $m = 1, 2, \ldots$; the best value for Π leads to the desired parameters. Consequently, μ, ζ and *m* are derived and implemented for the control design problem which is presented in 4.6.3.

4.6.2. Multiple Boundaries Feedback-Based Gating

The feedback gating concept introduced in section 4.6 is applied for a multiple concentric-boundary gating strategy. In large-scale urban networks, due to the heterogeneous spatial distribution of congestion, different regions may not reach the critical accumulation of the NFD simultaneously. Consequently, applying homogeneous gating strategy at the boundary of a large urban network could result in an immature gating action. Thus, in a joint research work with Prof. Nikolas Geroliminis and Mehmet Yildirimoglu (PhD student) at Urban Transport Systems Laboratory (LUTS), Ecole Polytechnique Federale de Lausanne (EPFL),

Switzerland, we propose a two-stage gating strategy which consists of two PI feedback controllers. The proposed gating strategy is tested on *scenario 3* and investigated in *study 3.* The first controller regulates the number of vehicles in the region where the first core of congestion starts creating (which may be determined based on existing historical traffic data and partitioning algorithms) at the start of the peak period, by modifying the traffic light settings in some of the junctions at the perimeter of this small zone. In highly congested urban networks, the congestion may not be mitigated by activating the gating perimeter control just at the small protected network (PN_1). This may be addressed via a second perimeter control activated at the border of a larger part of the urban network (PN_2). *Study 3* also proposes some extensions to the distribution of the ordered controller flow to the associated traffic signals in case of low demand or occurrence of spillback which will be discussed in Chapter 5. As mentioned before The investigations are conducted for the urban road network of San Francisco, USA, in a microscopic simulation environment under realistic traffic conditions.

If the number of vehicles in PN_1 or PN_2 (N_1 or N_2) are allowed to grow beyond certain limits, the PNs' exit flows $q_{out,1}$ or $q_{out,2}$, which may be seen as a portion of the total PN flow, decreases (according to the NFDs) due to link queue spillovers and gridlock. To avoid this

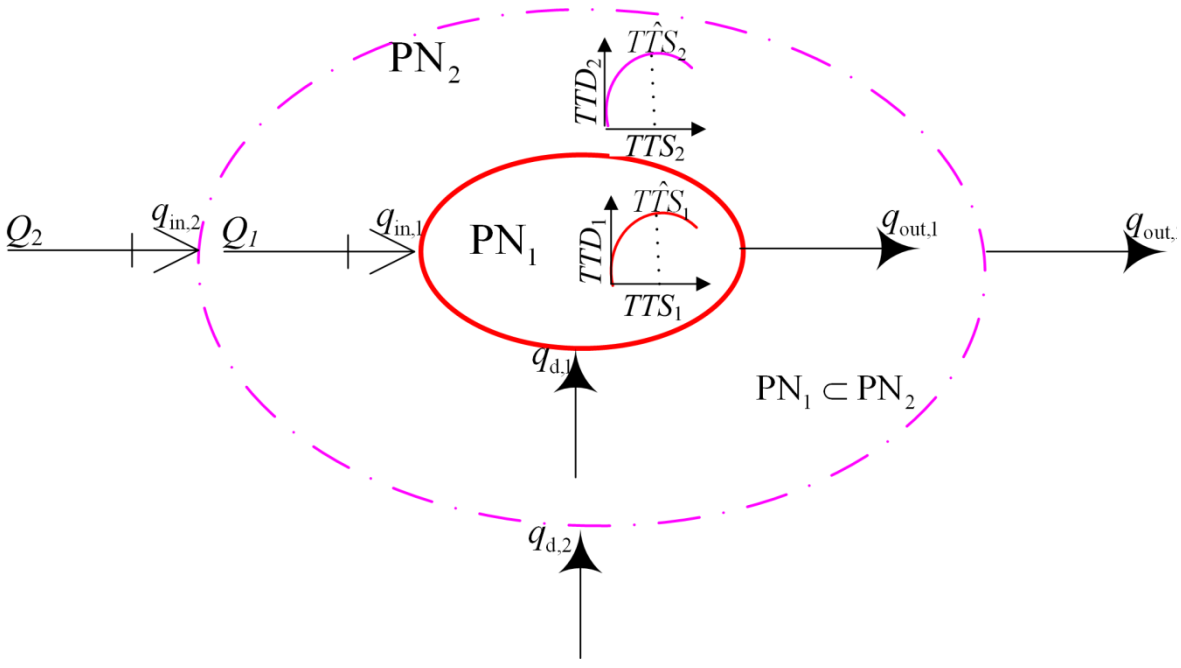

Figure 4-6 General scheme of the protected networks and the multiple concentric-boundary gating strategy at the perimeter of PN_2 (dashed pink line) and PN_1 (bold red line)

degradation, gating should reduce the PNs' inflows $q_{\text{in},1}$ and $q_{\text{in},2}$ appropriately, so as to maximize the PN throughput (see Figure 4-6).

The gating control problem is to regulate the TTS_1 and TTS_2 (number of vehicles) in the PN_1 and PN_2, respectively, so as to maintain them around their selected set-points $\hat{TTS}_1$ and $\hat{TTS}_2$ (see Figure 4-6), respectively, via appropriate manipulation of the gated flow ($q_{\text{g},n}$) in real time. By metering the flow at the perimeter of the PNs, queues start creating (Q_n).

To regulate TTS_n around the set-points, two PI regulators, the same type as (4-25), are implemented at the perimeter of each PN

$$q_{\text{g},n}(k) = q_{\text{g},n}(k-1) - K_{\text{P},n}\left[TTS_n(k) - TTS_n(k-1)\right] + K_{\text{I},n}\left[\hat{TTS}_n - TTS_n(k)\right] \qquad n = 1,2 \quad (4\text{-}43)$$

where *n* indicates the corresponding variables of the first and second controllers employed for controlling PN_1 and PN_2, respectively. K_{P} and K_{I} are the proportional and integral gains, respectively, which may be derived by the method presented in the previous section; in any case, feedback regulators are quite robust to moderate parameter value changes; nevertheless recent automatic fine-tuning tools could prove useful in field implementation [34].

It should be noted that the flows calculated by the regulator eq. (4-43) are constrained similarly to the bounding procedure described in section 4.6 for regulator (4-25) by pre-specified minimum and maximum values to account for physical or operational constraints.

4.6.3. Time-Delayed Feedback-Based Gating

Systems, in one sense, are devices that receive an input and produce an output. A system can be thought to operate on the input to produce the output. The output is related to the input by a certain relationship known as the system response. The system response usually can be modeled with a mathematical relationship between the system input and the system output. A step response in our case is possible in simulation, but virtually impossible in practice due to disturbances which would alter the step response. The μ, ζ and *m* which were derived from the model identification in section 4.6.1.1 are applied for the controller design in this section.

In all of the aforementioned gating strategies, gating is applied at the border of the protected network (PN). In this section, it is assumed that the gating strategy is applied further upstream of the PN and a robust feedback controller is designed by considering a time-delay

term, which corresponds to the travel time needed for gated vehicles to approach the PN (when the gating link is not directly at the PN perimeter).

According to eq. (4-39), we have a first order time-delayed system as following

$$\Delta TTS(k+1) = \mu \cdot \Delta TTS(k) + \zeta \cdot \Delta q_{\mathrm{g}}(k-m)$$

The z-transform function of the process is

$$P(z) = \frac{\zeta}{z^m (z-\mu)} \tag{4-44}$$

This corresponds to a continuous-time system (excluding the zero-order hold circuit) of

$$P(s) = \rho e^{-smT} \frac{\alpha}{s+\alpha} \tag{4-45}$$

where $\alpha = -\dfrac{\ln \mu}{T}$, $\rho = \dfrac{\zeta}{(1-\mu)}$ and T is the sampling time.

$$\Delta \dot{TTS} = -\alpha \cdot \Delta TTS + \rho \cdot \alpha \cdot q_{\mathrm{g}}(t-mT) \tag{4-46}$$

By applying the PI regulator (4-25), according to eq. (4-29), the controller z-transform function is

$$C(z) = \frac{z \cdot (K_{\mathrm{P}} + K_{\mathrm{I}}) - K_{\mathrm{P}}}{z-1} = K' \cdot \left(\frac{z - \dfrac{K_{\mathrm{P}}}{K'}}{z-1} \right) \tag{4-47}$$

where $K' = K_{\mathrm{P}} + K_{\mathrm{I}}$.

By closing the loop with C (controller) from (4-48) and P (process transfer function) from (4-44) we have

$$F_C = \frac{C(z) \cdot P(z)}{1 + C(z) \cdot P(z)} = \frac{K' \cdot \left(\dfrac{z - \dfrac{K_{\mathrm{P}}}{K'}}{z-1} \right) \cdot \left(\dfrac{\zeta}{z^m \cdot (z-\mu)} \right)}{1 + K' \cdot \left(\dfrac{z - \dfrac{K_{\mathrm{P}}}{K'}}{z-1} \right) \cdot \left(\dfrac{\zeta}{z^m \cdot (z-\mu)} \right)} \tag{4-48}$$

For $m = 0$, the reader is referred to dead-beat control design of section 4.6.1. Considering the time-delayed case ($m > 0$), a usual control design step, which reduces the complexity

(polynomial) of the closed-loop transfer function (F_C), is to specify the zero of the controller to be equal to the pole of the process, i.e. to set $K_{\mathrm{P}}/K' = \mu$; thus, the closed-loop transfer function (4-48) simplifies to

$$F_C = \frac{K'\zeta}{z^m\left(z-1\right)+K'\zeta} \tag{4-49}$$

The closed-loop transfer function (4-49) is identical with the one considered in Section 3.3 of [48], and hence, we can use the design rules of [49] to specify the second degree of freedom for our PI regulator for $m > 0$. Specifically, Figure 5 of [49] delivers the value of a parameter κ as a function of T/T_{D} . The correspondence between these quantities and our notation is $m = T_{\mathrm{D}}/T$ and $\kappa = K'\cdot\zeta\cdot m$. Thus, for any $m > 0$, we can use Figure 5 of [49] to obtain the corresponding value of κ and hence of K'. This value, combined with our above choice $K_{\mathrm{P}}/K' = \mu$, deliver the regulator parameter values displayed in Table 4-1.

For sufficiently long time-delays (i.e. $m \geq 4$), an alternative control design procedure for our discrete-time regulator is to consider a PI regulator design for the continuous-time system (4-46). With the approximation $\ln\mu \sim -\left(1-\mu\right)$, we obtain from (4-46)

$$\Delta T\dot{T}S = -\frac{1-\mu}{T}\cdot\Delta TTS + \frac{\zeta}{T}\cdot q_{\mathrm{g}}\left(t-mT\right). \tag{4-50}$$

By applying the Laplace transform (as in eq. (4-13)) to the continuous-time system

$$G\left(s\right) = \frac{\Delta TTS}{q_{\mathrm{g}}} = \frac{\frac{\zeta}{T}}{s+\frac{1-\mu}{T}} = \frac{\zeta T}{T\left(1-\mu\right)}\cdot\frac{1}{\frac{sT}{1-\mu}+1} \tag{4-51}$$

By applying the Chien et al. (1952) rules [10] (i.e. developed version of Ziegler and Nichols (1942) [63]) to this system, we obtain very similar values for K_{P} and K_{I} as in Table 4-1.

Table 4-1 Different values for K_{P} and K_{I} according to different time-delays (m)

m	K_{P}	K_{I}
0	μ/ζ	$(1-\mu)/\zeta$
1	$\mu/(3\zeta)$	$(1-\mu)/(3\zeta)$
2	$\mu/(5\zeta)$	$(1-\mu)/(5\zeta)$
3	$\mu/(6\zeta)$	$(1-\mu)/(6\zeta)$
>3	$\mu/(2m\zeta)$	$(1-\mu)/(2m\zeta)$

This confirms the pertinence of the previously derived rules, by use of a different design approach, for the cases with $m > 3$.

4.7. Gating Action at Gated Junctions/Links

According to Figure 4-7, the gating procedure at the gated signalized junctions can be listed as following

- The controller runs in background the whole period and is fed with real-time measurements of PN (*TTS*).
- A switch on/off logic decides for the implementation of the ordered flow (q_g) by the controller.
- During the switch on period, the ordered flow should be distributed among the gated junctions ($q_1,...,q_n$) where *n* is the number of gated junctions
- The distributed flows should be converted into green phases ($g_{g,1},...,g_{g,n}$) where g_g is the gated green.
- During the switch off period the fixed-time plan is implemented ($\bar{g}_1,...,\bar{g}_n$) where $\bar{g}$ is the nominal green and $\bar{q}$ is the flow which can be served during the nominal green.

4.7.1. Switch On/Off Logic

The switch on/off logic might be implemented only at the global level. Gating could be activated only within specific time-windows (e.g. at the peak periods) or if some real-time measurement-based conditions are satisfied. The regulator runs continuously in the background, but gating is actually activated only when *TTS* exceeds a threshold, that is selected slightly lower than (e.g. 85% of) the set point; and is de-activated when *TTS* falls below a 2nd, slightly lower threshold. Moreover, *TTS* smoothed trend, minimum non-switching time and presence of queue at the gated links could be also considered.

In this thesis, gating is activated when the *TTS* in PN exceeds 85% of $\hat{TTS}$ for more than two successive cycles and is deactivated when the *TTS* falls below 80% of $\hat{TTS}$ for more than 3 successive cycles. As long as no gating is applied (switch off), we assume that the considered

network operates under the fixed-time signal plan and the flow implemented is the nominal flow ($\bar{q}$) served during the nominal green.

4.7.2. Flow Distribution Among Gated Links

During the switch on period the ordered flow by the regulator should be distributed among the gated links (if more than one gated junction). The flow distribution problem is an important issue which might have a significant impact on the overall gating performance. Various methodologies may be applied to assign flows to different junctions.

4.7.2.1. Flow Distribution: *Version 1*

One could simply distribute the gated flow according to the saturation flow of the gated links, as following

$$\varphi_i = \frac{s_i}{\sum_{j=1}^{n} s_j} \tag{4-52}$$

$$q_i = \varphi_i \cdot q_g \tag{4-53}$$

where φ_i, s_i, s_j and q_i are distribution rate for the i^{th} gated junction, saturation flow of gated link i, saturation flow of the junctions (j=1,…, n; n is the total number of gated junctions) and the distributed gated flow of junction i, respectively. This flow distribution

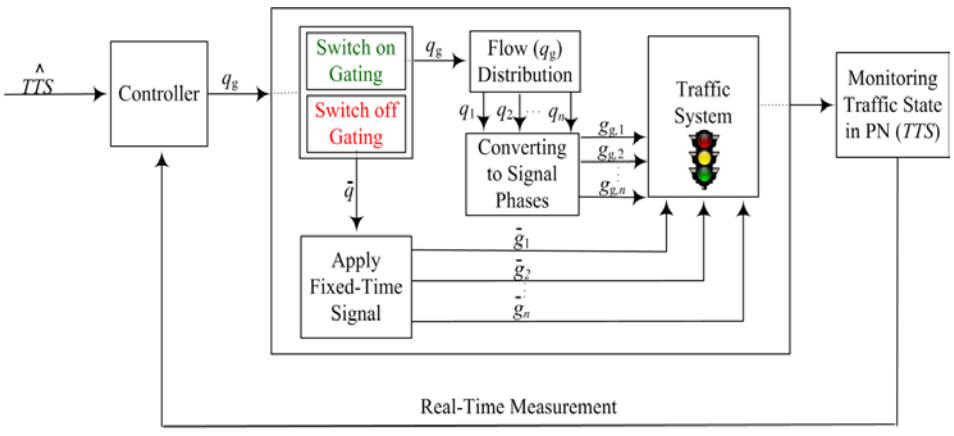

Figure 4-7 General gating procedure of a gated junction

version is applied for *study 1*, *study 2*, *study 4* and *study 5*, in this thesis.

A disadvantage of this policy could be ignoring the traffic state (i.e. queuing, lack of demand, waiting times etc.) at the gated links. This may lead to wasted green phases at junctions with spillback from downstream or lack of demand during specific time periods.

4.7.2.2. Flow Distribution: *Version 2*

To avoid the under-utilization of the gated traffic signals, in *version 1*, a new distribution policy is proposed which considers the real-time traffic state at the gated junctions. More specifically, the ordered flow (q_g) by the controller is split initially according to the respective saturation flow (by applying eq. (4-52) and (4-53)). But as discussed earlier, distributing the flows ordered by the regulators at the gated links without considering the possible spillback in the downstream links of the gated junctions and the queue length in the gated links in real-time may lead to wasted green time at the gated traffic lights. In other words, the controllers may introduce green phases, parts of which may not be used during the aforementioned situations. To reduce these impacts, in the new version, the number of vehicles waiting in the queue at the gated links are roughly estimated in every cycle and compared to the flow ordered by the controller and the actual flow passing the stopline. For this purpose, three loop detectors are installed in the gated links at distances of l_1 m(i.e. the queue length that can be dissolved by the minimum-green (e.g. 7s) in a cycle), l_2 m (in middle of the first and the third detector) and l_3 m (i.e. the maximum queue length dissolved during maximum-green in a cycle) from the stopline (see Figure 4-8). The estimated queue is obtained from the following equation:

$$Q_{ei}(k) = \left[\left(l_1 \cdot o_{e1i}(k)\right) + \left(l_2 \cdot o_{e2i}(k)\right) + \left(l_3 \cdot o_{e3i}(k)\right)\right] \cdot \mu_g \big/ \left(T \cdot 100\lambda\right) \cdot 3600 \qquad (4\text{-}54)$$

Where $Q_{ei}(k)$ is the estimated queue (in number of vehicles) of the *i* junction during *k*, $o_{e1i}(k)$, $o_{e2i}(k)$ and $o_{e3i}(k)$ are the measured time-occupancies (in %) in the gated link for the first, second and third detector of i^{th} junction during *k*, respectively, and μ_g is the number of lanes of the gated link. If the estimated queue is too short (i.e. less than the number of vehicles served by the minimum-green), e.g. due to low demand at the gated junctions, the assigned flow to these junctions will turn into the minimum flow (i.e. the flow served during

minimum green) for the next cycle; while the rest of ordered flow will be distributed among the rest of the gated junctions.

As pointed out before, spillback from downstream may at times reach the gated links, and, as a result, the ordered flow by the controller may not be fully utilized. To avoid this problem, junctions with spillbacks are identified in real-time by monitoring the estimated queue, actual flow passing the stop bar at the gated link and the calculated flow by the regulator each cycle. If the relative difference between the ordered and actual flows is more than a threshold, say 20%, and $Q_e > q_g$ in junction *i*, then the flow distribution algorithm will be alerted that the junction has over-spilling problem in the downstream link. Consequently, the calculated flow for that specific junction is reduced by some pre-fixed percentage, e.g. 20%. and this reduction is distributed among the other junctions. This version of flow distribution is investigated in *study 3* and tested on simulation *scenario 3*.

One of the disadvantages of this and other more sophisticated flow distribution versions may be the increased number of detectors in the network which might not be always economically feasible.

4.7.3. Converting the Distributed Flows into Green Phases

To execute the gating strategy at the gated junctions, the distributed ordered flow should be translated into green phases. This is done by

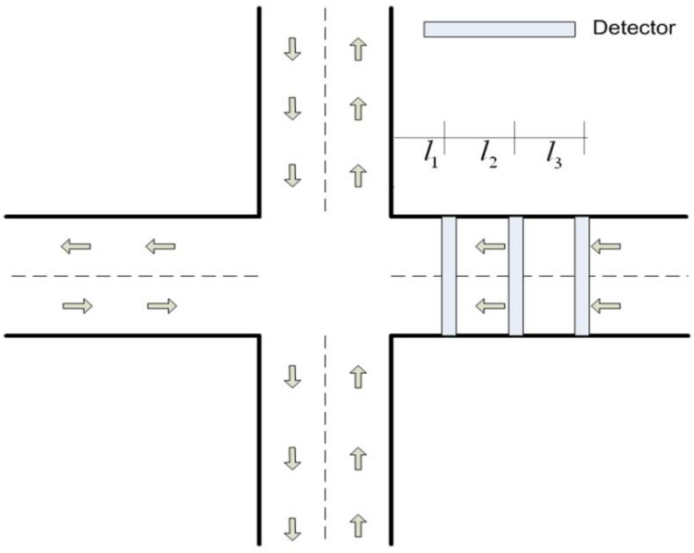

Figure 4-8 Detectors' location at the gated links

$$g_{g,i} = \frac{q_i \cdot C}{s_i} \tag{4-55}$$

where C and $g_{g,i}$ are the traffic light cycle length, and gated green time, respectively.

The fixed-time signal plan of the considered urban network is modified at the gating junctions when real-time gating actions are present, by modifying accordingly the involved signal stage (phase) durations. Specifically, the gating strategy determines the duration of one (gated) signal stage (within pre-specified bounds) in real-time; any (positive or negative) deviation of this stage duration from its fixed-time value is assigned to other stages of the same junction.

For a better understanding of what explained before, Figure 4-9 is demonstrated. It exhibits the fixed-time signal plan of the traffic light 3 at the perimeter of PN (shown in Figure 3-4).The *interphases* are shown by dark green along with the corresponding movements shown by arrows which remain constant during the control period. *Interphases 2* and *7* are all-red stages in which all movements at the junction receive red light. During gating action, the movement into the PN should be metered which belongs to *stage 5*. Thus, the gated signal stage (phase) is *stage 5* and any positive or negative deviation of *stage 5* during the gating activation, is assigned to *stage 8*. By this, the cycle length is kept constant during the gating period. Generally, the stage re-arranging should be such that $g_{\min,i} < g_{g,i} < g_{\max,i}$ where $g_{\max,i} = \sigma_{\text{gated},i} + \sigma_{\text{non-gated},i} - g_{\min,i}$ and σ is the stage duration. In case of existence of more than 2 stages, the positive or negative deviation of the gated stage either can be assigned to the stage which involves the movement of exiting flow from the PN or can be assigned proportionally to the other stages based on $\overline{g}_n$ (nominal green phases).

If a gated link (whose outflow enters the PN) receives its right-of-way in the same stage as an opposed-direction link (whose outflow exits the PN, e.g. in *stage 5* in Figure 4-9), an appropriate re-staging has been introduced to avoid unwanted and unnecessary delays at PN-exiting links during the gating activation. More specifically, an extra 0s stage is defined in the signal plan of that junction. This additional stage will be active during the gating period (in the signal cycle) in order to serve the green phase dedicated to the outflow exiting the PN. This phase separation is carried out as following

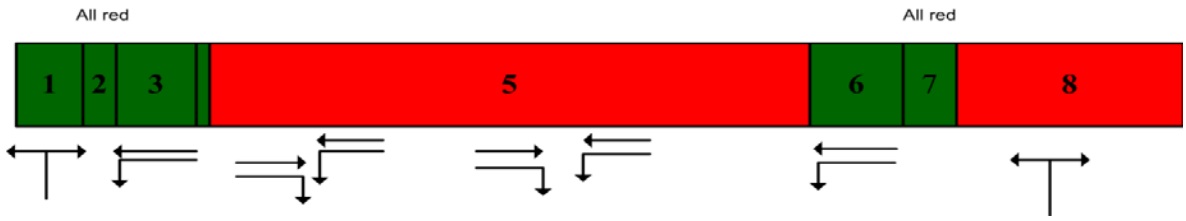

Figure 4-9 Signal plan of the junction 3 in Figure 3-4

$$g_{g,i} < \overline{g}_i \Rightarrow \begin{cases} \sigma_{\text{gated},i} = g_{g,i} \\ \sigma_{\text{added},i} = \overline{g}_i - g_{g,i} \\ \sigma_{\text{non-gated},i} = g_{\max,i} + g_{\min,i} - \overline{g}_i \end{cases} \quad (4\text{-}56)$$

$$\overline{g}_i \le g_{g,i} \Rightarrow \begin{cases} \sigma_{\text{gated},i} = g_{g,i} \\ \sigma_{\text{added},i} = 0 \\ \sigma_{\text{non-gated},i} = g_{\max,i} + g_{\min,i} - g_{g,i} \end{cases} \quad (4\text{-}57)$$

Note that, except *study 1*, all the studies (i.e. *study 2 to study 5*) include the latter re-staging concept.

Chapter 5

Discovery consists of seeing what everybody has seen,
and thinking what nobody has thought.

Albert Szent-Györgyi

5. Simulation-Based Results

The microscopic simulator AIMSUN is stochastic, thus different simulation runs (replications) with different random seeds may lead to different results. For this reason, it is common to use a number (10 in this thesis) of replications for each investigated scenario and then calculate the average value of the 10 runs for each evaluation criterion in order to compare different control scenarios with non-gated scenarios. Mainly, three performance indexes are used here (as provided by AIMSUN): the average vehicle delay per km and the mean speed, both for the entire implemented urban network (not only the PN); and the total number of vehicles that exit the overall network during the whole scenario. Moreover, *study 3* introduces a new performance index which may be interpreted as average *TTS*/*TTD* (s/km) or average unit travel time that considers the waiting time in the virtual queue (i.e. vehicles waiting out of the network) in the calculation of the *TTS*.

In this chapter, at the beginning, simulation *scenarios* (presented in chapter 3) applied for different studies are introduced briefly. Right after, the simulation results of the proposed feedback-based gating strategies in different studies, tested on different simulation *scenarios*, are presented.

5.1. Simulation Scenarios Applied for Different Studies

For each of the studies introduced in Chapter 2, a specific simulation scenario (presented in Chapter 3) is implemented. The proposed gating strategies in this thesis are studied on the following simulation scenarios:

- *Study 1*: Single boundary gating control by exploiting *complete* operational NFD $\rightarrow$ *Scenario 1*
- *Study 2*: Single perimeter gating control by applying *reduced* operational NFD $\rightarrow$ *Scenario 2*
- *Study 3*: Multiple concentric boundaries gating control $\rightarrow$ *Scenario 3*
- *Study 4*: Gating control remote form the PN $\rightarrow$ *Scenario 4*
- *Study 5*: Remote gating control of PN with bigger control steps $\rightarrow$ *Scenario 4*

5.2. *Study 1*: Single Perimeter Feedback-Based Gating

In this study, the proposed single perimeter gating control is tested on Chania urban network (Figure 3-4), by exploiting the complete NFD of the PN.

5.2.1. Complete Operational NFD of PN

Figure 5-1(a) displays the (operational) NFD for the Chania PN (assuming that all links are detector-equipped, i.e. $\mathbb{M} = \mathbb{Z}$) for the first 2 hours of the employed scenario, i.e. the period during which the network is filled, and the congestion is created; 10 different replications (each with different seed in AIMSUN) were carried out. To build a comprehensive NFD that includes free-flow conditions, the specified demand starts from very low values and increases gradually to levels that lead to heavy congestion in PN (as under typical real traffic conditions at the peak periods); eventually, the demand is gradually reduced until the network is virtually emptied at the end of the simulation (see Figure 5-1(b)). Figure 5-1(a) demonstrates that a fundamental diagram (inverse-U) shape is indeed occurring during the 2-h network filling period, with quite moderate scatter even across different replications; Figure 5-1(b) indicates that the inverse-U shape appears also during the decreasing demand period of 2 h, albeit with a visible hysteresis compared to the filling 2-h period. The hysteresis is limited and is due to different link load patterns that prevail in the emptying period compared to the filling period. Whatever the exact NFD (and despite some limited scatter), it can be seen in Figure 5-1 that the maximum *TTD* values in the diagram occur in a *TTS* region of 600 to 800 veh·h per h. If *TTS* (i.e. *N*) is allowed to increase beyond this limit, then *TTD* (and hence the PN throughput) decreases; this leads to an unstable escalation, as long as the PN inflows continue to be higher, that degrades increasingly the PN throughput and efficiency, leading them to accordingly low levels (or even to zero in the extreme total-

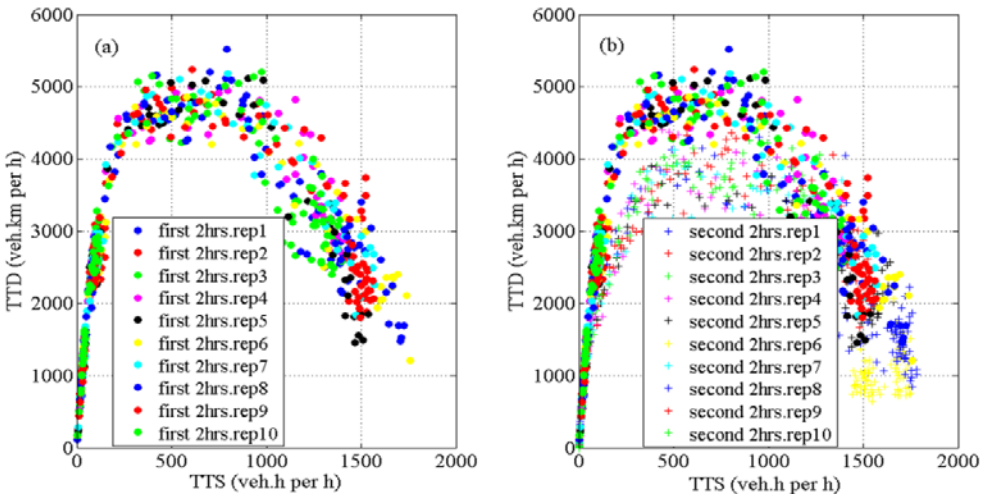

Figure 5-1(a) NFD of the PN for the first 2 hours for 10 replications; (b) NFD of PN for the 4 hours simulation for 10 replications

gridlock case). To avoid this unstable degradation and, in fact, maximize the PN throughput and efficiency, the PN's *TTS* should be maintained in the mentioned optimal range, and this is exactly the goal pursued in this work.

5.2.2. Non-Gating Scenario

Signal control for the non-gating case corresponds to the usual fixed-time settings used in the real Chania network. Table 5-1 displays the aforementioned evaluation indexes for every replication; as well as the average, maximum and minimum values of each index. Since AIMSUN is a stochastic tool, link over-spilling and partial gridlocks may be more or less pronounced in individual replications. In some replications (e.g. R2 and R6 in Table 5-1), the created congestion in the peak period leads to more serious gridlocks in the PN, consequently the delay is higher and the mean speed is lower than average, while the (lower) number of exited vehicles in these replications indicates that the network is not yet empty at the end of the simulation.

Table 5-1 Non-gating performance indexes for each replication

	R1	R2	R3	R.4	R5	R.6	R7	R8	R9	R10	Ave.	Max	Min
Delay (s)	513	722	362	360	431	674	467	372	350	424	467	722	350
Speed (km/h)	6.62	5.05	8.91	8.88	7.81	5.55	7.25	8.84	9.06	7.83	7.58	9.06	5.05
Vehicles out	14545	12997	14575	14593	14771	11092	14684	14719	14338	14844	14115	14844	11092

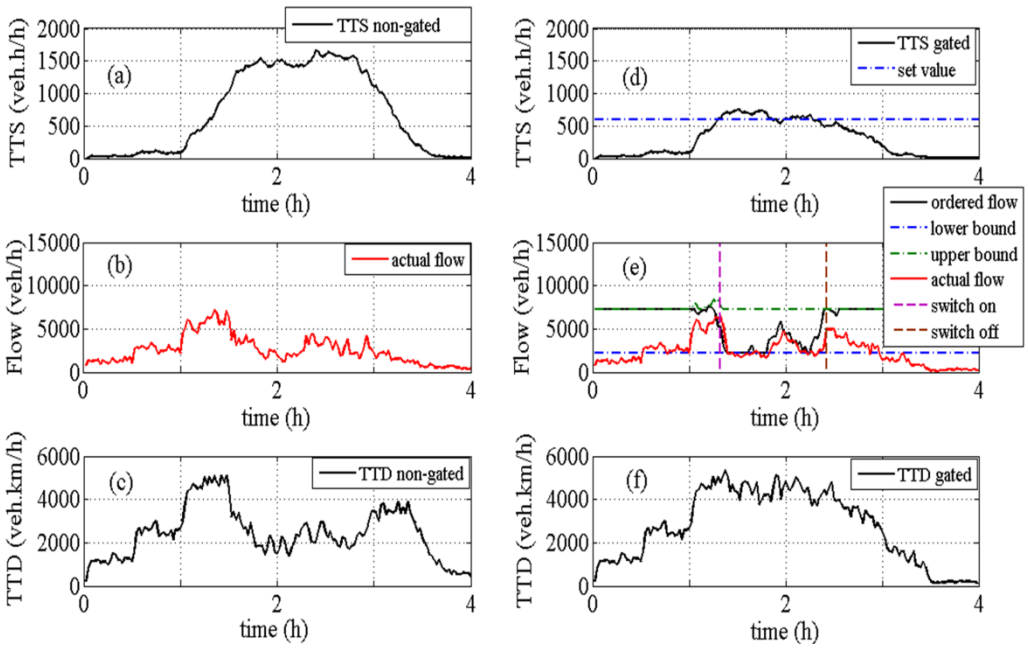

Figure 5-2 (a) PN's *TTS* vs. time in non-gating case; (b) actual PN inflow vs. time for the non-gating case; (c) PN's *TTD* vs. time for the non-gating case; (d) PN's *TTS*vs. time for the gating case; (e) ordered and actual PN inflow vs. time for the gating case; (f) PN's *TTD* vs. time for the gated case

To enable an illustrative comparison of the non-gating versus gating cases, the respective detailed results of Replication 1 are displayed in Figure 5-2, namely the PN's *TTS* (Figure 5-2(a) and (d)), the PN inflow (from 8 gated links) q_{in} (Figure 5-2(b) and (e)), and the PN's *TTD* (Figure 5-2(c) and (f)). Concentrating on the left column of Figure 5-2 (parts (a), (b), (c)), the 1st hour is characterized by a gradual increase of all three displayed quantities, as it is typical for increasing demand in under-saturated traffic conditions. At time $t = 1$ h, the abrupt increase of q_{in} leads to according increases of *TTS* and *TTD*, the latter reaching soon capacity values according to Figure 5-1(a), while the former is traversing the aforementioned critical region of [600, 800]. However, as q_{in} continues to be high, *TTS* continues to increase to very high values (i.e. the PN becomes increasingly crowded with vehicles); as a consequence, link over-spilling and gridlock lead to a sensible *TTD* reduction to low levels, that are persisting until $t = 3$ h, when the network starts de-congesting due to low demand. Remarkably, during the congested period $t \in [1.5\text{h}, 3\text{h}]$, the inflow q_{in} is also reduced due to over-spilling links of PN, i.e. as a result of the congestion that extends beyond the PN; but this reduction of the inflow comes too late, too little to reverse the already advanced PN degradation.

5.2.3. Gating Scenario

The usual off-peak signal plan is modified at the gating junctions when real-time gating actions are present, by modifying accordingly the involved signal stage durations. Re-arranging the signal stages is carried out according to section 4.7.3.

In *study 1*, since the gated junctions are at the perimeter of the PN (see Figure 3-4) the time delay τ is zero, hence the K_P and K_I values may be specified for the controlled system (4-24), (4-25) to exhibit a time-optimal dead-beat regulator behavior (see section 4.6.1). A least-squares parameter estimation is first conducted for μ and ζ in (4-24), using time-series of (q_{in}, *TTS*)-measurements within and around the critical *TTS*-range of [600, 800] (according to section 4.6.1.1). Once the parameters μ and ζ have been specified ($\mu = 0.807$ and $\zeta = 0.038$), the regulator parameters for (set-point) dead-beat behavior were calculated to be $K_P = 20$ h^{-1} and $K_I = 5$ h^{-1} (from Table 4-1, for m=0). Note anyhow that the feedback controller is quite robust to parameter variations, as it can be verified from the stability and optimality conditions provided.

The regulator's maximum and minimum bounds are visible in Figure 5-2(e). The regulator runs continuously in the background, but gating is actually activated only when *TTS* exceeds a threshold, that is selected slightly lower than (in this case 85% of) the set point; and is de-activated when *TTS* falls below a 2nd, slightly lower threshold. At all other times, fixed-time signal control is applied, as in the non-gating case. A set point of $\hat{TTS} = 600$ veh·h per h is selected for the gating operation.

By running AIMSUN with the control law (4-25) for the gated traffic signals, the results displayed in Table 5-2 are obtained for each of the 10 replications. The improvements, compared to Table 5-1, are significant, and, in fact, even the worse gating replication is superior to the best non-gating replication. An average speed increase of 40% results for the whole network (not just the protected part thereof).

Figure 5-2(d), (e), (f) display the detailed results of Replication 1 of the gating case and illustrate its way of functioning and impact. Traffic conditions are identical as in the non-gating case up to around $t = 1.2$ h, when gating is switched on (Figure 5-2(e)), as *TTS* approaches its set value; the gating regulator orders low inflow values to maintain *TTS* around its set point, and, as a consequence, *TTD* is maintained at high levels, in clear contrast

to the non-gating case. As mentioned earlier, the gating action creates temporary queues (and corresponding temporary vehicle delays) at the gated links; however, this proves highly beneficial for the PN throughput, and, as a consequence, the overall network delay (including the gating queue regions) is strongly reduced, as Table 5-2 indicates; in fact, even gated vehicles may have a net gain, as their temporary gating delays may be more than offset by enabled higher speeds once they enter the PN.

It is visible in Figure 5-2(e) that the q_{in} values ordered by the regulator, differ from the implemented ones for various reasons mentioned earlier, but this has a minor impact on the regulator's efficiency, as expected. At $t = 2.3$ h, *TTS* is moving to lower values, gating is switched off, and traffic flow returns to under-saturated conditions; in contrast to the non-gating case where over-saturated conditions are seen in Figure 5-2 to persist for 1 hour longer.

As mentioned before, Figure 5-2(e), illustrates the summation of ordered and actual flow in all eight gated links versus the simulation time period. The red line (actual flow) and the black line (ordered flow) are not matching all the time (see for instance [1.8h, 2h]) in Figure 5-2(e). Thus, for a higher resolution and better understanding of the traffic condition in each and every gated junction, numbered from 1 to 8 in Figure 3-4, Figure 5-3 is plotted. Figure 5-3 displays the distributed ordered flow and the actual flow versus time in every gated junction. The gap between the actual and ordered flow seems to be significant in junctions 1 and 2. In junction 1, at time t=1.8h the demand starts decreasing and the difference between the actual and ordered flow increases, consequently. On the contrary, junction 2 faces high demand during the gating period. This can be seen by the high values of actual flow after the deactivation of gating (i.e. at t=2.2h). According to the visual inspection during the simulation, the main cause of the gap between the red and black line in this case is spillback from the downstream junctions. In other words, due to the blocked downstream

Table 5-2 Performance indexes using the proposed gating control strategy

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Ave.	Max	Min
Delay (s)	293	287	310	298	317	310	328	314	299	292	304	328	287
Upgrade (%)	42.8	60.2	14.4	17.2	26.5	54.0	29.8	15.6	14.6	31.1	35.0	54.0	14.4
Speed (km/h)	10.3	10.6	9.9	10.4	10.0	10.0	9.4	9.9	10.3	10.4	10.1	10.6	9.4
Upgrade (%)	55.6	110	11.1	17.1	28.0	80.2	29.7	12.0	13.7	32.8	39.2	110	11.1
Vehicles out	14721	14521	14623	14516	14508	14569	14783	14632	14441	14515	14582	14783	14441

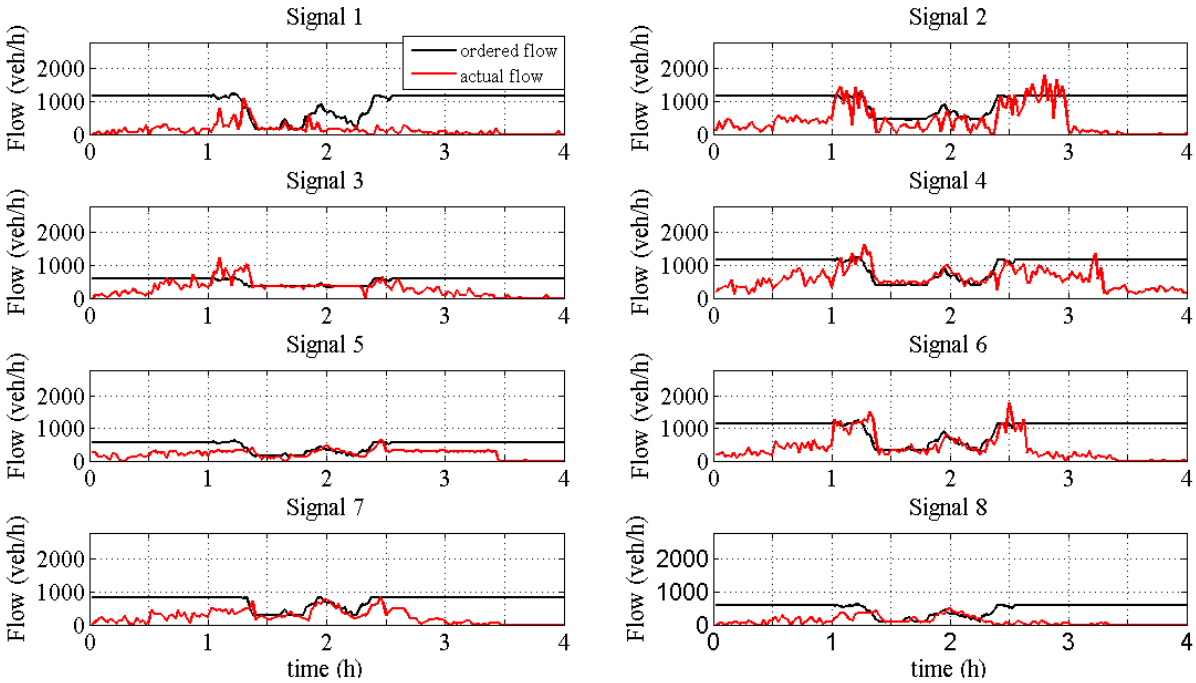

Figure 5-3 Ordered and actual flow at the eight gated junctions individually

links, this junction cannot serve as many vehicles as ordered by the regulator during some cycles (e.g. zero flow at t=1.4, 1.5 and 1.8). For the rest of the gated junctions, the actual and ordered flows almost match during the gating period. Again, the observed deviations of the ordered versus the actual flows are within reasonable bounds and do not affect the final gating outcome thanks to the feedback character of the regulator.

Figure 5-4 displays the simulation results obtained for Replication 1 using the bang-bang controller (4-26). The resulting delay is 341 s, which is higher than for any replication with the PI controller (4-25) (Table 5-2), but lower than for any replication without gating (Table 5-1). Figure 5-4(a) indicates that the bang-bang controller maintains the corresponding *TTS* near the set-point; while Figure 5-4(b) displays the expected bursty behavior of the bang-bang controller with regard to the controlled inflow.

Figure 5-5 displays the simulation results obtained for Replication 1 using the PI controller (4-25), but now with an increased lower bound of $q_{\min} = 4000$ veh/h (in place of 2180 veh/h used before). The regulator is seen to saturate at the new lower bound (i.e. it applies the maximum admissible gating action) for most of the active gating period. Figure 5-5 indicates that the actual inflow deviates slightly, but increasingly with time, from the ordered flow $q_{\min}$, which is due to spillback from downstream link queues at a couple of gating locations.

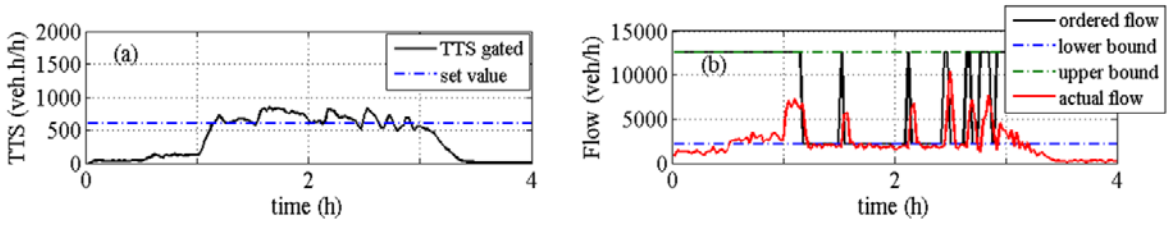

Figure 5-4 Bang-bang gating results: (a) *TTS* vs. time; (b) ordered and actual PN inflow vs. time

Overall, the constrained gating action is now not enough to maintain *TTS* close to its set-point because $q_{\min}$ is too high. Thus, except for two tiny periods of time at the start and the end, respectively, of the active gating period, *TTS* is essentially uncontrolled. Nevertheless, this constrained gating action is sufficient to enable lower *TTS*-values than without gating; in fact, *TTS* reaches a maximum of slightly above 1000 veh, as opposed to 1600 veh in the corresponding no-control case. The resulting delay is 330 s, which is higher than for any replication with lower $q_{\min}$ (Table 5-2), but lower than for any replication without gating (Table 5-1).

5.3. *Study 2*: Single Perimeter Gating Control by using Reduced NFDs

This section continues and extends the study carried out in section 5.2 by demonstrating that efficient feedback-based gating is actually possible with much less real-time measurements, i.e. at a drastically lower implementation cost. This is a significant addition that opens the way for real implementations of the method due to the substantially reduced cost implied. For this study, the simulation *scenario 2* is implemented.

5.3.1. Reduced Operational NFDs

The complete operational NFD of the Chania PN presented in Section 2 (assuming that all links are detector-equipped, i.e. $\mathbb{M} = \mathbb{Z}$), is obtained via a 4-hour AIMSUN simulation with dynamic traffic assignment based routing, and is displayed in Figure 5-6(a) and (b); ten different replications (each with different seed in AIMSUN) of the 4-hour scenario were carried out and are included (with different colors) in Figure 5-6. Figure 5-6(a) displays the (90 s cycle-based) measurement points for the first 2 hours of the employed demand scenario, i.e. the period, during which the network is filled, and congestion is created; while in

Figure 5-6(b) also the last 2-hour measurements (indicated by +), during which the network is emptied, are added. To build a comprehensive NFD that includes free-flow conditions, the specified demand starts from very low values and increases gradually to levels that lead to heavy congestion in PN (as under typical real traffic conditions at the peak periods); eventually, the demand is gradually reduced, until the network is virtually emptied at the end of the simulation (see Figure 5-6(b)). Figure 5-6(a) demonstrates that a fundamental diagram (asymmetric inverse-U shape) is indeed occurring during the 2-h network filling period, with moderate scatter even across different replications; Figure 5-6(b) indicates that the inverse-U shape appears also during the decreasing demand period of 2 h, albeit with a visible hysteresis compared to the filling 2-h period. The hysteresis has a limited width and is due to different link load patterns that prevail in the emptying period compared to the filling period. What is most relevant for gating control and can be clearly seen in Figure 5-6(a) and (b), is that the maximum *TTD* values in the diagram occur in a *TTS* region of 600 to 900 vehh per h. If *TTS* (i.e. *N*) is allowed to increase beyond this limit, then *TTD* (and hence the PN throughput) decreases; this leads to an unstable escalation, as long as the PN inflows continue to be high, that degrades increasingly the PN throughput and efficiency, leading them to accordingly low levels (or even to zero in the extreme total-gridlock case).

It should be noted that the complete operational NFD calls, by its definition, for traffic measurements collected from each and every (even secondary, non-signalized) link in the network, which is unusual in practice (although this may change in future, thanks to probe car measurement technologies). Real networks may feature traffic detectors at the links approaching signal-controlled junctions, e.g. to facilitate real-time UTC; in other cases, only some strategic detectors are installed at few network links, e.g. as a basis for real-time selection of pre-specified signal plans. Last not least, in the case of urban networks without any monitoring equipment, the following important questions arise: Is it necessary to equip all links of a PN with detectors in order to enable efficient feedback gating actions that

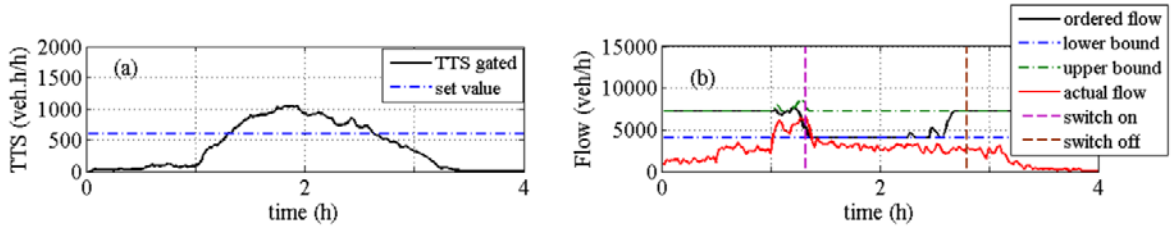

Figure 5-5 Higher lower bound results: (a) *TTS* vs. time; (b) ordered and actual PN inflow vs. time

exploit the NFD notion according to? If not, how many links should be equipped? Appropriate answers to these questions have obvious implications for the cost, and hence the applicability, of feedback-based gating strategies. It is the main focus of this part of the study to investigate these significant questions and their implications.

As a starting point for addressing these questions, reduced operational NFDs, i.e., based on fewer measurements from the Chania PN, have been derived (see Figure 5-6(c)-(h)). At a first step of this investigation, all links approaching signal-controlled junctions inside the PN have been selected for deriving a correspondingly reduced operational NFD. This leads to a number of 57 out of 165 links, i.e. 35% of the measurements (see Figure 5-7 for the location of the detectors in the PN) used for a complete operational NFD (see Figure 5-6(c) and (d)). We will call this reduced operational diagram NFD(35%) for brevity.

Since 57 links may still be a high number in practice for this size of urban network, we proceed, in the next steps of our investigation, to further reductions of the percentage of monitored PN links to 10% (17 links) and to 5% (8 links) of the whole link population in the PN, where the 5% links is a subset of the 10% links (see Figure 5-7). The specific selection of equipped links (out of the 57 signal-controlled links) was based on visual inspection of the protected network during the simulation. More specifically, the selection included, in a first step, the most critical links, i.e. links where the congestion starts spreading at the peak period; in a second step, the signal-controlled PN links which are deemed most representative for the traffic conditions in the PN were added to complete the targeted percentage of equipped links. The obtained reduced NFDs are displayed in Figure 5-6(e) and (f) and Figure 5-6(g) and (h) for 10% and 5% of measurement-equipped links, respectively. Interestingly, a similar attempt to derive representative NFDs by use of a sub-set of network link measurements was undertaken independently by [45]; albeit without actually exploiting the resulting reduced NFDs for gating control as in this work.

Clearly, the range of the obtained *TTS* and *TTD* values across the four diagrams (Figure 5-6(a)-(h)) is different, because of the corresponding differences in the number of measurements that are summed up in eq. (2-1) and (2-2). We will call the (reduced) quantities of a NFD(x%), *TTS(x%)* and *TTD(x%)* for brevity. The displayed NFDs for different percentages of measurements (100%, 35%, 10% and 5%), indicate that the respective maximum *TTD* occurs in a *TTS* region of 600 to 900, 350 to 450, 150 to 200, 90 to 130 vehh per h, respectively.

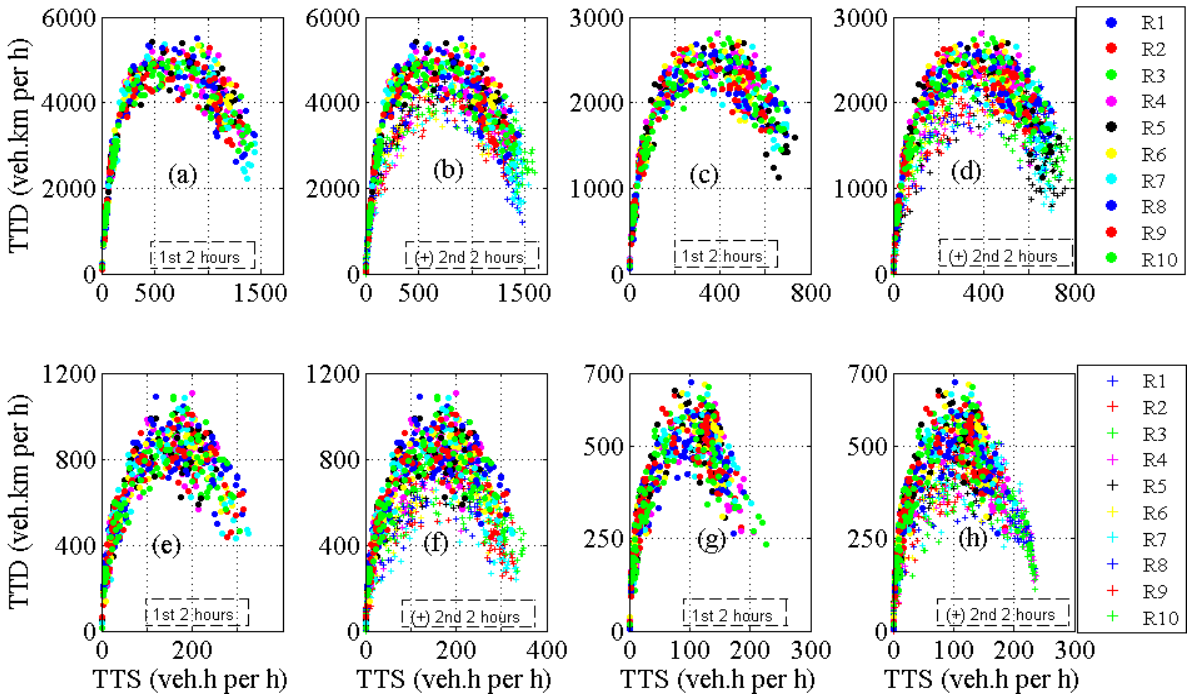

Figure 5-6(a) NFD of the PN for the first 2 hours for 10 replications; (b) NFD of PN for the 4 hours simulation for 10 replications (100% measurements); (c) NFD of the PN for the first 2 hours for 10 replications; (d) NFD of PN for the 4 hours simulation for 10 replications (35% measurements); (e) NFD of the PN for the first 2 hours for 10 replications; (f) NFD of PN for the 4 hours simulation for 10 replications (10% measurements); (g) NFD of the PN for the first 2 hours for 10 replications; (h) NFD of PN for the 4 hours simulation for 10 replications (5% measurements)

At this point, it is important to highlight a potential dilemma that stems from the following facts: (a) The targeted goal is to maximize the *TTD* of the protected network as a whole, not merely the *TTD* occurring at a subset of PN links; but (b) we wish to achieve this goal on the basis of measurements collected from a subset of PN links, i.e. based on the knowledge of the accordingly reduced NFD. To address this dilemma, we need to investigate if, at the time the reduced *TTS* (with less measurements) is within its critical value range, the corresponding *TTS(100%)* is within its own critical value range as well. This investigation confirmed that in all cases and replications, the critical *TTS* ranges are indeed reached approximately simultaneously, specifically within the time interval of [1.4h, 1.5h]. Note that this outcome is in agreement with a sufficient condition for the existence of an NFD, namely the space-homogeneity of congestion in network links (see [26]). In fact, if this homogeneity condition holds true, then any subset of network links is representative for the congestion status in the whole network.

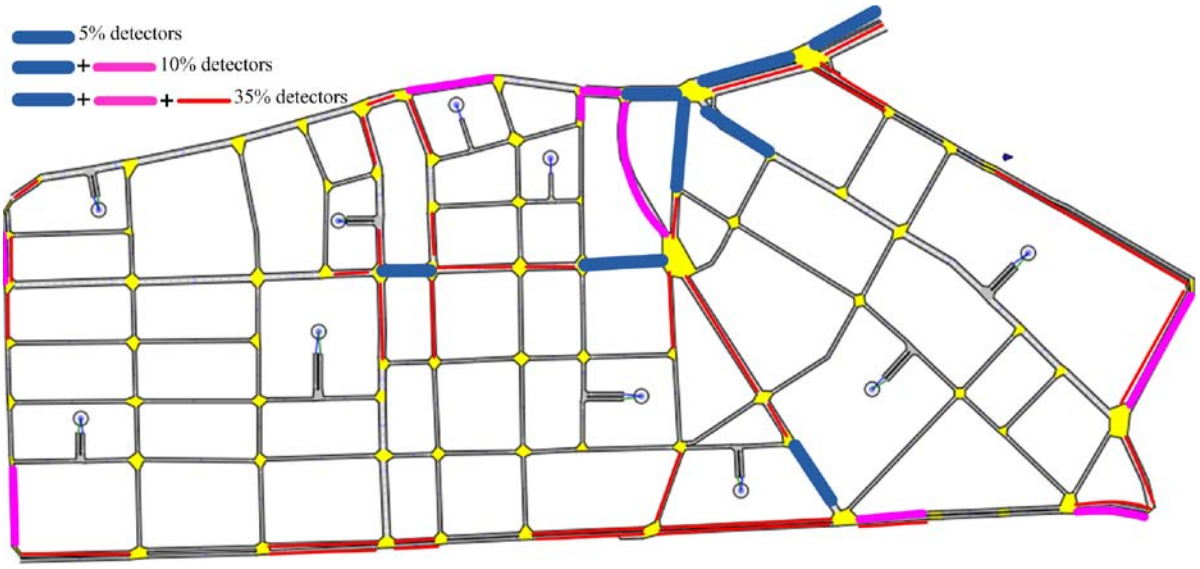

Figure 5-7Links considered for deriving reduced operational NFD: The thick blue lines indicate the links used for the case of 5% measurements; the thick blue and pink medium-width lines show the links considered for the case of 10% measurements; and the thin red lines plus the aforementioned colors reflect the case with 35% measurements

The finding of the above investigation is important for the practical deployment of NFD-motivated feedback gating strategies, because it suggests that throughput-maximizing gating can actually be applied on the basis of a small percentage of detector-equipped links, i.e. with far less cost than what would be necessary for obtaining the complete operational NFD of the protected network. More specifically, the investigation indicates that targeting a reduced *TTS* (with less measurements) to be in its optimal region, is practically equivalent to targeting the complete *TTS(100%)* to be in its own optimal region, which is known (by the NFD notion) to lead to a maximum PN throughput. The testing of this practically significant outcome via actual application of feedback gating actions under different instrumentation scenarios will be considered in the next sections.

5.3.2. Non-Gating Scenario

The applied signal control in the non-gating case corresponds to the usual fixed-time signal settings used in the real Chania network during the peak period. Table 5-3(a) displays the evaluation indexes for every replication. The last four columns display the average, maximum, minimum and standard deviation (S. D.) for each index. The standard deviation

Table 5-3 (a) Non-gating performance indexes for each replication; (b) Performance indexes using the proposed gating control strategy with different measurement percentages (*: in 1000)

		Index	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Ave.	Max	Min	S.D.
(a) Non-gated		Delay (s/km)	427	327	271	321	279	289	458	253	352	438	342	458	253	75
		Speed (km/h)	7.2	8.9	10.4	9.1	10.2	9.9	6.8	11.0	8.4	7.0	8.9	11.0	6.8	1.5
		Veh. out*	15.6	15.7	15.7	15.9	15.7	15.8	15.7	15.9	15.8	16.0	15.8	16.0	15.6	0.1
(b) %Measurement	100%	Delay (s/km)	244	239	223	242	234	249	257	236	241	239	241	257	223	8.5
		Speed (km/h)	11.3	11.5	12.1	11.4	11.7	11.1	10.9	11.6	11.4	11.5	11.5	12.1	10.9	0.3
		Veh. out*	15.7	15.7	15.6	16.0	15.6	15.9	15.8	16.0	15.8	15.9	15.8	16.0	15.6	-
	35%	Delay (s/km)	248	219	215	238	217	241	230	226	229	249	231	249	215	11.8
		Speed (km/h)	11.1	12.0	12.4	11.5	12.2	11.4	11.8	12.0	11.8	11.1	11.7	12.4	11.1	0.42
		Veh. out*	15.8	15.8	15.7	15.9	15.7	15.8	15.7	15.9	15.7	15.9	15.8	15.9	15.7	-
	10%	Delay (s/km)	241	229	239	251	236	246	258	237	235	262	243	262	229	10.0
		Speed (km/h)	11.4	11.8	11.5	11.1	11.6	11.2	10.8	11.5	11.6	10.7	11.3	11.8	10.7	0.34
		Veh. out*	15.9	15.6	15.8	16.0	15.7	15.8	15.7	15.8	15.9	16.0	15.8	16.0	15.6	-
	5%	Delay (s/km)	249	220	238	262	238	251	241	238	254	232	242	262	220	11.9
		Speed (km/h)	11.2	12.3	11.5	10.7	11.5	11.1	11.4	11.5	11.9	11.8	11.5	12.3	10.7	0.44
		Veh. out*	15.8	15.7	15.7	16.0	15.8	16.0	15.8	15.8	15.9	16.0	15.9	16.0	15.7	-

was included as an indicator reflecting the reliability of the traffic conditions across the different replications. In fact, the reliability of the traffic conditions, i.e. their similarity from day to day, is deemed to be a significant network quality criterion because it enables the road users to plan their trips and activities in an accordingly reliable way. Although different AIMSUN replications are not necessarily equivalent to different days in a real network, a possible reduction of the standard deviation in the gated case could be interpreted as an ability of the control action to reduce the aforementioned stochastic or chaotic traffic characteristics, hence to potentially increase the reliability, of the network traffic process.

To enable an illustrative comparison of the non-gating versus gating cases, the detailed results of one specific replication (R2) (for the 100% measurement case) are displayed in Figure 5-8, namely the PN's *TTS* (Figure 5-8(a) and (d)), the PN inflow (from the 8 gated links) q_{in} (Figure 5-8(b) and (e)), and the PN's *TTD* (Figure 5-8(c) and (f)). Note that q_{in} and q_{g} (Figure 2-2) are identical in the present application, as gating is applied directly at the PN boundary. Concentrating for now on the left column of Figure 5-8 (parts (a), (b), (c)), the first hour is characterized by a gradual increase of all three displayed quantities, as it is typical with increasing demand in under-saturated traffic conditions. At time $t = 1$ h, the observed stronger increase of q_{in} leads to an according surge of *TTS* and *TTD*, the latter

reaching soon capacity values according to Figure 5-6(a) (for the 100% measurement case), while the former is traversing the aforementioned critical region of [600, 900].

However, as the inflow q_{in} continues to be high, *TTS* continues to increase to very high values (i.e. the PN becomes increasingly crowded with vehicles); as a consequence, link over-spilling and gridlock phenomena appear in the PN and lead to a sensible *TTD* reduction to lower levels, that are persisting until about $t = 2.8$ h, when the network returns to the critical and eventually the under-saturated ranges, thanks to the decreasing demand. It should be noted that, during the congested period $t \in [1.5\text{ h}, 2.8\text{ h}]$ (when *TTS* values are higher than 900 veh), the inflow q_{in} is also reduced at some PN entrance links due to over-spilling links of PN, i.e. as a result of the congestion that extends beyond the PN; but this "natural" gating is too late, too little to counter the occurred PN degradation caused by its overcrowding with vehicles.

5.3.3. Gating Scenario

In this section, we present the results of feedback gating applied to the same Chania network scenario as the mere fixed signal control above. It is recalled that a similar investigation conducted in section 5.2 demonstrated the usefulness of feedback gating when using 100% of PN link measurements; while the focus of the present investigation is on demonstrating the gating control action and comparing its impact on the network performance indexes, when using decreasing percentages of real-time measurements, i.e. on the basis of accordingly reduced implementation requirements and cost.

As pointed out in section 4.7, the usual peak-hour fixed-time signal plan of the Chania urban road network is modified at the gating junctions when real-time gating actions are present, by modifying accordingly the involved signal stage durations (see section 4.7.3 for more details). Note that the new re-staging procedure introduced in section 4.7.3, for separating the gated right-of-way from the opposite-direction right-of-way in the same stage is implemented in this study.

The various reduced-measurement gating cases call for accordingly reduced set-points $\hat{TTS}$ to feed the regulator (4-25). In fact, the set-points for the investigated cases of 100%, 35%, 10% and 5% measurements may be taken from the corresponding NFDs displayed on Figure 5-

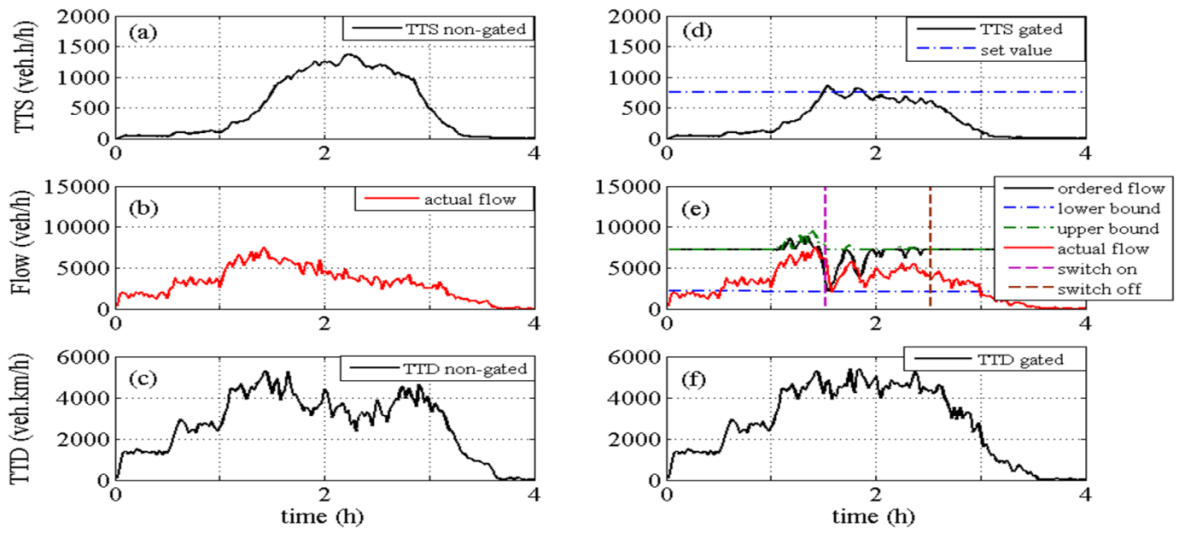

Figure 5-8 (a) PN's TTS vs. time in non-gating case; (b) actual PN inflow vs. time for the non-gating case; (c) PN's TTD vs. time for the non-gating case; (d) PN's TTS vs. time for the gating case; (e) ordered and actual PN inflow vs. time for the gating case; (f) PN's TTD vs. time for the gated case (100% measurements)

6(a)-(h). Specifically, the utilized respective $\hat{TTS}$ values are 750, 350, 170 and 105 veh·h per h.

The regulator parameters K_P and K_I were given the same values as in previous section for the 100% measurement case, i.e. $K_P = 20\,\text{h}^{-1}$ and $K_I = 5\ \text{h}^{-1}$. For the reduced measurement cases, these values were increased proportionally to the corresponding decrease of the set-point $\hat{TTS}$,to counterbalance the lower values of the *TTS*-measurements while calculating the inflows $q_{in}(k)$ with the regulator (4-25). Thus, to obtain the regulator gain values in the case of 35% measurements, given the approximate drop by half in the set-point (i.e. $\hat{TTS}(100\%)/\hat{TTS}(35\%) \approx 2$), the K_P and K_I values of the 100% case were multiplied by 2 (leading to $K_P = 40\,\text{h}^{-1}$ and $K_I = 10\,\text{h}^{-1}$). In the same way, for the case of 10% measurements, the values $K_P = 70\,\text{h}^{-1}$ and $K_I = 17\,\text{h}^{-1}$ were obtained. For the case of 5% measurements, the same policy was pursued at a first stage; however, when the resulting gain values were actually applied, the obtained closed-loop behavior was found to exhibit relatively strong time-variations of the resulting input signal $q_{in}(k)$. To explain this behaviour, one must consider that the time delay between the gating action and its impact at the measurement locations becomes increasingly non-negligible as the measurement percentage decreases. This means that the dynamic structure of the (q_g , *TTS*) SISO process under control changes

accordingly and increasingly in the reduced measurement cases. The regulator structure (4-25) is valid also in presence of time delays, but its gain values should be generally smaller in presence of a time-delay for damped closed-loop behaviour. Thus, the regulator gain values were eventually manually fine-tuned in the case of 5% measurements ($K_P = 60$ and $K_I = 15$) to obtain a smoother control action.

The regulator's maximum and minimum bounds are the same for all gating cases and are visible in Figure 5-8(e) (as well as in Figure 5-9(b), (g) and (l)). The regulator runs continuously in the background, but gating is actually activated only when *TTS* exceeds a threshold, that is selected slightly lower than (in this case 85% of) the set point; and is de-activated when *TTS* falls below a 2nd, slightly lower threshold and the relative difference between the ordered and actual flow is more than 30%. At all other times, fixed-time signal control is applied, as in the non-gating case.

Figure 5-8 (second column) displays the gating results for the case of 100% measurements, which can be directly contrasted to the (same-replication) non-gating results displayed in the first column of the same figure. It may be seen that the traffic conditions are identical as in the non-gating case up to around $t = 1.5\,\text{h}$, when gating is switched on (Figure 5-8(e)), as *TTS* approaches its set-point; the gating regulator orders low inflow values to maintain *TTS* around its set point, and, as a consequence, *TTD* is maintained at high levels (see Figure 5-8(f)), in clear contrast to the non-gating case (see Figure 5-8(c)). It is visible in Figure 5-8(e) that the actual inflow, shown by the red line, deviates at times from the q_{in} values ordered by the regulator (shown by the black line). By closer inspection of the results (not shown here), this deviation is mainly due to two specific gated links, whose demands are not always sufficient to create the respective inflows assigned to them. Remarkably, this deviation has only a minor impact on the regulator's ability to maintain *TTS* close to its set-point, thanks to the feedback character and the chosen structure of the regulator eq. (4-25). At $t = 2.5\,\text{h}$, *TTS* is returning to lower values, and the relative difference between the actual and ordered inflows grows higher, than the respective thresholds, hence gating is switched off, and the traffic flow returns to under-saturated conditions due to lower demand; this is in contrast to the non-gating case, where over-saturated traffic conditions are seen in Figure 5-8(a) to persist for almost half an hour longer.

To further illustrate the beneficial impact of the gating strategy on the traffic flow throughput in the PN, the NFD for the case of 100% measurements is displayed in Figure 5-10 for the

control case. The change, compared to its counterpart in Figure 5-6 is obvious. Under the action of the feedback gating regulator, the PN *TTS* is operated around its set-point; as a consequence the observed throughput degradation of the non-gating case is avoided and the throughput of the protected network is kept at the desired level (maximum *TTD*).

For the cases with reduced measurements, the gating case results (again for the same specific replication R2 for each case) are displayed in Figure 5-9 specifically, each of the three columns of Figure 5-9 (i.e. Figure 5-9((a)-(e)), ((f)-(j)) and ((k)-(o), respectively) displays the detailed results of replication R2 for the respective gating cases with 35%, 10% and 5% of measurements, to illustrate and compare their way of functioning and impact.

The first row of Figure 5-9 (i.e. Figure 5-9(a), (f) and (k)) displays the *TTS(x%)* trajectories, for x=35, 10, 5, produced (via eq. (2-1)) from the available measurements within PN; we recall that this is indeed the quantity feeding the regulator eq. (4-25) for each respective reduced-measurement case. Clearly, the *TTS* values are accordingly smaller when fewer measurements are summed up in eq. (2-1). In all three cases, the regulator is seen to keep the *TTS* values close to the respective set-point (blue dot-dashed line) during the gating period.

The second row of Figure 5-9 (i.e. Figure 5-9(b), (g) and (l)) displays the ordered (black curve) and actual (red curve) inflow to PN for each respective reduced-measurement case. The basic structure of these diagrams is similar across the three cases as well as with Figure 5-8(e), i.e. the 100% measurement case. Specifically, there is a gating activation period (between the dashed vertical lines), which may be slightly shorter or longer, depending on the specific development of the traffic conditions reflected in the real-time measurements of each case; within the gating period, the regulator (4-25) determines the displayed gated inflow appropriately, so as to maintain *TTS* close to its set-point. As in the discussed 100% measurement case, the deviation of ordered from actual inflows is mainly due to lack of demand at two gated links, but this fact has hardly any effect on the gating regulator's efficiency in maintaining *TTS* near its set-point.

The third row of Figure 5-9 (i.e. Figure 5-9(c), (h) and (m)) displays the *TTD(x%)* trajectories, for *x*=35, 10, 5, produced (via eq. (2-2)) from the available measurements within PN; as with *TTS*, the magnitude of the *TTD* values depends obviously on the number of measurements summed up in eq. (2-2). Thus, in contrast to the third row of Figure 5-8, it is not possible here to see directly whether the PN's *TTD* is actually maintained at capacity levels, because the measurements capture only a portion of the total *TTD*. To enable the complete assessment, it is easy, in the present simulation environment, to calculate and display, in the fourth and fifth rows of Figure 5-9, the full (100% measurement) *TTS* and *TTD* trajectories, respectively, which result under the reduced-measurement gating of each figure

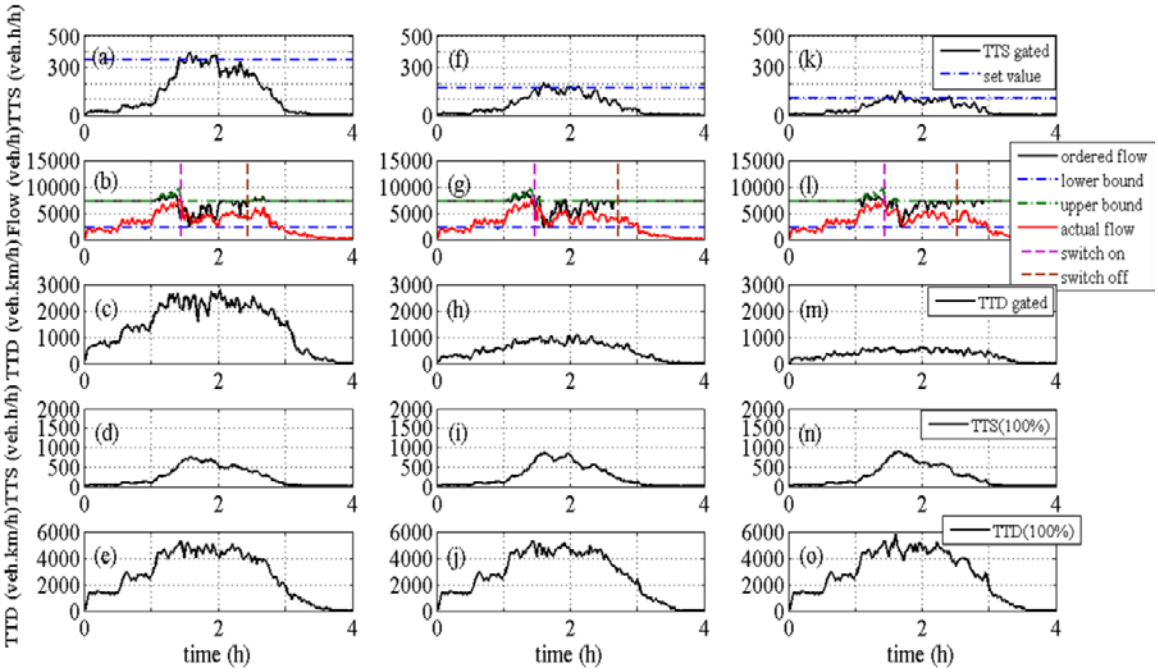

Figure 5-9(a) PN's TTS(35%) vs. time for the gating case (35% measurements); (b) ordered and actual PN inflow vs. time for the gating case (35% measurements); (c) PN's TTD(35%) vs. time for the gated case (35% measurements); (d) PN's corresponding TTS(100%) vs. time for the gating case (35%measurements); (e)corresponding PN's TTD(100%) vs. time for the gated case (35% measurements); (f) PN's TTS(10%) vs. time for the gating case (10% measurements); (g) ordered and actual PN inflow vs. time for the gating case (10% measurements); (h) PN's TTD(10%) vs. time for the gated case (10% measurements); (i) corresponding PN's TTS(100%) vs. time for the gating case (10% measurements); (j) corresponding PN's TTD(100%) vs. time for the gated case (10% measurements); (k) PN's TTS(5%) vs. time for the gating case (5% measurements); (l) ordered and actual PN inflow vs. time for the gating case (5% measurements); (m) PN's TTD(5%) vs. time for the gated case (5% measurements); (n) corresponding PN's TTS(100%) vs. time for the gating case (5%measurements);(o) corresponding PN's TTD(100%) vs. time for the gated case (5% measurements)

column. By inspection of the corresponding diagrams, we can state that:

- The *TTS*(100%) trajectories lie within the critical value range of [600, 900] during the gating activation period, for all reduced-measurement gating cases. This confirms (now for a prolonged period of time thanks to gating) the observation of Section 5.3.1, where it was found that the critical ranges for all *TTS*(*x*%), *x*=100, 35, 10, 5, are attained roughly at the same time.

- The *TTD*(100%) trajectories reach capacity values, as in the 100% measurement case of Figure 5-8(f), during the gating activation period, for all reduced-measurement gating cases. This confirms that the concept of feedback gating can be applied successfully to maximize the PN throughput based on far less real-time measurements and cost than needed to capture the complete traffic state of the PN.

We now turn our attention to the detailed performance index values obtained in the 10 replications of each (full or reduced-measurement) gating case, as displayed on Table 5-3(b). The last four columns of Table 5-3 report the average, maximum, minimum and standard deviation (S.D.) values, respectively, for each index and for each gating scenario. Furthermore, Table 5-4 displays the achieved gating improvements of the average performance indexes over the non-gating case of Table 5-3(a), for each gating scenario.

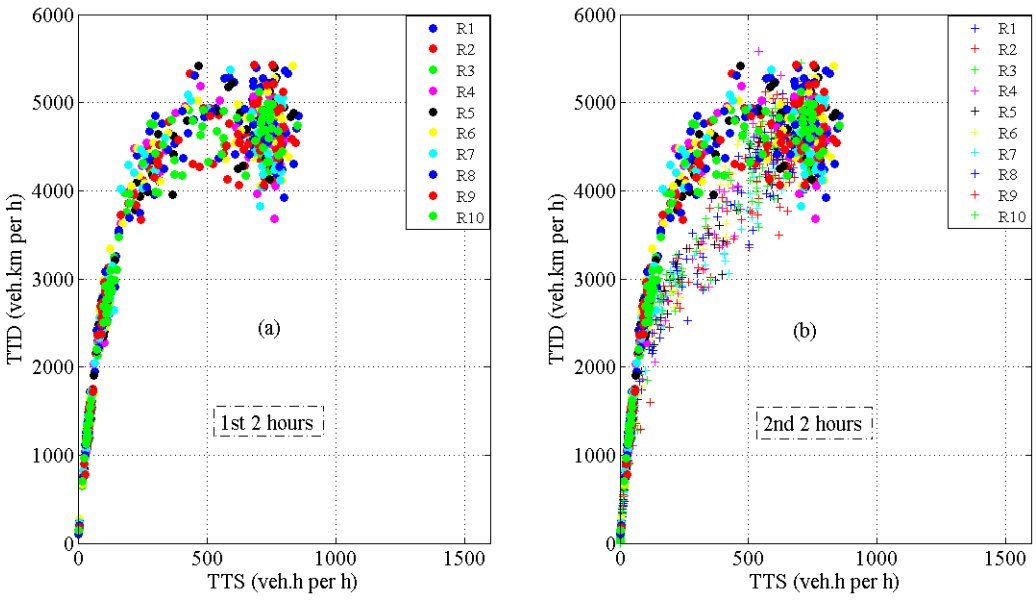

Figure 5-10 The operational NFD(100%) after using the gating strategy; (a) for the first 2 hours and (b) the 4 hours simulation

Table 5-4 Improvements of gating cases versus the non-gating case for different measurement percentages

	Changes (%)			
%Measurement	Ave. Delay	Ave. Speed	Delay S. D.	Speed S. D.
100%	-29.5	29.2	-87.9	-78.3
35%	-32.2	31.4	-83.3	-70.5
10%	-28.9	27.1	-85.8	-76.0
5%	-29.1	29.2	-84.3	-74.4

To start with, the average number of served vehicles over the whole scenario duration is virtually identical across all (gating or non-gating) cases. This was expected, since the underlying demand scenario is the same; the initial traffic state is identical (empty network); and the final traffic state is identical (virtually empty network); for all investigated cases.

The average delay and mean speed show clear and comparable improvements in all gating cases compared to the non-gating case, which are due to the avoidance of PN degradation. Specifically, the average delay improvements range from 28.9% (in the 10% measurement case) to 32.2% (in the 35% measurement case); while the average speed improvements range from 27.1% (in the 10% measurement case) to 31.4% (in the 35% measurement case). The slight improvement variations across different gating cases are deemed to reflect, except for stochastic effects, the impact of the specific locations of the chosen measurements that are utilized for gating. This issue is further discussed in Section 5.3.4.

A particularly interesting finding in the reported results concerns the S.D. index, which, as mentioned earlier, may reflect to some extent the reliability of the traffic conditions. Table 5-4 reports huge reductions of S.D. in all gating cases, in the order of 85% for the delays and of 75% for the speeds, compared to the non-gating case. In fact, if a network degradation is allowed to occur (as in the non-gating case), the underlying overspill and partly gridlock conditions seem to give rise to a huge variety of quite different potential evolution paths of the traffic state, that may be triggered by small initial deviations, i.e. a chaotic-like dynamic behaviour. On the other hand, the avoidance of overspill, gridlock and the resulting network degradation in the gating cases, renders the traffic state evolution and performance more stable and predictable, despite the existence of stochastic variations across different replications. Although a quantification of the related reliability improvements is not possible on the basis of the present simulation-based investigations, we expect that similar

phenomena may be encountered in the daily variation of traffic conditions in real conditions as well.

5.3.4. Discussion and Practical Implications

Study 2 shows that the feedback gating procedure and algorithm proposed in section 5.2, which was based on the provision of full real-time information from all PN links, may be applied equally efficiently also with much less real-time measurements pertaining to a small subset of PN links. Obviously, this finding has significant implications for the implementation requirements and cost of the method in practice. Specifically, feedback gating may be applied with a small percentage (up to 5% in this work) of real-time measurements that are used to feed the regulator (4-25). The set-point $\hat{TTS}$ included in (4-25) may be readily obtained on the basis of the same small percentage of detector-equipped links, by tracing the corresponding reduced NFD before the gating application, as e.g. in Figure 5-6(g) for the 5% measurement case.

At this point, it is important to stress the importance of a proper selection of measurement links, particularly in cases of very few measurement links. The successful results reported in Section 5.3.3, even for 5% of measured links, may not materialize for all possible measurement link selections. As mentioned earlier, reduced measurement links were selected to reflect early increasing queuing in the PN, something that seems easy to also pursue in practice, based on known recurrent daily traffic patterns in the PN. Nevertheless, it is interesting to investigate cases of other measurement link selections, and, in particular, to demonstrate that proper set-point specification for feedback gating may be virtually impossible in cases of really inappropriate link selection. Figure 5-11 displays the operational NFDs obtained (without gating) by use of 5% measurements (8 links), albeit with different link selections than in Figure 5-6. Specifically, Figure 5-11(a) and (b) (for network filling and emptying, respectively) are based on 4 of the previously considered 5% links (which feature early queuing) plus 4 PN links which are “late congested”, i.e. feature significant queuing only when the whole PN gets congested; while Figure 5-11(c) and (d) reflect the NFD for 8 exclusively late-congested links. Remarkably, the NFD scatter in the latter case is so strong, that no reasonable *TTS* set-point can be determined; note that the strong scatter remains even if we calculate *TTD* averages for each *TTS* region of 1 vehh per h (not shown here). Thus, feedback gating based on this “NFD” does not seem feasible. On the other hand, a

sufficiently well-formed NFD results from the mixed selection of (early-plus-late) measurement links according to Figure 5-11(a) and (b), with a maximum-*TTDTTS*-range of [50-90] veh·h per h. Note that the time period of reaching this *TTS*-range in the non-gating case largely overlaps, but does not completely coincide, with the time-period of reaching the optimal *TTS*-range in the 100% measurement case.

Figure 5-12 displays the gating results obtained (for Replication R2) for the case of mixed-selection 5% measurements; while Table 5-5 displays the related detailed and averaged performance index values; for a set-point of 70 veh·h per h and regulator gain values of $K_P = 70\ h^{-1}$ and $K_I = 20\ h^{-1}$. The obtained results are seen to be roughly equivalent to the

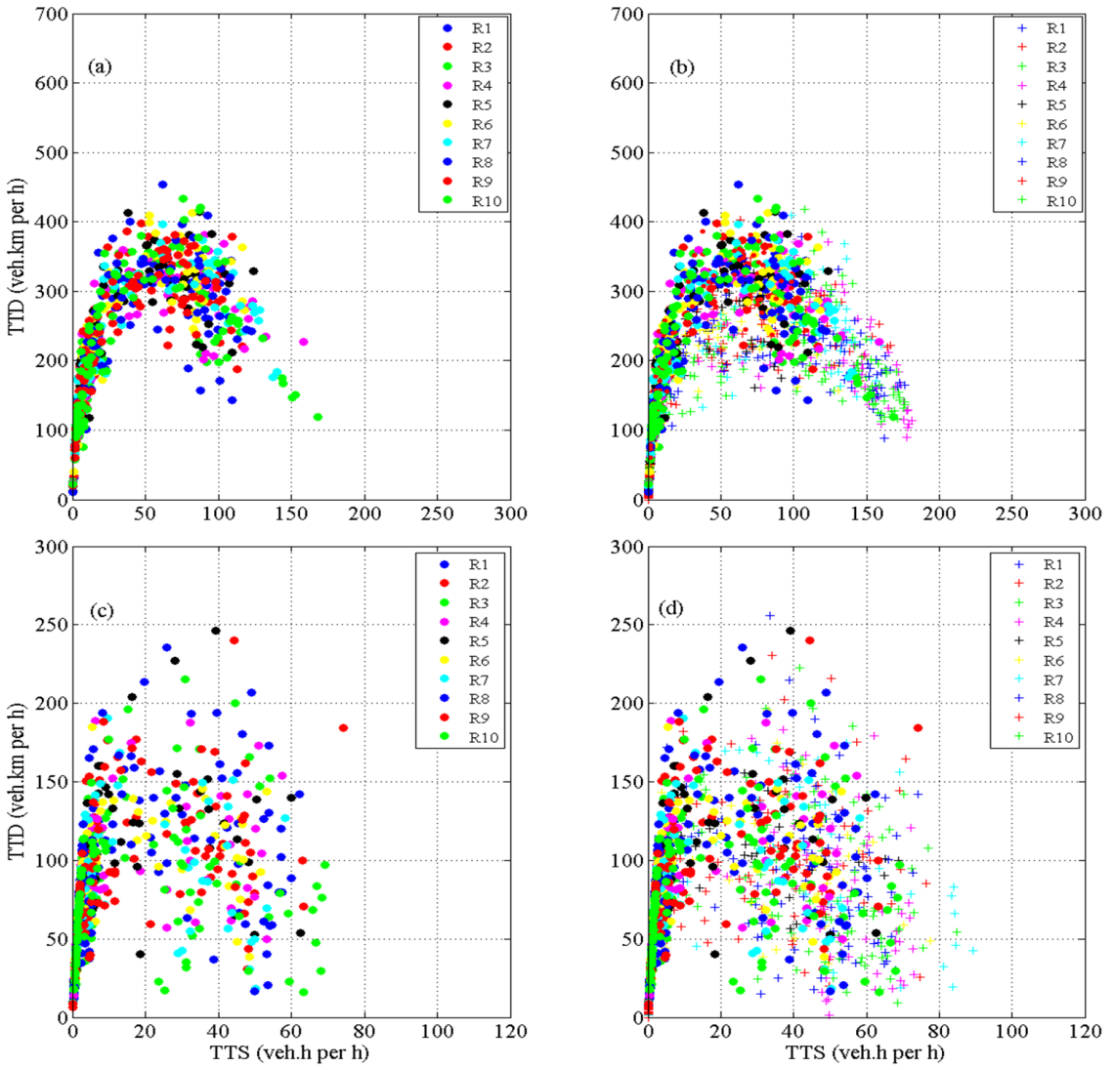

Figure 5-11(a) NFD of the PN for mixed 5% measurements for the first 2 hours for 10 replications; (b) NFD of PN for mixed 5% measurements for the 4 hours simulation for 10 replications; (c) NFD of the PN for late-congested 5% measurements for the first 2 hours for 10 replications; (d) NFD of PN for late-congested 5% measurements for the 4 hours simulation for 10 replications

results of section 5.2, which indicates that even a non-perfect mix of few measurement links may suffice for efficient feedback gating.

The above results indicate that the proper selection of, even very few, PN measurement links for feedback gating is not really difficult, at least in the here utilized simulated environment with real-time route assignment. Whether this also applies to real networks, should be specifically investigated with real data.

It should be noted, however, that the outlined field deployment procedure would feature a difference compared to the present simulation-based investigation, where we had the possibility to also obtain the full NFD of the PN (via emulated detector measurements in all links) and to check whether the critical *TTS* range of the reduced NFD is attained simultaneously with the critical *TTS* range of the full NFD. This test is not possible in a real deployment environment where only a small percentage of link measurements would be actually available. In other words, the readily obtained reduced NFD (as in Figure 5-6(g)) would include a critical *TTS* range where the reduced (measured) *TTD* is maximized; but there would be no direct way to check, that the corresponding traffic state actually maximizes the full PN throughput in practical deployments. Thus, although an efficient selection of link measurements proved easy in the present simulation-based environment, the question of how to best select the specific links to be equipped for feedback gating application in practice remains relevant and should motivate future research.

5.4. *Study 3*: Multiple Concentric Boundary Gating

Given that in large metropolitan urban networks the congestion spreads mostly heterogeneously over the network, in which case a homogenous gating strategy may not be the optimal solution, this section proposes a new gating strategy which implements the aforementioned feedback-based gating strategy, along with considering the heterogeneity of a large-scale urban network. This multiple concentric-boundary gating strategy, initially applies the gating concept at the border of the region where the first core of congestion starts;

Table 5-5 Performance indexes using the mixed 5% link selection (*: in 1000)

Index	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Ave.	Max	Min	S. D.
Delay (s/km)	233	231	236	235	226	251	258	221	229	266	238	266	221	14
Speed (km/h)	11.7	11.8	11.6	11.7	12.0	11.1	10.9	12.2	11.9	10.6	11.6	12.2	10.6	0.50
Veh. out*	15.9	15.8	15.6	15.9	15.7	16.0	15.8	15.9	15.7	15.9	15.8	16.0	15.6	-

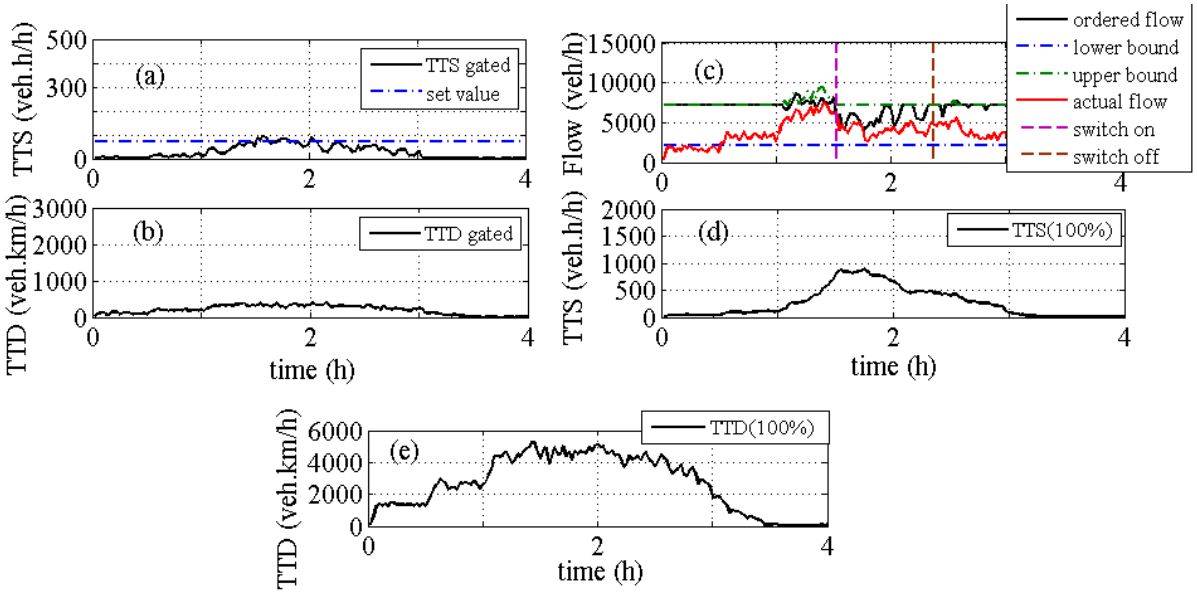

Figure 5-12 Feedback gating results for mixed 5% measurements: (a) PN's *TTS(mixed 5%)* vs. time; (b) ordered and actual PN inflow vs. time; (c) PN's *TTD(mixed5%)* vs. time; (d) PN's corresponding *TTS(100%)* vs. time; (e) corresponding PN's *TTD(100%)* vs. time

eventually, as congestion continues to expand, the border of an extended network part becomes the second perimeter for a second (independent) gating control.

As discussed in section 4.7.2, distributing the flow ordered by the controllers is a crucial issue in the gating procedure. During the activation of the controllers, in one or more gated junctions, the actual flow may differ substantially from the flow ordered by the regulators. This may occur because of low demand in the gated link or due to downstream spillover at the gated junctions. Thus, ignoring the real-time traffic state in gated links may result in unutilized portions of green phases and, consequently, in unnecessary delays imposed on the vehicles in the non-gated directions. Therefore, as mentioned earlier, in *study 3*, a new flow distribution policy is proposed, which considers the cases of low demand (by monitoring the queue length in the gated link in real-time) and spillback simultaneously. The gating strategy proposed in *study 3* has led to significant improvements versus the previously proposed single boundary gating control and the no-gating scenarios. The proposed gating strategy implements *scenario 3* as test-bed.

The results of 10 different replications for each investigated scenario are presented in Table 5-6 and then each evaluation criterion is compared across different scenarios. Since all the simulation scenarios (*scenario 1* to *scenario 4*) are in the environment of AIMSUN 6,

which does not consider the waiting time in the virtual queue in the calculation of the delay (s/km) at the end of the simulation, an additional performance index is introduced in this study. This performance index may be interpreted as average *TTS*/*TTD* (s/km) or average unit travel time that considers the waiting time in the virtual queue in the calculation of the *TTS*. Beside this, to compare the new performance index with the reported index (delay) in previous scenarios, the average vehicle delay per km is also reported here. It must be emphasized, that average *TTS*/*TTD* is provided for the entire San Francisco urban network (not only the PN); thus, the reported improvements due to gating are net benefits, because the delays at the gated links are actually included in the performance index. The improvements of the proposed multiple boundary gating control are compared with single boundary (as in section 5.2) and the non-gating scenario.

5.4.1. Network Fundamental Diagram of Protected Networks (PNs)

The NFDs of the San Francisco PN_1 and PN_2 introduced in chapter 3 (assuming that all links are detector-equipped, i.e. $\mathbb{M} = \mathbb{Z}$), is obtained via a 5-hour simulation with DTA-based routing and is displayed in Figure 5-13. Ten different replications (each with a different random seed) of the 5-hour scenario are included (with different colors) in Figure 5-13. Figure 5-13(a) and (b) display the (60s) measurement points for $PN_2$and PN_1 respectively. In this study, the demand profile is trapezoidal, starting from low values and increasing gradually to levels that lead to heavy congestion in PN (as under typical real traffic conditions at the peak periods); ultimately, the demand is gradually reduced, until the

Table 5-6 Performance indices of non-gated (NG) and control strategies (SC, TC) and corresponding improvements

	TTS/TTD (s/km)			Improvement (%)		Delay (s/km)			Improvement (%)	
	NG	SC	TC	SC	TC	NG	SC	TC	SC	TC
R 1	1179	1151	1134	2.43	3.87	509	426	416	16.31	18.27
R2	1257	1171	1138	6.86	9.46	513	426	397	16.96	22.61
R 3	1213	1153	1103	4.92	9.08	506	433	401	14.43	20.75
R 4	1470	1265	1140	13.95	22.45	607	455	411	25.04	32.29
R 5	1320	1255	1135	4.88	13.97	541	459	394	15.16	27.17
R 6	1550	1266	1255	18.32	19.06	628	464	446	26.11	28.98
R 7	1414	1230	1219	13.00	13.76	554	417	413	24.73	25.45
R 8	1391	1201	1157	13.68	16.86	574	419	419	27.00	27.00
R 9	1497	1345	1207	10.12	19.36	618	479	414	22.49	33.01
R 10	1398	1231	1174	11.95	16.00	581	446	425	23.24	26.85

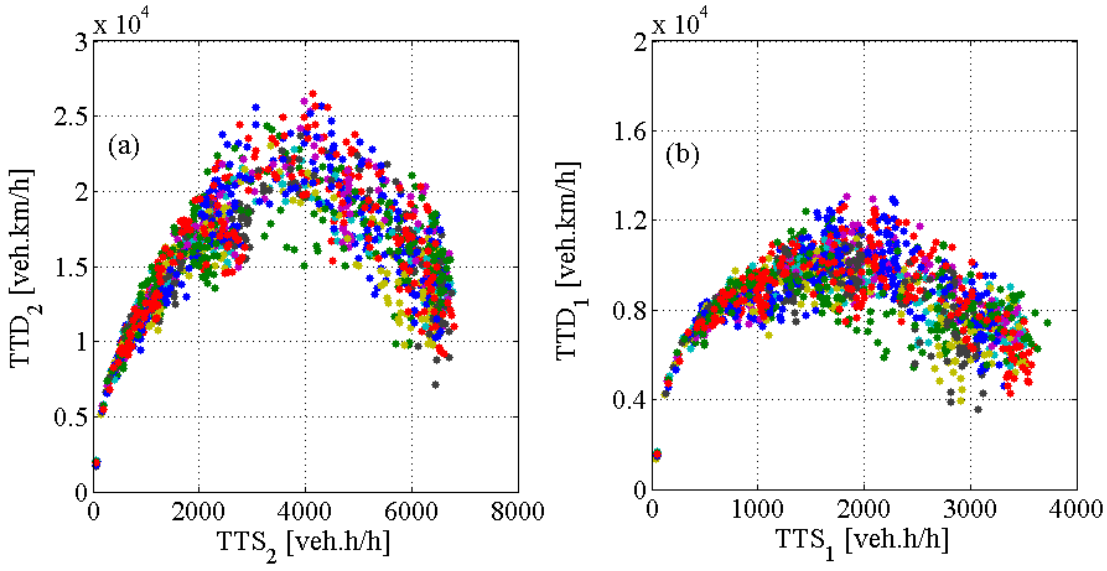

Figure 5-13(a) NFD of PN_2; (b) NFD of PN_1 for 10 replications

network is virtually emptied at the end of the simulation (see Figure 5-13). The most relevant issue for gating control, which can be clearly seen in Figure 5-13, is that the maximum *TTD* values in the diagram for PN_2 and PN_1 consistently occur in a limited *TTS* range of 3500 to 4000 and 1700 to 2000 vehh per h, respectively. If the *TTS* values exceed these ranges, the *TTD* decreases drastically and can lead the PN throughput to lower levels or even to a gridlock state. A thorough observation of the NFDs of PN_1 and PN_2 revealed that PN_2 reaches the maximum *TTD* region in its NFD approximately with a time gap of 20 minutes (i.e. 20 cycles) later than PN_1. This confirms the need of a multiple-boundary gating strategy for a large-scale urban network like the one addressed in this section to avoid gating delays for some vehicles without any reason. Thus, the new proposed gating strategy attempts to maintain the traffic state of PN_1 in the maximum throughput condition, and, if not completely successful in mitigating the overall congestion during the peak period, the second gating boundary at PN_2 turns on. In other words, the two-stage gating concept tries to maintain the traffic conditions in both PN_1 and PN_2 in the critical range (i.e. *TTS* with maximum *TTD*) of their NFDs by employing two separate feedback regulators at the borders of PN_1 and PN_2 respectively (see section 4.6) .

5.4.2. Control Scenarios

In this sub-section, the simulation results of three control scenarios are presented; non-gating (NG) scenario or fixed signal control, single-controller (SC) scenario or single-boundary

Table 5-7 Performance indices average improvement of control strategies

	TTS/*TTD* (s/km)			Improvement (%)	
	NG	SC	TC	SC	TC
Avg.	1369	1227	1166	10.01	14.39
COV	0.09	0.05	0.04	-	-
	Delay (s/km)			Improvement (%)	
	NG	SC	TC	SC	TC
Avg.	563	442	414	21.15	26.24
COV	0.08	0.05	0.04	-	-

gating, and two-controller (TC) strategy or multiple boundary gating scenario. In the NG scenario, a real fixed-time signal plan applied in the San Francisco network during the peak period is utilized. This scenario is used as base-scenario to compute the benefits from the proposed gating strategy. The controller implemented in SC strategy monitors TTS_2 measurements in PN_2 and regulates the inflows through 15 intersections located at the boundary of PN_2. TC case reflects a new gating strategy; two controllers that may work separately or in conjunction. The first gating controller employs TTS_1 measurements (in real-time) from PN_1, and regulates traffic signal timings according to the ordered flow values at 9 intersections located on the bold red line in Figure 3-5. The second gating controller uses TTS_2 measurements from PN_2, and regulates the inflow at 11 intersections located on the pink line ($PN_2 \backslash PN_1$) in Figure 3-5.

K_P and K_I are derived by manual fine-tuning, after calculating them from parameter estimation procedure introduced in section 4.6.1.1. Thus, in SC scenario, the regulator parameters are $K_P = 20\ h^{-1}$ and $K_I = 5\ h^{-1}$ and in TC strategy, regulator parameters are calculated to be $K_{P1} = 12\ h^{-1}$ and $K_{I1} = 3 h^{-1}$ for PN_1, $K_{P2} = 16\ h^{-1}$ and $K_{I2} = 4\ h^{-1}$ for $PN_2 \backslash PN_1$. Set-points are selected to be $\hat{TTS}_1 = 1750$ veh·h/h for PN_1, and $\hat{TTS}_2 = 3500$ veh·h/h for PN_2. In this study, the regulators are activated when TTS_1 and TTS_2 exceed a certain threshold, 85% of $\hat{TTS}_1$ and 85% of $\hat{TTS}_2$, respectively, and it is turned off when they are below a 2nd slightly lower threshold (see section 4.7.1).

Table 5-6 introduces the results obtained with three different control strategies for the 10 replications (*TTS/TTD*, delay and individual improvements). In all the replications, SC strategy produces better results than NG, and TC scenario better than SC regarding both *TTS*/*TTD* and average delay. According to Table 5-7, the average improvements of SC over

NG case are 10% and 21% for *TTS*/*TTD* and average delay, respectively. These improvements rise to 14% and 26%, respectively for *TTS*/*TTD* and average delay, with the implementation of the new multiple-boundary gating strategy. Nevertheless, the benefit from gating considering adverse effects of gated vehicles, especially with TC strategy, is quite significant. In fact, it implies that even gated vehicles may have benefits from gating strategy, as their waiting times can be compensated by higher speeds inside the network enabled by the gating strategy (slower is faster effect). Note that incorporation of only 5 intersections (1^{st}, 2^{nd}, 3^{rd}, 8^{th} and 9^{th} intersections on the red bold line in Figure 3-5) into the gating strategy has substantially increased the improvements produced by SC strategy.

Apparently, the NG strategy is not able to maintain a reliable network performance during peak hours, as indicated by coefficient of variance (COV) values. The reliability of the network is positively affected by SC strategy, and it is further improved by TC gating, see COV values in Table 5-7.

Figure 5-14 displays *TTS* evolution of R4 in PN_1 and PN_2 in a comparative way, which indicates that congestion does not evolve in the network in a homogenous manner. Especially at the loading stage, PN_1 and PN_2 are not homogeneously congested; note that the curves are not close to the proportionality line. This clearly demonstrates that PN_1 and PN_2 have different congestion levels for a certain period during the simulation. Therefore, a single gating activity at the boundary of the entire network can be improved by a finer strategy which includes two controllers that work in conjunction and operate all regions approximately at their optimum *TTS*. On the other hand, Figure 5-14 clarifies the fact that TC gating is able to render the network homogenous and maintain different parts at the capacity flow. Note that the dots in Figure 5-14 concentrate on the intersection of two set points in case of TC strategy.

Figure 5-15(a) and (b) display *TTS* evolution over time for three control scenarios in PN_2 and in PN_1, respectively. Note that both TTS_2 and TTS_1 are identical until $t \cong 1.1$ h for all strategies, where gating action starts in TC scenario. Apparently, SC is able to keep PN_2 at the desired level $\hat{TTS}_2$. However, the same statement does not hold for PN_1; TTS_1 values do not stay around the set point (see Figure 5-15(b)). Figure 5-15(a) indicates that TTS_2 measurements with TC strategy show a very similar behavior to the ones with SC strategy. However, PN_2 with TC gating switches from capacity flow to uncongested regime earlier than SC scenario, which indicates that the new strategy is able to dissipate congestion faster.

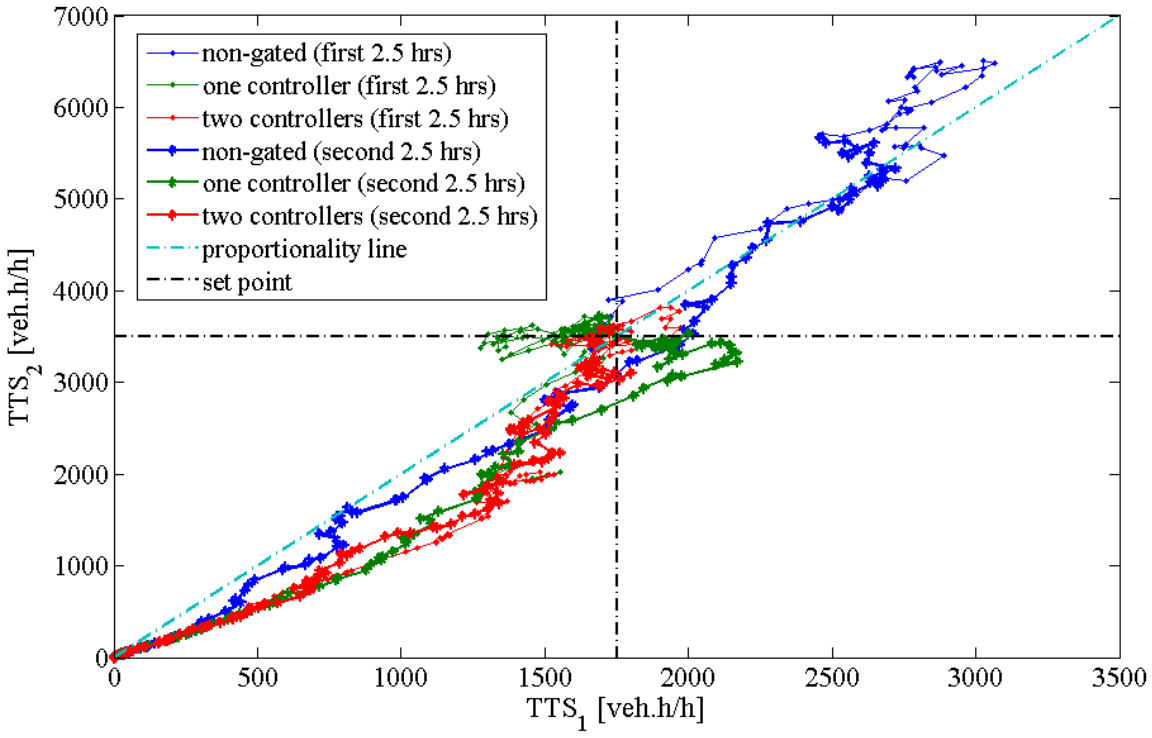

Figure 5-14 *TTS* Evolution in PN_1 and PN_2

Figure 5-16(a) and (b) introduce controller action results for $PN_2\backslash PN_1$ and PN_1 in R4, respectively. Regulators' activation and deactivation times, along with lower and upper bounds are also presented. Note that, for the periods where the controllers are off, the flows ordered by the fixed signal timing plan are displayed in Figure 5-16. The inflow introduced in Figure 5-16(a) represents the number of vehicles crossing the pink border (through 11 gated junctions) in Figure 3-5 to enter the network ($PN_2\backslash PN_1$), and the one presented in

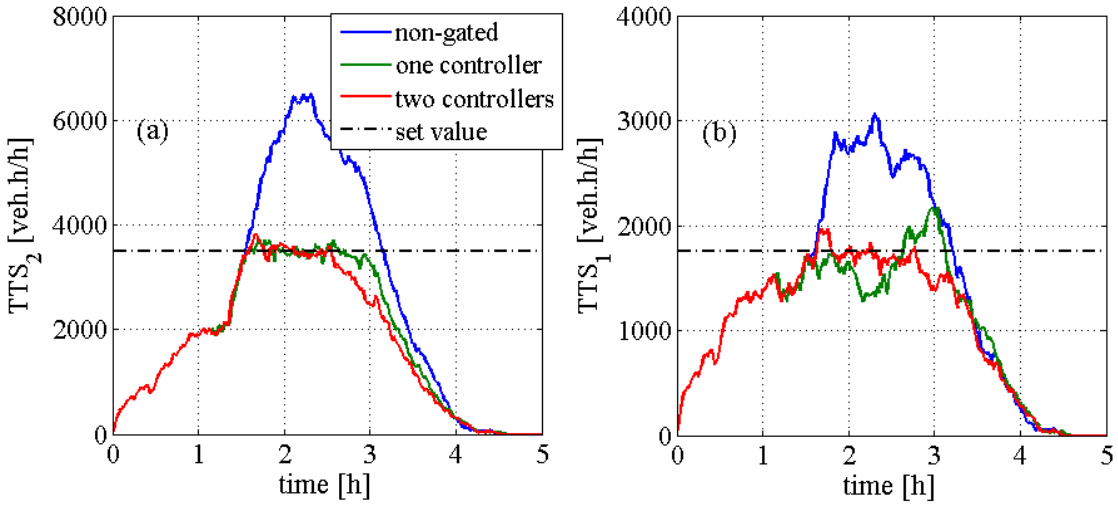

Figure 5-15 (a) TTS2 vs. Time; (b) TTS1 vs. time

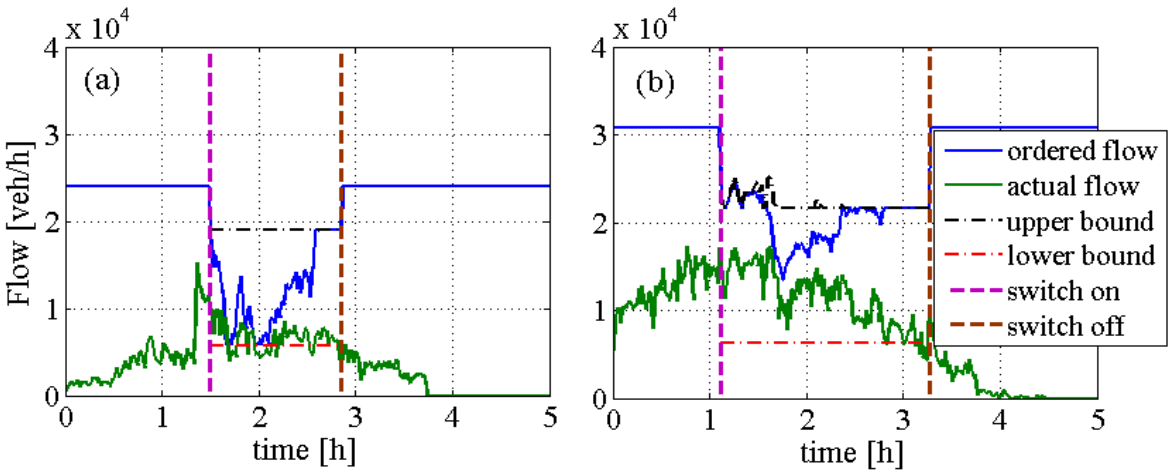

Figure 5-16(a) Ordered and actual inflow in PN2\PN1 two controllers; (b) ordered and actual inflow in PN1 two controllers

Figure 5-16(b) denotes the number of vehicles that cross the red bold line (through 9 gated junctions) to enter PN_1. Gating controller for PN_1 switches on first at $t \cong 1.1$ h as TTS_1 measurements exceed the corresponding threshold. Controller regulates traffic signal timings at the boundary of PN_1, to keep TTS_1 close to $T\hat{T}S_1$. At $t \cong 1.5$ h, TTS_2 measurements exceed the defined threshold, and gating at $PN_2\backslash PN_1$ is also switched on. For the period, $t \in [1.5, 2.9]$ h, two controllers work in conjunction to reach desired levels of $T\hat{T}S_1$ and $T\hat{T}S_2$. At $t \cong 2.9$ h, TTS_2 falls below the defined threshold, and the gating controller switches off for $PN_2\backslash PN_1$. For $t \in [2.9, 3.2]$ h, the gating controller at PN_1 works individually to keep PN_1 at the capacity flow. Then, the system switches back to fixed signal timing configuration at $t \cong 3.2$ h, as traffic levels in both parts are uncongested. Note in Figure 5-16(a) and (b) that ordered flows differ from actual flows. In case of idle gating, where the controller is not active, the big gap between ordered flows (from fixed signal timing configuration) and the actual flows is mainly due to low demand. On the other hand, in case of gating, the gap occurs mainly due to spillbacks and gridlocks at the gating junctions.

5.4.3. Spillback and Low-Demand Actions

Flow distribution policy implemented in this study prevents wasted green time at gating junctions where demand is not sufficient to reach ordered flow, and it readjusts signal timings at intersections under spillback influence, leading to a smoother traffic flow. Figure 5-17(a) and (b) display ordered and actual flows along with estimated queue size for the 7th

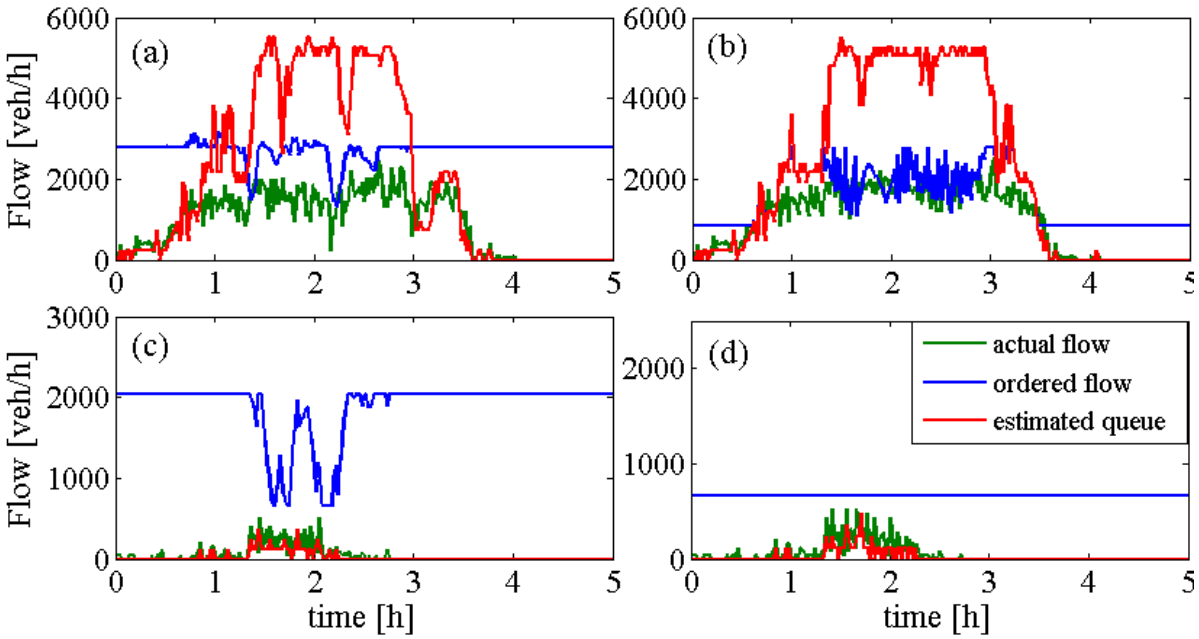

Figure 5-17(a) Ordered and Actual Inflow for 7th Intersection on the Red Line without Spillback Action; (b)– with Spillback Action; (c) Ordered and Actual Inflow for 8th Intersection on the Pink Line – without Low Demand Action; (d)– with Low Demand Action

intersection on the red line (see Figure 3-5) without and with spillback action, respectively. The gap between actual and ordered flows in Figure 5-17(a) for the period $t \in [1.2, 2.1]$ h clearly indicates that spillbacks occur at the gated junction. Although there is enough demand at the gated junction (i.e. estimated queue size is bigger than ordered flow), inflow assigned by the original demand distribution algorithm cannot be satisfied. On the other hand, the proposed algorithm with the spillback action is able to adjust ordered flow values and to reduce the size of the aforementioned gap. Note that spillback action presented in this study cannot guarantee that actual and ordered flows are identical. However, empirical observations indicate that this algorithm is able to significantly decrease the gap, although not fully removing it (see Figure 5-17(b)).

Figure 5-17(c) and (d) present estimated queue size, ordered and actual flows for the 8th intersection on the pink line (see Figure 3-5) without and with low-demand action, respectively. In order not to waste green time, the current distribution algorithm replaces maximum green time in the original algorithm with the necessary signal timing for the estimated queue size at the gated junction. In case it is less than the specified minimum green time, lower and upper bounds (i.e. minimum and maximum green times) are equated to each other (see Figure 5-17(d)). Even though the demand at the gated junction is very low, the

uniform demand distribution algorithm assigns high inflow values (see Figure 5-17(c)). On the other hand, the proposed approach monitors the conditions on the link in real-time, and tunes the inflow values by bringing the phase duration down to its lower bound. The new flow distribution policy leads to significant reduction of wasted green times at low-demand junctions, and uses the excess green times to serve the queued vehicles at other intersection approaches. Note that actual flow at this intersection is not significantly affected by the new approach (see Figure 5-17(c) and (d)), as the links connected to this junction do not suffer from spillback issues, and the gated demand can freely enter the network even within minimum green duration. Note that ordered flows presented in Figure 5-17 are applied only when the controller is on; otherwise, fixed signal timings are employed.

5.5. *Study 4*: Feedback-Based Gating Remote from PN

In all the previous studies (i.e. *study 1to study 3*), gating was applied directly at the border of the protected network (PN), i.e. the network part to be protected from over-saturation. In other words, to implement gating, the usual traffic light settings have been modified at (one or more) junctions at the boundary of the PN. In *study 4*, the developed feedback-based gating concept is partly applied at junctions located further upstream of the PN. This induces a time-delay, which corresponds to the travel time needed for gated vehicles to approach the PN. The resulting extended feedback control problem can be tackled by use of a PI (Proportional-Integral) regulator as well, albeit with different gain values compared to the case without time-delay. The reported results show a stable behavior and improved mobility of the overall network in terms of mean speed and travel time. It should be noted that simulation *scenario 4* is applied for this study.

5.5.1. NFD of PN

Figure 5-18(a) and (b) display the complete operational NFD (90 s cycle-based measurement points) obtained for the PN of Figure 3-6 for the loading (first 2 hours) and the whole (i.e. 4-hour) AIMSUN simulation period (including the network recovery period, indicated by +), respectively. As pointed out also in Chapter 3, the DTA based-routing is activated during the simulation runs; ten different replications (each with different seed in AIMSUN)of the 4-hour scenario were carried out and are included (with different colors) in Figure 5-18. A trapezoidal traffic demand profile, starting from very low values and increasing gradually to

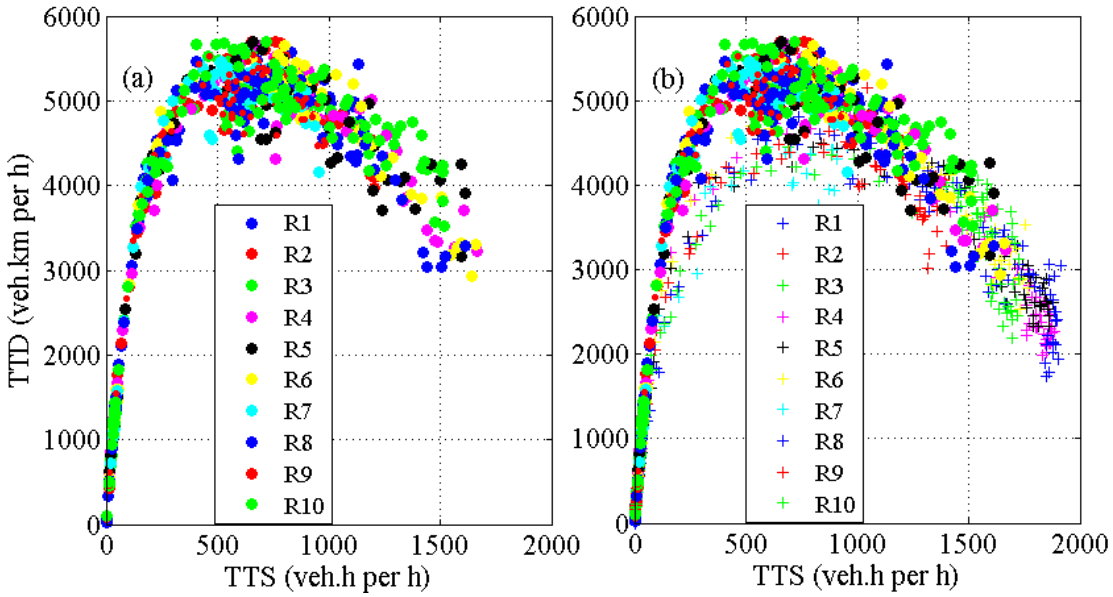

Figure 5-18 (a) NFD for loading PN for 10 replications; (b) NFD for loading and unloading for 10 replications

levels that result in congestion within PN (as under typical real traffic conditions at the peak periods) is introduced to the network; ultimately, the demand is gradually reduced, until the network is virtually emptied at the end of the simulation (see Figure 5-18(b)). Figure 5-18(a) demonstrates that a fundamental diagram (asymmetric inverse-U shape) is indeed occurring during the 2-h network filling period, with low scatter even across different replications; Figure 5-18(b) indicates that the inverse-U shape appears also during the decreasing demand period of 2 h, albeit with a visible hysteresis compared to the filling 2-h period. The maximum *TTD* values in the diagram occur in a *TTS* region of 500 to 800 veh·h per h. These phenomena may be exploited for gating control by maintaining the *TTS* in the aforementioned range, so as to maximize the throughput in the PN. An uncontrolled *TTS* may lead to the decrease of *TTD* (and hence of the PN throughput), if the optimal range of *TTS* is exceeded.

5.5.2. Non-Gating Case

As pointed out before, the traffic signal plan in the non-gating case is set according to the fixed-time signal settings utilized in the real Chania network during the peak period. Table 5-8(a) displays the aforementioned indexes for every replication (R). The last two columns display the average and standard deviation (S. D.) for each index. The standard

Table 5-8 (a) performance indexes results for non-gated scenario; (b) results for the gated scenario (*: in 1000)

Scenario	Indexes	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Ave.	S. D
(a)	Delay (s/km)	230	260	285	421	405	352	187	413	214	369	314	84
Non- gated	Speed (km/h)	11.6	10.6	9.9	7.2	7.4	8.3	13.4	7.3	12.2	8.0	9.6	2.2
	Vehicles out*	12.8	12.9	12.6	12.8	12.8	13.0	12.8	12.8	12.7	12.8	12.8	-
(b)	Delay (s/km)	213	232	189	214	202	220	180	226	187	206	206.9	16.5
	Change (%)	-7.4	-10.8	-33.7	-49.2	-50.1	-37.5	-3.7	-45.3	-12.6	-44.2	-34.0	-80.4
Gated	Speed (km/h)	12.2	11.4	13.3	12.1	12.7	11.9	13.8	11.7	13.4	12.5	12.5	0.8
	Change (%)	5.2	7.5	34.3	68.1	71.6	43.4	3.0	60.3	9.8	56.3	30.3	-65.1
	Vehicles out	12.8	12.9	12.6	12.8	12.8	13.0	12.8	12.8	12.8	12.8	12.8	-

deviation is considered as an index of reliability of the traffic conditions over different runs of simulation. It is observed that the non-gating scenario leads to high S.D. for both delay and mean speed, i.e. 84 (s/km) and 2.2 (km/h), respectively.

The detailed results of one specific replication (R5) are displayed in Figure 5-19, for an illustrative comparison of the non-gating versus gating cases. Figure 5-19(a) and (d) display the PN's *TTS* and the total flow served by the gated junctions (from the 7 gated links) q_{g} and the PN's *TTD* are shown in (Figure 5-19(b) and (e)) and (Figure 5-19(c) and (f)), respectively. Note that q_{in} and q_{g} are different in the present application, as gating is applied partly further upstream (gating points 1, 2, 3 and 5) of the PN perimeter. Focusing on the left column of Figure 5-19 (i.e. (a), (b), (c)), with increasing demand, all three displayed quantities are gradually increasing, as typical in under-saturated conditions. At time $t = 0.6$ h, the increased demand leads to the observed surge of *TTS* and *TTD*, the latter reaching soon capacity values according to Figure 5-18(a), while the former is traversing the aforementioned critical region of [500, 800]. As the demand keeps increasing, *TTS* continues to increase and reaches very high values (2000 veh per h); as a consequence, link over-spilling and gridlock phenomena appear in the PN and lead to a sudden drop in *TTD* values (see Figure 5-19(c)), that are persisting until about $t = 2.6$ h; eventually, in the period $t \in [2.6\mathrm{h}, 3.3\mathrm{h}]$, *TTD* increases again thanks to the reduced congestion due to the decreasing demand at the end-stage of simulation.

5.5.3. Gating Scenario

In this section, the simulation results while applying the gating control strategy to the same Chania network as for the non-gating case are presented. The set-point for the regulator may be taken from the NFD displayed on Figure 5-18. Specifically, the utilized $\hat{TTS}$ value is 750 veh·h per h. The utilized regulator parameters K_P and K_I were obtained after a least square estimation (see section 4.6.1.1) by applying the time-series of (q_g, *TTS*) data around the *TTS*-critical (750 veh.h/h) and considering $m = 3$. To specify m, the longest path from the remote gating positions (i.e. junctions 1, 2, 3 and 5 in Figure 3-6) to the PN is found. The distance from junction 5 to the PN is about 900m which, by considering an average speed of 12km/h (including waiting times at three signalized junction in between and the congestion occurring in this path during the peak period) for the vehicles traveling to the PN, may reach a value of 270s as the travel time of a vehicle to PN. Consequently, the gating action may face a delay of 270s or three signal cycles to reach the PN. This time-delay term is considered by m in the least-squares procedure in section 4.6.1.1. Thus, in the first step μ and ζ are defined; $\mu = 0.769$ and $\zeta = 0.012$ and by applying Table 4-1 for $m = 3$; we derive $K_P = 10$ $h^{-1}$and $K_I = 3$ h^{-1}. In cases where the gating positions are located at different distances from the PN (as in this study), the aforementioned methodology can be implemented to derive the controller parameters; but this might be slightly conservative, because, the longest travel time to the PN is considered as the time-delay of the system. Moreover, the farthest gating junction may not be the most influential among the other gating positions, since this depends highly on the demand profile and the flow which is served in that gated junction. Thus, one may use the method discussed in section 4.6.1.1 by identifying the ersatz model (i.e. $\Delta TTS(k+1) = \mu \cdot \Delta TTS(k) + \zeta \cdot \Delta q_{in}(k-m)$), in place of the complicated real model, with a certain level of error and carry out the estimation procedure for different values of *m*; for $m = 1$: $\mu = 0.812$, $\zeta = 0.023$ and $\Pi = 29429$; for $m = 2$: $\mu = 0.781$, $\zeta = 0.027$ and $\Pi = 33704$; for $m = 3$: $\Pi = 23674$. Consequently, the minimum value for Π is derived from the parameter estimation with $m = 3$.

Figure 5-19 (second column) displays the results for the gated case (same replication as for non-gating), which can be directly compared to the non-gating results displayed in the first column of the same figure. Up to around $t = 1.1$h, the traffic conditions are identical as in the non-gating case. When gating is activated (shown by the dashed violet vertical line in

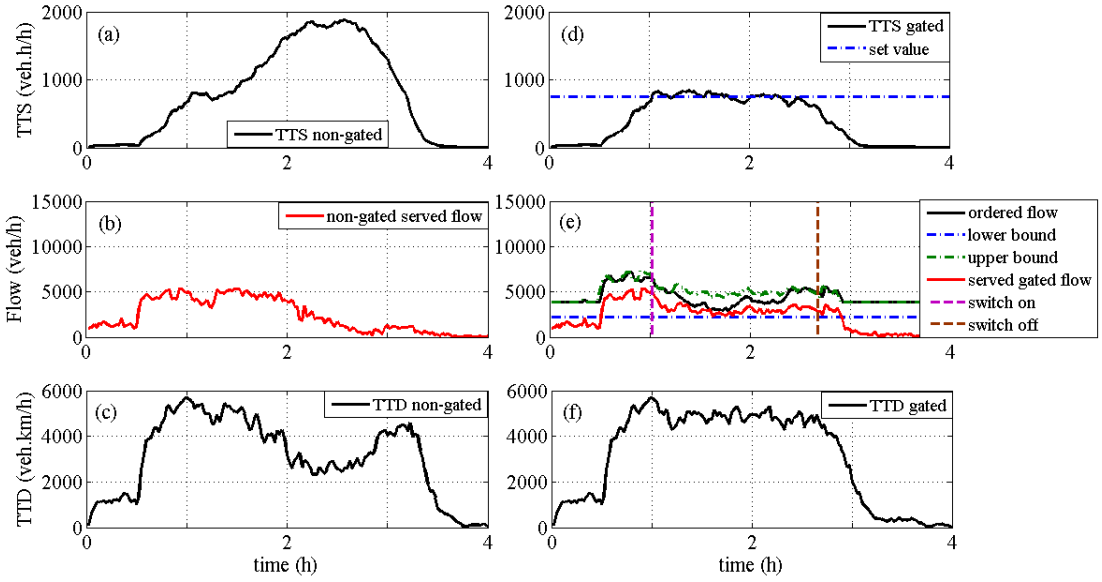

Figure 5-19(a) PN's TTS vs. time in non-gating case; (b) total served flow by the gated junctions vs. time for the non-gating case; (c) PN's TTD vs. time for the non-gating case; (d) PN's TTS vs. time for the gating case; (e) served and ordered flow vs. time for the gating case; (f) PN's TTD vs. time for the gated case

Figure 5-19(e)), as *TTS* approaches its set-point ($T\hat{T}S = 750$), the gating regulator orders lower gating flow values to maintain *TTS* around its set point, and, as a consequence, *TTD* is maintained at high levels (see Figure 5-19(f)), avoiding the discussed drop of *TTD* in time period $t \in [1.8\text{h}, 2.6\text{h}]$ in the non-gating scenario (see Figure 5-19(c)). It is visible in Figure 5-19(e) that the served gated flow, shown by the red line, deviates at times from the values ordered by the regulator (shown by the black line). A thorough inspection of the total served gated flow (not shown here) revealed that this gap is mainly due to some specific gated links, whose demands are not always sufficient to create the respective inflows assigned to them. As it was expected, thanks to the chosen feedback structure of the controller, this deviation has virtually no influence on the regulator's performance for keeping PN's *TTS* close to the pre-specified set-point. As we reach the end period of the simulation, *TTS* is returning to lower values, and the relative difference between the actual and ordered gated flows grows higher, hence gating is switched off, and the traffic flow returns to under-saturated conditions due to lower demand. Interestingly, the congestion period in the gated scenario is reduced by almost 40 minutes (see Figure 5-19(a) and (d)) compared to the non-gating case.

Table 5-8(b) displays the achieved gating improvements of the average performance indexes over the non-gating case of Table 5-8(a). As expected, because of the identical traffic demand in both scenarios and the fact that the network is evacuated at the end of the simulation, the total number of vehicles served in both cases is close to each other. There are remarkable improvements in the overall mean speed and the average delay in order of 34% and 30%, respectively. Table 5-8 reports also huge reductions of S. D., i.e. 80% for the delay and of 65% for the speed, compared to the non-gating case.

5.6. *Study 5*: Increasing Control Step

Further investigation on the control action in presence of time-delay has been carried out by compensating τ with application of a bigger control step for the regulator eq. (4-25). For the same scenario discussed in section 5.5 (*scenario 4*), the gating control action is implemented for a time step of five cycles (i.e. 450s). In other words, during the gating action is updated with new measurements every five cycles. By considering this time step, each time the regulator orders a flow; this flow is implemented for five consecutive cycles, before it is updated again. Thus, in this case, we may assume that the control system is operating with virtually zero time-delay. By using time-series of (q_g , *TTS*) data around the *TTS*-critical (750 veh here) and considering *m*=0 for the estimation procedure, we end up in $\mu = 0.760$ and $\zeta = 0.011$. By applying Table 4-1, for *m*=0, we have $K_P = \mu/\zeta = 65$ and $K_I = 1 - \mu/\zeta = 20$.

Table 5-9 displays the detailed simulation results for the non-gated and gated scenarios of this study. Interestingly, the results indicate similar efficiency compared to the simulation results in *study 4.* It should be emphasized that this is an important finding, especially for the case of real-filed implementation. This study proves that for an efficient gating control, the control action can be executed even in a bigger time step. According to Table 5-9, significant improvements in the overall mean speed and the average delay in order of 34% and 30%, respectively, have been achieved.

The results of the non-gated (first column) and gated scenario (second column) are demonstrated in Figure 5-20 for replication 10. The first column was discussed in details in the previous section. The second column displays the results for the gated case (same replication as for non-gating), by considering the aforementioned parameters in the regulator's equation. As expected, the regulator attempts to maintain the *TTS* of the PN (see

Figure 5-20(d)), close to the pre-specified set-value (i.e. 750 veh.). After activation of gating control, (shown by the dashed violet vertical line in Figure 5-20(e)), as *TTS* approaches its set-point, the gating regulator orders lower gating flow values (2200 veh./h) to keep *TTS* around its set point, and, as a consequence, *TTD* is maintained at high levels (see Figure 5-20(f)). Since the control step is five cycles in this case, the activation of the gating strategy (i.e. exceeding 85% of $\hat{TTS} = 750$ after 2 cycles, defined in the algorithm), starts slightly later than in Figure 5-19. However, the controller acts perfectly and keeps the *TTS* pretty close to the critical value. After applying the gating control, the overall delay and mean speed are reduced to 212 s/km and 12.3 km/h, respectively.

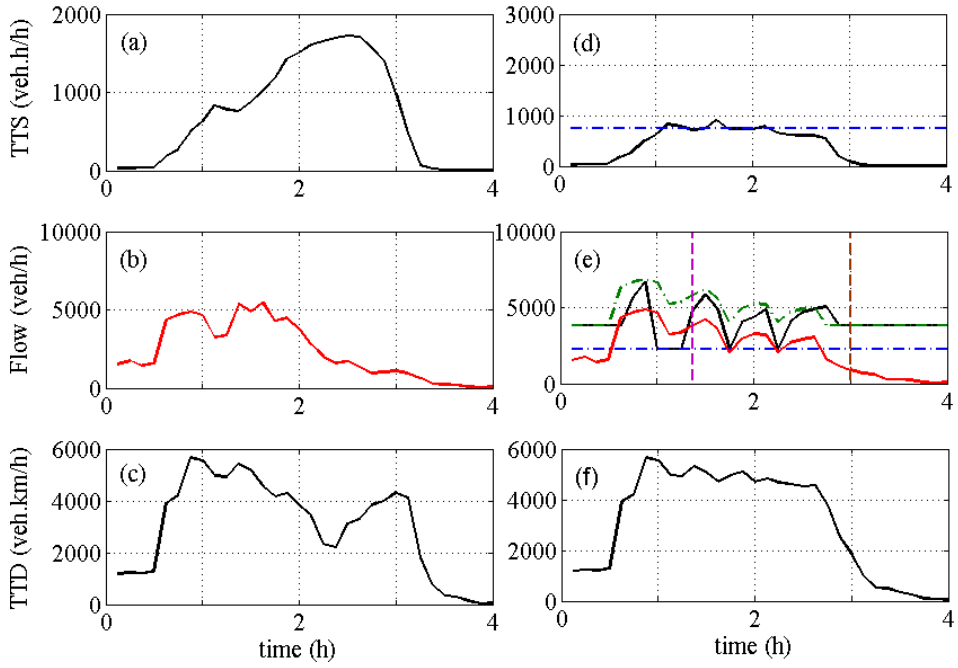

Figure 5-20 Results of gating with bigger control steps: (a) PN's *TTS* vs. time in non-gating case; (b) total served flow by the gated junctions vs. time for the non-gating case; (c) PN's *TTD* vs. time for the non-gating case; (d) PN's TTS vs. time for the gating case; (e) served and ordered flow vs. time for the gating case; (f) PN's TTD vs. time for the gated case

Table 5-9 Results for the gated scenario (*: in 1000) for bigger control steps

Scenario	Indexes	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Ave.	S. D
Gated	Delay (s/km)	184	189	192	181	185	197	187	219	188	212	193.4	11.4
	Change (%)	-20.0	-27.3	-32.6	-57.0	-54.3	-44.0	0.0	-47.0	-12.1	-42.5	-38.3	-85.9
	Speed (km/h)	13.6	13.2	13.2	13.7	13.5	12.9	13.4	12	13.4	12.3	13.1	0.5
	Change (%)	17.2	24.5	33.3	90.3	82.4	55.4	0.0	64.4	9.8	53.8	36.8	-75.2
	Vehicles out*	12.8	12.9	12.6	12.8	12.8	13.0	12.8	12.8	12.8	12.8	12.81	-

Chapter 6

The important thing is never to stop questioning.

Albert Einstein

6. Conclusion and Future Work

In the final chapter, the findings and results of this thesis are summarized, along with comments on further research on the topic. Section 6.1 gives a summary of the mains findings. The main contributions of this thesis are summarized in section 6.2. Finally, the aspects which should be considered for further research are presented in section 6.3.

6.1. Concluding Remarks

Gating aims at protecting urban road networks from over-saturation, or, more specifically, at maximizing the network throughput. Based on the previously developed concept of a network fundamental diagram (NFD), an operational urban NFD has been defined to enable simple, practicable and efficient gating control, potentially even by use of a very limited amount of real-time measurements. A simple (nonlinear and linearized) control design model, incorporating the operational NFD, has been developed, which allows for the gating problem to be cast in a proper feedback control design setting. This allows for application and comparison of a variety of linear or nonlinear, feedback or predictive (e.g. Smith predictor, internal model control and other) control design methods from the Control Engineering arsenal; among them, a simple but efficient

PI feedback regulator was developed and successfully tested in a fairly realistic microscopic simulation environment. More specifically, two different urban road networks (i.e. Chania

and San Francisco) were modeled in the microscopic simulator AIMSUN as test-beds for this thesis, to protect their most sensitive part from spillovers, gridlock, and the resulting strong degradation. Application of the developed gating strategies is demonstrated to lead to significant improvements compared to the non-gating control scenarios in the investigated studies (*study 1* to *5*).

The protected networks in these investigations were selected in an ad hoc way, based on related experience with the real traffic conditions. Further investigations and, hopefully, field implementations, with different network types and sizes, as well as different demand and congestion patterns and different gating locations may shed more light on the most beneficial practices to be applied.

It must be stressed that gating can only be successful in reducing the overall network delay if a couple of conditions are actually met in the network to be protected from oversaturation. Firstly, a congestion-caused degradation, i.e. a reduction of throughput (e.g. due to link spillover and gridlock), must actually occur without gating, else there would be no potential for improvement. Secondly, the occurring degradation must be (at least partly) reversible if the number of vehicles in the network is maintained at a certain optimal level; in other words, the targeted high efficiency and throughput must be sustainable, not merely transient phenomena.

In *study 1*, the measurements of all links within the protected network are fed to the regulator; study 2 revealed that this is not necessary, and that gating may be applied similarly efficiently with far less real-time measurements. However, this may open the question on the sort of measurements that are most beneficial for gating and hence contradict to some extent the usage of the NFD, which, by definition, reflects the traffic conditions on a whole network, not only on selected parts thereof.

In *study 2*, it is confirmed that reduced NFDs, obtained with far less measurements than in a complete NFD, exhibit a critical range of traffic states that is virtually equivalent to their counterparts of the complete NFD. Subsequently, the usage of less real-time measurements for feeding the PI regulator was investigated thoroughly and revealed virtually equivalent gating results compared to the case of 100% link measurement availability; and significant improvements (in the order of 30% in mean speed and delay) compared to the non-gating control case when using any of the aforementioned percentages of measurements. These improvements stem from a sensible increase of the network throughput enabled by the

feedback gating action. Interestingly, the gating action was also found to drastically reduce the standard deviation of the performance indexes across different replications, compared to the non-gating case, which is likely to lead to improved reliability of the traffic conditions in practice.

In case of big heterogeneous urban networks, different regions may exceed critical range of accumulation at different times. Therefore, a single-controller based gating strategy may cause unnecessary delays for the vehicles that are bound to uncongested parts of the network. The approach presented in *study 3* includes two feedback regulators that are associated with different regions in the network and that work either in separate or in coordination. The urban road network of San Francisco is modeled as a study area for the new gating strategy. Application of the new gating strategy based on two controllers has resulted in significant improvements over the single-perimeter gating strategy and non-gating cases. Moreover, a new flow distribution policy has been introduced, by monitoring the queuing at the gated junctions, to avoid under-utilized green phases caused by downstream spill-back and/or lack of demand.

In *study 4*, it is demonstrated that efficient feedback-based perimeter control may be applied not only at the boundary, but also further upstream of the PN. Based on restrictions which might be faced for implementing gating control at the perimeter of the PN in different urban networks, such as unavailability of proper links to store the gated vehicles (queuing) or sufficient number of signalized junctions, the traffic may be also metered at some junctions remote from the border of the PN. Significant improvements in terms of mean speed and average delay have been achieved by applying the gating strategy.

Last but not least, *study 5* proved that the real-time gating action can be executed at a slower pace. In other words, this finding shows that even with a bigger control step (not one signal cycle as in *study 1* to *4*), the gating strategy acts efficiently and leads to significant improvements in mean speed and delay at the overall network level. This may be very interesting for the urban traffic management authorities; because it shows that the signal plan can be modified less frequently during the gating action.

6.2. Main contributions

The main contributions of this thesis can be summarized as following:

- A control-design model exploiting the NFD and an appropriate feedback-based gating regulator was developed for the first time in the frame of the present thesis research.
- A thorough study on NFDs derived with subsets of links in the network has been carried out.
- It is proved that an efficient feedback-based gating is possible with much less real-time measurements.
- A multiple concentric boundary gating strategy is introduced which implements the aforementioned feedback-based gating strategy, along with considering the heterogeneity of a large-scale urban network.
- Different methods for the flow splitting at the gated junction are proposed and applied in the microscopic simulation environment.
- A robust feedback controller, by considering time-delay on the system (due to gating action not at the border of protected network), is designed.
- It is shown that the feedback gating works properly with much longer time-steps.

6.3. Further Research

Further investigations and, hopefully, field implementations, with different network types and sizes, as well as different demand and congestion patterns and different gating locations may shed more light on the most beneficial practices to be applied, e.g. for the selection of appropriate measurement locations. Further research directions include comparison with more comprehensive traffic-responsive signal control strategies, introducing more efficient queue balancing and management policies at the gating positions and possibly designing a hybrid control strategy which combines the gating concept with a traffic responsive control strategy downstream of the gated links. This may lead to increased junction capacity of the downstream links and might be beneficial in reducing spillback.

Bibliography

[1] Aboudolas, K., Zheng, N., and Geroliminis, N., 2014. Perimeter flow control for bi-modal urban road networks. 93th TRB Annual Meeting, Washington, DC, USA.

[2] Aboudolas, K., and Geroliminis, N., 2013. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. Transportation Research Part B 55, 265-281.

[3] Aboudolas, K., Papageorgiou, M., Kouvelas, A., and Kosmatopoulos, E., 2010. A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks. Transportation Research Part C 18 (5), 680-694.

[4] Abu-Lebdeh, G., and Benekohal, R.F., 1997. Development of traffic control and queue management procedures for oversaturated arterials. Transportation Research Record 1603, 119-127.

[5] Barceló, J., and Casas J., 2004. Heuristic DynamicAssignment Based on AIMSUN Microscopic Traffic Simulator. Proceedings of TRISTAN V.

[6] Beard, C., and Ziliaskopoulos, A., 2006. A system optimal signal optimization formulation. 85th TRB Annual Meeting, Washington, DC, USA.

[7] Bretherton, D., Bowen, G., and Wood, K., 2003. Effective urban traffic management and control: Recent developments in SCOOT. 82nd TRB Annual Meeting, Washington, DC, USA.

[8] Buisson, C., and Ladier, C., 2009. Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams,Transportation Research Record 2124, 127–136.

[9] Chang, T.H., and Sun, G.Y., 2004. Modeling and optimization of an oversaturated signalized network. Transportation Research Part B 38, 687-707.

[10] Chien, K. L., Hrones, J. A., Reswick, J. B., 1952. On the Automatic Control of Generalized Passive Systems. Transactions of the American Society of Mechanical Engineers 74, 175-185.

[11] Daganzo C. F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches.Transportation Research Part B 41 (1), 49-62.

[12] Daganzo, C. F., and Geroliminis, N., 2008. An analytical approximation for macroscopic fundamental diagram of urban traffic. Transportation Research Part B 42 (9), 771-781.

[13] Daganzo, C. F., Gayah, V. and Gonzales, E., 2011. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability,Transportatio Research Part B 45, 278–288.

[14] De Schutter, B., and De Moor, B., 1998. Optimal traffic light control for a single intersection. European Journal of Control 4 (3), 260-276.

[15] Diakaki, C., Papageorgiou, M., and Aboudolas, K., 2002. A multivariable regulator approach to traffic-responsive network-wide signal control. Control Engineering Practice 10 (2), 183-195.

[16] Dinopoulou, V., Diakaki, C, and Papageorgiou, M., 2005. Application and evaluation of the signal traffic control strategy TUC in Chania. Journal of Intelligent Transportation Systems 9 (3), 133–143.

[17] Fadali, S., 2009. Digital Control Engineering Analysis and Design. USA: Elsevier.

[18] Farges, J.L., Henrry, J.J., and Tufal, J., 1983. The PRODYN real-time traffic algorithm. 4th IFAC Symposium on Transportation Systems. Baden-Baden, Germany, 307-312.

[19] Farhi, N., 2008. Modélisation minplus et commande du trafic de villes régulières (Ph.D. Thesis). Université de Paris I-Panthéon-Sorbonne, Paris, France.

[20] Franklin, G. F., Powell, J. D., and Emami-Naeini, A., 1994. Feedback Control of Dynamic Systems, Third Edition,Addison-Wesley.

[21] Godfrey, J. W., 1969. The mechanism of a road network.Traffic Engineering Control 11, 323–327.

[22] Gal-Tzur, A., Mahalel, D., and Prashker, J. N., 1993. Signal design for congested networks based on metering. Transportation Research Record 1398, 111-118.

[23] Gartner, N.H., 1983. OPAC: a demand-responsive strategy for traffic signal control. Transportation Research Record 906, 75-84.

[24] Gartner, N.H., and Wagner, P., 2004. Analysis of traffic flow characteristics on signalized arterials. Transportation Research Record 1883, 94-100.

[25] Gayah, V., and Daganzo,C.F., 2011. Clockwise hysteresis loops in the Macroscopic Fundamental Diagram:An effect of network instability. Transportation Research Part B 45, 643-655.

[26] Geroliminis, N., and Daganzo, C.F., 2008. Existance of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B 42 (9), 759-770.

[27] Geroliminis, N., and Sun, J., 2011. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transportation Research Part B 45 (3), 605-617.

[28] Geroliminis, N., Haddad, J., and Ramezani, M., 2012. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach. IEEE Trans. on Intelligent Transportation Systems 14 (1), 348-359.

[29] Helbing, D., 2009. Derivation of a fundamental diagram for urban traffic flow. The European Physical Journal B 70 (2), 229-241.

[30] Hunt, P.B., Robertson, D.I, Bretherton, R.D., and Royle, M.C., 1982. The SCOOT on-line traffic signal optimization technique. Traffic Engineering and Control 23 (4), 190-192.

[31] Ji, Y., Daamen, W., Hoogendoorn, S., Hoogendoorn-Laser, S., and Qian, X., 2010. Investigating the shape of the macroscopic fundamental diagram using simulation data. Transportation Research Record 2161, 40-48.

[32] Ji, Y., and Geroliminis, N., 2012. On the spatial partitioning of urban transportation networks. Transportation Research Part B 46 (10), 1639–1656.

[33] Knoop, V.L., Van Lint, J.W.C., and Hoogendoorn, S.P., 2012. The macroscopic fundamental diagram used for control using subnetwork accumulation, Transportation Research Records 2315.

[34] Kouvelas, A., Aboudolas, K., Kosmatopoulos, E.B., and Papageorgiou, M., 2011. Adaptive performance optimization for large-scale traffic control systems. IEEE Transactions on Intelligent Transportation Systems 12 (4), 1434–1445.

[35] Lia, Y., Xua, J., and Shena, L., 2012. A Perimeter Control Strategy for Oversaturated Network Preventing Queue Spillback. Procedia-Social and Behavioral Sciences 43, 418-427.

[36] Lieberman, E., Chang, J., Bertoli, B., and Xin, W., 2010. New signal control optimization policy for oversaturated arterial systems. 89th TRB Annual Meeting, Washington, DC, USA.

[37] Lo, H.K., 1999. A novel traffic signal control formulation. Transportation Research Part A 33 (6), 433-448.

[38] Lo, H.K., Chang, E., and Chan, Y.C., 2001. Dynamic network traffic control. Transportation Research Part A 35 (8), 721–744.

[39] Lowrie, P.R., 1982. SCATS: the sydney co-ordinated adaptive traffic system-priciples, methodology, algorithms. IEEE International Conference on Road Traffic Signalling, London, England, pp. 67-70.

[40] Luk, J., and Green, D., 2010. Balancing traffic density in a signalized network. Austroads Research Report AP-R369/10, Sydney, Australia.

[41] Mahmassani, H., and Peeta, S., 1993. Network performance under system optimal and user equilibrium dynamic assignments: Implications for ATIS Transportation Research Record 1408, 83-93.

[42] Mahmassani, H., J. Williams, and R. Herman, 1987. Performance of urban traffic networks in Proceedings of the 10th International Symposium on Transportation and Traffic Theory. Amsterdam, The Netherlands.

[43] Mazloumian, A., Geroliminis, N. and Helbing, D., 2010. The spatial variability of vehicle densities as determinant of urban network capacity. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368(1928), 4627-4647.

[44] Mirchandani, P., and Head, L., 1998. RHODES-a real-time traffic signal control system: architecture, algorithm. TRISTAN III (Triennial Symposium on Transportation Analysis), San Juan, Puerto Rico, vol. 2.

[45] Ortigosa, J., Arnet, K., Menendez, M., 2012. Studying the feasibility of a MFD control in Zürich. 1st European Symposium on Quantitative Methods in Transportation Systems, Lausanne, Switzerland.

[46] Papageorgiou, M., Vigos, G., 2008. Relating time-occupancy measurements to space-occupancy and link vehicle-count. Transportation Research Part C 16 (1), 1-17.

[47] Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A., Wang, Y., 2003. Review of road traffic control strategies. Proceedings of the IEEE 91, 2043-2067.

[48] Papageorgiou, M., Haj-Salem, H. and Blosseville, J.M., 1991. ALINEA: A local feedback control law for on ramp metering. Transportation Reseach Record 1320, 58–64.

[49] Papageorgiou, M., and Messmer, A., 1985. Continious-time and discrete-time design of water flow and water level regulators. Automatica 21(6), 649-661

[50] Papamichail, I., Papageorgiou, M., 2011. Balancing of queues or waiting times on metered dual-branch on-ramps. IEEE Transactions on Intelligent Transportation Systems 12 (2), 438-452.

[51] Putha, R., Quadrifoglio, L. and Zechman, E., 2010. Using ant optimization for solving traffic signal coordination in oversaturated networks. 89th TRB Annual Meeting, Washington, DC, USA.

[52] Roess, R. P., Prassas, E. S., and Mcshane, W. R., 2004. Traffic Engineering. New Jersey: Pearson Education.

[53] Saberi, M. and Mahmassani, H., 2012. Exploring the properties of network-wide flow-density relations in freeway networks. Transportation Research Record 2315, 153–163.

[54] Seborg, D., Edgar, T.F., Mellichamp D.A., 1989. Process Dynamics and Control. New York: Wiley.

[55] Strating, M., 2010. Coordinated signal control for urban networks by using MFD (M.Sc. Thesis). Delft University of Technology, Delft, the Netherlands.

[56] TSS transport simulation systems, 2008. AIMSUN User Manual Version 6, Barcelona, Spain.

[57] Vigos, G., Papageorgiou, M., 2010. A simplified estimation scheme for the number of vehicles in signalized links. IEEE Trans. Intelligent Transportation Systems 11, 312-321.

[58] Wang, Y., Papageorgiou, M., Gaffney, J., Papamichail, I., Rose, G., Young, W., 2010. Local ramp metering in random-location bottlenecks downstream of a metered on-ramp. Transportation Research Record 2178, 90-100.

[59] Wang, Y. and Papageorgiou, M., 2006. Local ramp metering in the case of distant downstream measurements,Proc. 2006 IEEE Intelligent Transportation Systems Conference, Toronto, Canada.

[60] Wood, K., Bretherton, D., Maxwell, A., Smith, K., Bowen, G., 2002. Improved traffic management and bus priority with SCOOT (TRL Staff Paper PA 3860/02). London, UK: Transport Research Laboratory.

[61] Zhang, Y., Bai, Y. and Yang, X.G., 2010. Strategy of traffic gridlock control for urban road network. China Journal of Highway Transportation, 23, 96-102.

[62] Zheng, N., Aboudolas, K., and Geroliminis, N., 2013. Investigation of the existence of city-scale three-dimensional macroscopic fundamental diagrams for bi-modal traffic. In 16th IEEE International Conference on Intelligent Transportation Systems,TheHague,The Netherlands.

[63] Ziegler, J. G., and Nichols, N. B., 1942. Optimum settings for automatic controllers. Transactions ASME 64, 759-768.