Spatio-temporal Traffic Flow Estimation and Optimum Control in Sensor-Equipped Road Networks

Shangbo Wang

Electrical and Data Engineering

University of Technology Sydney

This dissertation is submitted for the degree of

Doctor of Philosophy

Faculty of Engineering and
Information Technology

04 December 2018

I would like to dedicate to my parents, Guanfeng Wang and Ping Yang, for their endless love and support.

Declaration

Shangbo Wang declares that this thesis, is submitted in fulfillment of the requirements for the award of PhD degree, in the Electrical and Data Engineering/Faculty of Engineering and Information Technology at the University of Technology Sydney. This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Shangbo Wang
04 December 2018

Acknowledgements

"*Es ist absolut möglich, dass jenseits der Wahrnehmung unserer Sinne ungeahnte Welten verborgen sind*" - Albert Einstein

Under the guidance of philosophical truth, this thesis is accomplished during my PhD study in Electrical and Data Engineering, Faculty of Engineering and Information Technology at University of Technology Sydney.

Firstly, I would like to express my deepest gratitude to my academic advisor, Prof. Guoqiang Mao, and my academic co-advisor A/Prof. Mehran Abolhasan for their mentor-ship, and support during my PhD study, which has been a thoughtful and rewarding journey.

I would like to thank all my colleagues in my research group: Jieqiong Chen, Tian Ding, Peibo Duan, Junnan Yang, Wenwei Yue and Yimeng Feng. The great time we had together has made this PhD journey more memorable.

I would like to thank my parents. They are always supportive and encouraging with their best wishes.

Abstract

With rapid urbanization, ITS (Intelligent Transportation Systems) has been deployed in some metropolitan areas to relieve traffic congestion and traffic accidents by traffic flow prediction and optimum traffic control. Due to temporary deployment of sensors, sensor malfunction and lossy communication systems, data missing problems has drawn significant attention from both academia and industry. Missing traffic data problem has negative impact on traffic flow prediction and optimum traffic control because ATIS (Advance Traveler Information Systems) and ATMS (Advance Traffic Management Systems) both rely on reliable, accurate and consistent traffic data measurements. Furthermore, adaptive traffic control is the most effective method to relieve traffic congestion and maximize road capacity. In this thesis, an Optimum Closed Cut (OCC) based spatio-temporal imputation technique was proposed, which can fully exploit the spatial-temporal correlation and road topological information in urban traffic network. The road topological information and flow conservation law can be explored to further improve the estimation performance while reducing the number of sensors involved in the data imputation, hence improving the computational efficiency. Besides, this thesis investigated the fundamental limits of missing traffic data estimation accuracy in urban networks using the spatio-temporal random effects (STRE) model. Furthermore, a hybrid dynamical system was investigated, which incorporates flow swap process, green-time proportion swap process and flow divergence for a general network with multiple OD pairs and multiple routes. A novel control policy was proposed to fill the gap by only adjusting the green-time proportion vector, and a sufficient condition was derived for the existence of equilibrium of the dynamical system under the mild constraints that (1) the travel cost function and stage pressure function should be continuous functions; (2) the flow and green-time proportion swap processes project all flow and green-time proportion vectors on the boundary of the feasible region onto itself. The condition of unique equilibrium was derived for fixed green-time proportion vector and it is shown that with varying green-time proportion vector, the set of equilibria is a compact, non-convex set, and with the same partial derivative of travel cost function with respect to the flow and green-time proportion vectors. Finally, the stability of the proposed dynamical system was proved by using Lyapunov stability analysis.

Publication

1. S. Wang and G. Mao, “Missing Data Estimation for Traffic Volume by Searching an Optimum Closed Cut in Urban Networks”, IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 1, pp. 75-86, Jan. 2019. doi: 10.1109/TITS.2018.2801808.

2. S. Wang and G. Mao, “Fundamental Limits of Missing Traffic Data Estimation in Urban Networks”, IEEE Transactions on Intelligent Transportation Systems, pp. 1 - 13, April 2019, doi: 10.1109/TITS.2019.2903524

3. S. Wang, G. Mao and J. A. Zhang, “Joint Time of Arrival Estimation for Coherent UWB Ranging in Multipath Environment with Multi User Interference”, IEEE Transactions on Signal Processing, vol. 67, no. 14, pp. 3743- 3755, July 15, 2019. doi: 10.1109/TSP.2019.2916016

4. S. Wang, Changle Li, Wenwei Yue, G. Mao, “Network Capacity Maximization Using Route Choice and Signal Control with Multiple OD Pairs”, IEEE Transactions on Intelligent Transportation Systems, doi: 10.1109/TITS.2019.2909281

5. Peibo Duan, Guoqiang Mao, Bin Zhang, Changsheng Zhang, Shangbo Wang, “STARIMA-based Traffic Prediction with Time-Varying Lags”, in the 19th IEEE Intelligent Transportation Systems Conference, pp. 1610- 1615, 2016

6. Peibo Duan, Guoqiang Mao, Shangbo Wang, Wenwei Yue, “A Unified Spatio-Temporal Model for Short-term Traffic Flow Prediction”, 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 2018, pp. 1652- 1657.

7. Y. Feng, G. Mao, B. Cheng, B. Huang, S. Wang , J. Chen, "MagSpeed: A Novel Method of Vehicle Speed Estimation Through A Single Magnetic Sensor", accepted to appear in 22nd International Conference on Intelligent Transportation Systems (ITSC), Auckland, New Zealand

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Transportation systems play an important role in economic growth of all nations. With rapid urbanization, congestion, accidents, and pollution issues due to transportation are becoming more severe as a result of the rapid increase in various travel demands. It was reported in [24] that in the United States, commuters spend approximately 42h a year stuck in traffic, drivers waste more than 3 billion gallons of fuel per year, having a total nationwide price tag of $160 billion, equivalent to $960 per commuter. Such problems will become worse in the future because of population growth and the increasing migration to urban areas in many countries around the world as reported by the United Nations Population Fund [59] and the Population Reference Bureau [6]. Furthermore, severe traffic congestion is imminent in metropolitan areas with growing transportation demand and urbanization trend. Traffic congestion imposes several adverse effects on urban communities, such as increase of travel time, unproductive wasted fuel, high possibility of traffic accidents, noise pollution, etc. To relieve traffic congestion in metropolitan areas, high annual budget is spent by local and federal governments to improve transportation facilities. For instance, the Texas transportation institute estimated that the cost of congestion in US was more than $120 billion in 2011 [24]. Furthermore, congestion annually costs Europe about 1% of its gross domestic product (GDP) that is projected to increase by about 50% to nearly 200 billion Euro by 2050 [14]. Therefore, there is a strong motivation to improve the safety, efficiency, and alleviate traffic congestion in transportation networks.

To resolve such issues, ITS (Intelligent Transportation Systems) has been deployed in some metropolitan areas in recent decades, which have the functionality of integrating a broad range of systems, including sensing, communication, information dissemination, and traffic

control [55]. The main components of ITS are: data sensing, data transmission and data analysis. Data sensing traditionally utilizes inductive loop detectors or magnetic detectors to detect the presence of vehicles or the vehicle speed based on the induced current in the loop with passing vehicles [35, 52]. With development of sensing technology, radar sensors, video cameras and GPS devices are also widely being considered for use in traffic data collection. Data transmission components of ITSs help transmit the data provided by sensors to operation centers for analysis and evaluation, and disseminate analyzed results, and/or management/control strategies to travelers and infrastructures. Data analysis components of ITSs analyze the data measured by sensors and create various management/control strategies, such as the Advance Traveler Information Systems (ATIS), which acquire, analyze and present information to assist travelers navigating from the source to the destination, and the Advance Traffic Management Systems (ATMS), which integrate various technology to improve road traffic flow and road safety. For example, analysis of road-condition images from drivers taken automatically from smartphone applications can be used to determine available roadside parking in real time [73]. Zhou et al. proposed the strategy to predict bus arrival time from information transmitted by bus passengers through mobile phone signals across different cell towers [80]. Fu and Yang proposed bus-holding control strategies based on real-time bus location information to regulate bus headway at specific stops [23]. Kurkcu et al. proposed an extended virtual sensor framework by using open data source and social media data for incident detection, which is the crucial first step of incident-management procedures [30].

To provide accurate travel time, route guidance information and control strategies to travelers and infrastructure, ATIS and ATMS both rely on reliable, accurate and consistent traffic data provided by deployed sensors [70]. Missing data problem, where some subsets of traffic data become missing, has greatly hindered the collection and subsequent analysis and evaluation. Traffic data may become missing due to temporary deployment of sensors, sensor malfunction or lossy data transmission systems. Specifically, due to high deployment costs, permanent traffic sensors may be installed on a subset of roads only [7] and some other roads may only be equipped with temporary sensors, which can provide traffic data within limited time periods. Furthermore, failures, caused by sensor malfunction and lossy data transmission systems, may also result in incomplete traffic data [12, 39]. It was reported in [32] that at hundreds of detection points within PeMS (Performance Measurement System) traffic flow database, more than 5% of data are missing. The missing data has severe impacts on many ITS applications, most of which rely on reliable, accurate and complete data [27, 60, 64]. For instance, traffic flow prediction relies on the complete historical data and the prediction performance will reduce sharply with incomplete data . Therefore, developing methodologies to precisely estimate the missing data, i.e. traffic data imputation, is an important task.

A number of imputation methods have been proposed in recent decades. Existing imputation techniques can be generally classified into three categories: interpolation based, prediction based and statistical learning based methods . Some well-known interpolation based methods include correlative kNN (k Nearest Neighbor) scheme [8], sectional kNN scheme [56] and LLS (Local Least Squares) scheme [10]. Prediction based methods have Auto-regressive Integrated Moving Average (ARIMA) [21, 79], Seasonal ARIMA (SARIMA) [67], Space-Time ARIMA (ST-ARIMA) [25, 72]. The most frequently used statistical learning methods are Probabilistic Principal Component Analysis (PPCA) [39], Kernel Probabilistic Principal Component Analysis (KPPCA) [32] and tensor completion techniques [57]. Most existing traffic data imputation methods suffer from the following shortcomings: 1) there are few studies trying to find the optimum subset of sensors before data imputation. It is well known that choosing all sensor-measurements may improve the accuracy of imputation but significantly increase the computational complexity, which consequently result in the imputation method becoming non-scalable; 2) the spatial correlation of the traffic data has not been fully utilized; 3) previous work mostly neglects the road topological information, which can be further exploited to improve the accuracy of the traffic data imputation.

Vehicles traveling through a specific road during a certain time interval can be classified into three portions: i) vehicles arriving from some neighboring roads equipped with sensors (termed *measured roads*); ii) vehicles coming from some neighboring roads without sensors (termed *unmeasured roads*), iii) vehicles coming from sources or traveling to sinks within some specific sections (termed *flow generation or dissipation*). The first portion can be read from the sensors while sum of the second and the third one can be estimated from the empirical data. The road topology gives the sufficient information about on-ramp and off-ramp of each measured and unmeasured roads and thus can be exploited to alleviate the missing data problem. Intuitively, when drawing a closed circuit or a closed cut, on a road map, the total amount of *long-term* traffic entering into the closed circuit must be equal to the total amount of *long-term* outgoing traffic. This implies that traffic on the roads intersected by the closed circuit must be correlated. Indeed, an equality can be established that relate traffic on those intersected roads. Motivated by the intuition and the aforementioned shortcomings of existing data imputation techniques, in this thesis, an Optimum Closed Cut (OCC) based spatio-temporal imputation technique is proposed, where the OCC satisfies the following conditions: a) the closed cut intersects the target road; and b) traffic on other intersected roads has the maximum correlation with that on the target road; and c) the number of intersected roads is minimized. The proposed technique utilizes both the road topological information and the spatio-temporal correlation among road traffic for imputation, while using a minimum number of sensor measurements. Therefore, it strikes a fine balance between imputation accuracy and

computational complexity.

Furthermore, existing research efforts mostly deal with developing certain data-driven or model-driven methods for data imputation. There is a lack of study on the achievable estimation accuracy and the conditions to achieve accurate estimation results. In this thesis, imputation accuracy is characterized in terms of a performance measure called the Squared Flow Error Bound (SFEB) and the fundamental limits of missing data estimation is studied. The Fisher matrix is an important tool for evaluating the accuracy of the parameter estimation technique. The inverse of the Fisher matrix yields the Cramer-Rao Lower Bound (CRLB), which provides asymptotically a lower bound for the co-variance matrix of unbiased estimators [29]. The fundamental limits of missing data estimation accuracy is investigated under different scenarios: the Fisher matrix is a full-rank matrix, which is the sufficient condition to find an efficient estimator, or a singular matrix, with which an efficient estimator can be found under certain conditions. It also investigates the impact of relationship between the number of missing points and rank of the Fisher matrix on the fundamental limits. It shows the performance optimality of the spatio-temporal Kriging estimator proposed for missing data estimation.

Almost all proposed historical neighboring imputation methods, interpolation based methods and prediction based methods only dealt with either single missing data imputation or the case that the number of missing data is not large. The imputation performance of the proposed methods greatly depend on the surrounding observed data of the missing points. Thus, their performances suffer when the missing ratio goes high. Although the proposed PPCA/KPPCA based imputation methods utilize all observed data to fit missing points and theoretically can cope with multiple detector cases, the aforementioned limitations remain. In a presence of a large data set with a high missing ratio, some available data may have weak spatio-temporal correlation with the missing data. PPCA/KPPCA based approaches have dominant overfitting problem. The adopted interpolation based approaches, i.e., original kNN or correlative kNN, are not able to work well because a large portion of missing data leads to difficult selection of nearest neighbors . Intuitively, the missing data physically close to the observed data has a relatively high correlation and thus can be estimated with relatively high accuracy, then the estimation results can be considered as additional observed data in the next iteration for estimating missing data further away. Motivated by the intuition and the aforementioned shortcomings of existing imputation techniques, in this thesis, I propose an iterative multiple-point spatio-temporal Kriging techinque, which can greatly reduce the computational complexity when the size of the observed parameter vector becomes large and is proved to be optimal in estimation accuracy under certain conditions.

On the other hand, ITSs have been implemented in many countries to reduce traffic conge-

stion by improving traffic flows in transportation networks and managing the demand. How to economically utilize the spatio-temporal resources of roads and traffic signal settings to maximize road network capacity becomes an important issue. In the existing literature, capacity of a transportation network is defined as the maximum link dynamical route flow that the network can accommodate under a given travel demand. Without consideration of control policies such as route choice and traffic signal control, the capacity maximization problem becomes a max-flow min-cut problem that can be directly drawn from network flow and graph theory [1, 5]. In practice, however, the maximum throughput between one pair of OD (Origin-Destination) nodes may deviate from sum of the flows on the bottlenecks because of traffic signal control applied on each intersection and drivers' route choice.

To achieve an optimal throughput, there are some recent research efforts studying the impact on network capacity posed by optimizing isolated or coordinated traffic signals, and drivers' route [31, 49, 51, 71]. They mainly aim at answering the following questions: by a given traffic assignment model, 1) does the Wardrop equilibrium exist? 2) is the Wardrop equilibrium unique? 3) is the dynamical system stable? To this end, Smith et al. proposed P0 control policy [50] and proved that by using P0 control policy on a simple network, Wardrop equilibrium can be found by intentionally constructing bottleneck delay on each link, that is, the maximum network throughput can be achieved at a quasi-dynamic user equilibrium [49, 51]. Xiao and Lo investigated the interaction between adaptive traffic control and day-to-day route choice adjustments of travelers in a dynamical system setting [71]. Yang et al. modeled the capacity and level of service, and investigated the additional demand tolerance of urban networks [74]. Le et al proposed a decentralized traffic signal control policy to optimize the throughput on a very simple one junction network [31]. Smith proposed P0 control policy and showed that under natural conditions a simple network is stable under this policy, and this policy can achieve capacity maximization [48]. Chen and Kasikitwiwat proposed two approaches for assessing the value of capacity flexibility based on the concept of reserve capacity and demand changes, respectively. The authors considered the capacity flexibility influenced by user's free choice of routes and destinations [11]. Chiou presented a new solution that simultaneously maximizes increase in travel demands and minimizes link capacity expansion investment [13].

To the author's knowledge, almost all existing research efforts focus on seeking an optimal control policy to optimize throughput by considering only one OD pair in a simple network, without consideration of the flow divergence among different OD pairs, i.e., traffic may switch to another link/route (Fig. 1.1). P0 control policy proposed by Smith [49–51] has been verified to outperform other control policies such as equi-saturation policy [66] because it can maximize capacity at a quasi-dynamic user equilibrium on a simple network, which consists

of only one OD pair with a group of routes and does not have any flow divergence at each route. In an urban network, there may be multiple OD pairs, each of which is connected via multiple routes that may intertwine with those affiliated to different OD pairs (Fig. 1.2). In the presence of flow divergence among different OD pairs, the route flow swap process and traffic signal assignment process become much more intricate compared to those for a simple network because the traffic signal assignment process can also occur between different OD pairs, rather than just different routes for the same OD pair; the route flow swap process should take into account each link's cost on the same route. In summary, no existing work investigated the conditions for the existence and uniqueness of equilibrium and stability of the dynamical systems when flow swap process, green-time proportion swap process and flow divergence among different OD pairs are simultaneously considered.

1.2 Main Contribution

This thesis aims at tackling the problems for traffic data estimation and proposing an optimum traffic control policy for a realistic scenario to maximize the network capacity. Specifically, the main contributions of this thesis are as follows:

- An Optimum Closed Cut (OCC) based spatio-temporal estimation technique is proposed, which explicitly incorporates road topology information into data imputation while using a small number of sensor measurements;

- A graph-based technique is developed to select the OCC that achieves an optimum trade-off between the number of employed sensor measurements and the set of measured roads whose traffic has maximum correlation with that of the target road;

- The fundamental limits of traffic flow data estimation are investigated, and the Squared Flow Error Bound (SFEB) is derived for the cases that Fisher matrix is a full-rank and singular matrix, respectively;

- Optimal and iterated spatial-temporal Kriging estimators are proposed, and the conditions of optimality for iterated spatial-temporal Kriging estimator are given;

- A novel control policy is proposed for the general network to maximize the network capacity with a steady demand and vertical queue, and the superiority of the novel control policy over P0 control policy is shown;

- Based on P0 control policy and the novel control policy, a hybrid traffic assignment model incorporating the flow swap process, green-time proportion swap process and

flow divergence is proposed for the general network;

- The conditions for the existence and uniqueness of Wardrop equilibrium, and the characteristics of the set of equilibria are given and the stability of the dynamical system is proved by using Lyapunov stability analysis.

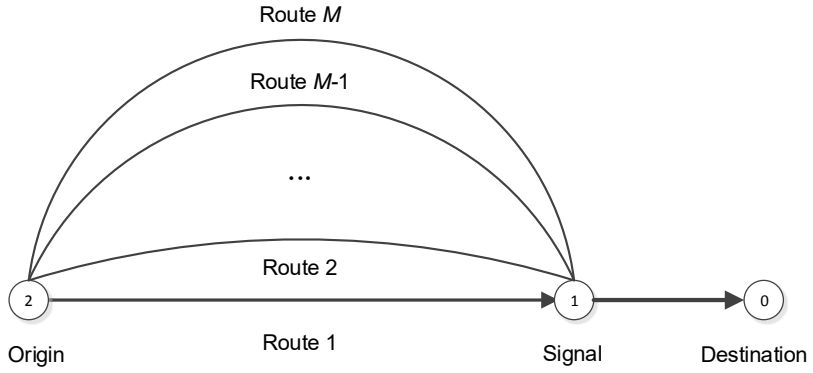

Figure 1.1: A M route signalized network proposed by Smith. Only one OD pair was considered

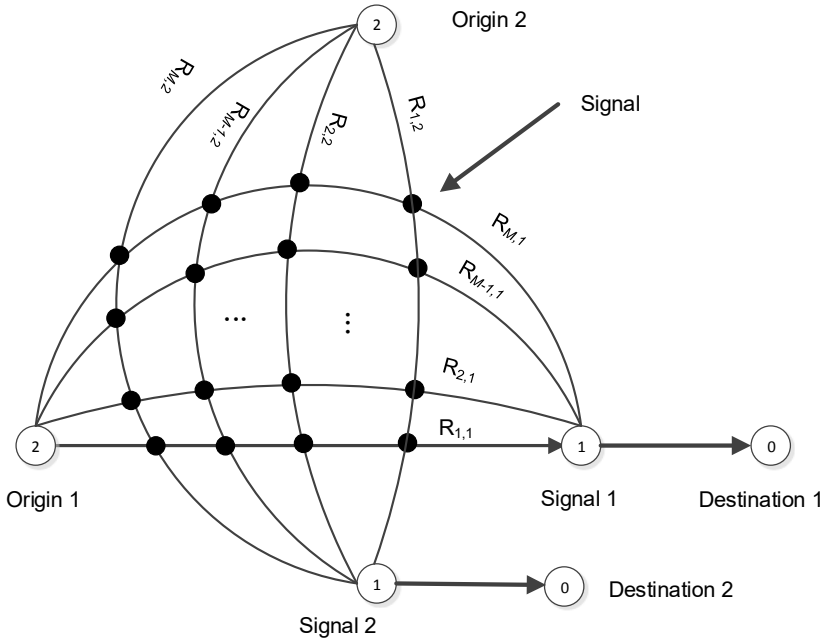

Figure 1.2: A signalized network with two OD pairs, where each intersection is equipped with traffic signal

1.3 Outline of the Dissertation

This thesis contains six chapters. Chapter 1 introduces the motivation and the main contributions of this thesis. Chapter 2 reviews the related work. Chapter 3 proposes the novel missing data estimator utilizing the topological information in senor-equipped road networks. Chapter 4 investigates the fundamental limits of missing traffic data estimation and proposes the iterated spatio-temporal Kriging estimator. Chapter 5 investigates the optimum traffic control policy in senor-equipped road networks. Chapter 6 concludes the thesis.

Chapter 2

Literature Review

2.1 Overview

This section is to review the state-of-the-art traffic data estimation technique and traffic control strategies with the emphasis of their drawbacks.

2.2 Literature Review for Missing Traffic Data Estimation

A number of imputation methods have been proposed in the recent decade. Existing imputation techniques can be generally classified into three categories: interpolation based, prediction based and statistical learning based methods [32]. Some well-known interpolation based methods include correlative kNN (k Nearest Neighbor) scheme [8], sectional kNN scheme [56] and LLS (Local Least Squares) scheme [10]. Prediction based methods only rely on known traffic prediction methods, e.g., Auto-regressive Integrated Moving Average (ARIMA) [21, 79], Seasonal ARIMA (SARIMA) [67], Space-Time ARIMA (ST-ARIMA) [25, 72]. The most frequently used statistical learning methods are Probabilistic Principal Component Analysis (PPCA) [39], Kernel Probabilistic Principal Component Analysis (KPPCA) [32] and tensor completion techniques [3, 57].

There are some studies having been carried out in exploiting spatial information to improve estimation performance. Tak et al. proposed sectional kNN method, which estimates missing data based on road sections sharing the same traffic property [56]. Cai et al. introduced the correlative kNN model, which was superior than the original kNN model because it replaces physical distances by the equivalent distances, which are determined by both the physical distance and the correlation coefficient between the historical traffic data of the two roads [8]. In [57] and [3], the authors explored the ability of tensor based method for multi-loop sensor's

missing data imputation, which completes the missing data by tensor decomposition. Qu et al. proposed the PPCA (Probabilistic Principal Component Analysis) based method, which integrated MLE (Maximum Likelihood Estimation) into traditional PCA (Principal Component Analysis) approach [39]. Li et al. compared PPCA method and KPPCA (Kernel Probabilistic Principal Component Analysis) method, which assumes a nonlinear relationship between observed samples and latent variables [32]. Wang and Pagageorgiou utilized the macroscopic traffic flow model and the extended Kalman-filtering (EKF) method to estimate the freeway traffic state [65]. The considered freeway is subdivided into N segments. Traffic flow at boundary of each segment and some important parameters constitute the state vector. The key differences from the technique proposed in this thesis are that [65] mainly utilized the time evolution and measurements to estimate the state vector whereas this thesis focuses on utilizing spatial-temporal correlation to estimate the missing data. Ng proposed a strategy, which aims at determining the smallest subset of links in a traffic network for counting sensor installation in order to infer flows on all remaining links [37]. Ng presented the condition that all link flows can be inferred and proposed the inference method. Viti et al. studied the network sensor location problem (NSLP), which considered the case that the variables are partially observed [63]. Castillo et al. dealt with the over-specified network observability problem, which aims at determining link flow based on a subset of observed OD pair and link flows [9]. The key difference between the technique presented in this thesis and those in [9, 37, 63] is that [9, 37, 63] dealt with network observability problem, which aims at optimizing the sensor location and determining link flow based on a subset of observed OD-pair and link flows, whereas my technique tries to utilize the spatio-temporal correlation between each crossed link and the target link to estimate the missing data caused by temporary sensor failure based on the empirical measured data.

2.3 Literature Review for Optimum Traffic Control

There are some literature dealing with traffic control strategies. Smith et al. presented a more special dynamical models of day-to-day rerouting and green-time response. They showed that with responsive P0 control policy, a simple network can achieve the maximum capacity at a quasi-dynamic user equilibrium, and queues and delays in the simple network remain bounded in natural dynamical evolution [48, 51]. They showed in [49] that applying P0 control policy in a simple network, the convergence of stage green-time can be guaranteed in the vertical queuing case. Xiao and Lo addressed the equilibrium stability and convergence of the proposed joint dynamic traffic system by analyzing the Jacobian matrix associated with each fixed point in a simple network [71]. Liu and Smith introduced the restrictive proportional adjust-

ment process (RPAP) for route swaps, which is more restrictive for defining routes compared to the original proportional adjustment process (PAP) proposed by [47]. Yang et al. adopted the model of equilibrium trip distribution/assignment with variable destination costs (ETDA-VDC) to assign the flow on each route when the maximum OD flows are obtained. They proposed a bi-level optimization model where the upper-level sub-problem aims at maximizing the sum of total trip generations and the lower-level sub-problem is the flow assignment. They considered the case of multiple OD pairs. However, the signal control and interference among multiple OD pairs were not considered [74]. Le et al. proposed a decentralized signal control policy, which adopted the BackPressure scheme [58] to allocate the proportion of the common cycle length to each phase [31]. The main objective of that paper was to prove the stability (limited queue length) of the proposed signal control policy. Route was not taken into account and capacity maximization at equilibrium was not investigated. Chen and Kasikitwiwat provided a quantitative assessment of capacity flexibility for the passenger transportation network using bi-level network capacity model [11]. Variations in demand pattern and volume were allowed in that paper. The utilized bi-level optimization model was similar to that proposed in [74]. Signal control was not considered. Chiou aimed at optimizing traffic assignment by minimizing travelers' delay and maximizing reserve capacity [13] using the TRANSYT model [62]. Nie proposed a class of bush-based algorithms for the user equilibrium traffic assignment problem, which is in fact a variant of Beckmann's seminal formulation [4, 38]. Mounce and Carey proposed the route swap algorithm that the swap rate is proportional to the flow on the more costly route multiplied by the cost difference between two routes, and proved that with the proposed route swap algorithm, the dynamical system can converge to a global equilibrium [36]. The aforementioned review did not investigate the capacity maximization at equilibrium by simultaneously considering flow swap and green-time proportion swap with flow divergence among different OD pairs.

2.4 Gaps in the Literature

Most existing traffic data imputation methods suffer from the following shortcomings: 1) there are few studies trying to find the optimum subset of detectors before imputation. It is well known that choosing all detector measurements may improve the accuracy of imputation but significantly increase the computational complexity, which consequently results in the imputation method becoming non-scalable; 2) the spatial correlation of the traffic data has not been fully utilized; 3) previous work mostly neglects the road topological information, which can be further exploited to improve the accuracy of the traffic data imputation. In summary, the aforementioned literature for missing traffic data estimation reveals that most existing studies

do not consider the problem of finding the optimum set of sensors for imputation. They either collected traffic data from all sensors or considered the sensors satisfying some given (often arbitrarily set) conditions.

Furthermore, almost all proposed historical neighboring imputation methods, interpolation based methods and prediction based methods only dealt with either single missing data imputation or the case that the number of missing data is not large. The imputing performances of the proposed methods greatly depend on the surrounding observed data of the missing points. Thus, their performances suffer when the missing ratio goes high. Although the proposed PPCA/KPPCA based imputation methods utilize all observed data to fit missing points and theoretically can cope with multiple detector cases, the aforementioned limitations remain. In a presence of a large data set with a high missing ratio, some available data may have weak spatio-temporal correlation with the missing data. PPCA/KPPCA based approaches have dominant overfitting problem. The adopted interpolation based approaches, i.e., original kNN or correlation kNN, are not able to work well because a large portion of missing data leads to difficult selection of nearest neighbors.

Existing research efforts on traffic data imputation mostly deal with developing certain data-driven or model-driven methods for missing data estimation. There is a lack of study on the achievable estimation accuracy and the conditions to achieve accurate estimation results. CRLB, as a well-known criterion, has been widely used in various applications for performance evaluation. For instance, Shen et al. derived the Squared Position Error Bound (SPEB) via equivalent Fisher information (EFI) for a general framework and cooperative wide-band localization systems [43, 44]. They introduced the notion of EFI to reduce the matrix dimensions by using Schur complement and investigated the cases that a priori knowledge of the channel parameters and agent's position to the agent's localization information is available or unavailable. Yu and Dutkiewicz derived CRLB for mobile tracking in non-line-of-sight (NLoS) environment when measurements of distance, heading angle, and velocity are employed to estimate the mobile position [76]. The results were only derived for the scenario where the range bias error is Rayleigh distributed and the heading and velocity errors are Gaussian distributed. Liang et al. proposed the posterior CRLB for mobile tracking problem in mixed line of sight (LoS) and NLoS conditions [33]. Rashid et al. derived CRLB for a general multi-channel spaceborne synthetic aperture radar system for ground moving target indication, which can be readily applied to an arbitrary antenna switching configuration [40]. Therefore, in this thesis, I characterize estimation accuracy in terms of a performance measure called the Squared Flow Error Bound (SFEB) and focus on the study of the fundamental limits of missing data estimation accuracy by using Fisher matrix to derive SFEB. The Fisher matrix is an important tool for evaluating the accuracy of the parameter estimation technique. The

inverse of the Fisher matrix yields the Cramer-Rao Lower Bound (CRLB), which provides asymptotically a lower bound for the covariance matrix of unbiased estimators. I investigate the fundamental limits of missing data estimation accuracy under different scenarios: the Fisher matrix is a full-rank matrix, which is the sufficient condition to find an efficient estimator or a singular matrix, with which an efficient estimator can be found under certain conditions. I also investigate the impact of relationship between the number of missing points and rank of the Fisher matrix on the fundamental limits.

For aforementioned literature for optimum traffic control, almost all existing research efforts focus on seeking on optimal control policy to optimize throughput by considering one or multiple OD (Origin-Destination) pair, without consideration of the flow divergence among different OD pairs caused by the turning matrix. Flow divergence is defined by the traffic switching to another link/route midway in its route when the routes of multiple OD pair intersect each other. P0 control policy proposed by Smith [49–51] has been verified to outperform other control policies such as equi-saturation policy [66] because it can maximize capacity at a quasi-dynamic user equilibrium on a simple network, which consists of multiple OD pairs with a group of routes and does not have any flow divergence at each route. In an urban network, there may be multiple OD pairs, each of which is connected via multiple routes that may interwine with those affiliated to different OD pairs. In the presence of flow divergence among different OD pairs, the route flow swap process and traffic signal assignment process become much more intricate compare to those for a simple network because the traffic signal assignment process can also occur between different OD pairs, rather than just different routes for the same OD pair; the route flow swap process should take into account each link's cost on the same route. In summary, no existing work investigated the conditions for the existence and uniqueness of equilibrium and stability of the dynamical systems when flow swap process, green-time proportion swap process and flow divergence among different OD pairs are simultaneously considered.

Chapter 3

Novel Missing Data Estimation Strategy in Sensor-Equipped Road Networks

3.1 Overview

This chapter introduces the novel missing data estimation strategy, which incorporates topological information to improve the estimation actuary. Firstly, a formal definition of the problem being considered in this chapter is given. Then, the OCC based spatio-temporal estimation technique is proposed and finally, the proposed estimation strategy is evaluated by comparing to some existing methods. Various missing type and missing ratios are observed in performance comparison among the estimation strategies in terms of MAPE (Mean Absolute Percentage Error) and RMSE (Root Mean Squared Error).

3.2 Problem Formulation

Suppose that there is a total of N_l links in a sensor-equipped road network, and there are N_{m} roads equipped with permanent or temporary detectors while the rest have no detector. Furthermore, each detector measures the data with the same sampling rate and delivers maximum M data points each day, and there are K observed days. Then, the traffic data can be viewed as a tensor $T \in \mathbb{R}^{N_{\mathrm{m}} \times M \times K}$. Denote the set of missing data in T by Q_{miss} and let N_{miss} be cardinality of Q_{miss}. Each element of Q_{miss} can be represented as q_{nmk} (true value of the missing data), where the subscripts n, m and k are the n-th road, m-th data point and k-th day, respectively. The missing data imputation aims at finding a function of the available measured data $f_{nmk}(T \backslash Q_{\mathrm{miss}})$ to obtain the most likely estimates of Q_{miss} to minimize MAPE (Mean Absolute Percentage Error) and RMSE (Root Mean Squared Error) of the estimates, defined

by

$$\text{MAPE} = \frac{1}{N_{\text{miss}}} \sum_{q_{nmk} \in Q_{\text{miss}}} \frac{| f_{nmk}(T \backslash Q_{\text{miss}}) - q_{nmk} |}{q_{nmk}} \tag{3.1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{q_{nmk} \in Q_{\text{miss}}} | f_{nmk}(T \backslash Q_{\text{miss}}) - q_{nmk} |^2}{N_{\text{miss}}}} \tag{3.2}$$

3.3 OCC based Strategy

Prior to explaining the OCC based strategy, I will firstly review the SRE (Spatial Random Effects) model, which has been applied in [70] and [26]. Then, the SRE model will be applied to the OCC search algorithm. Cressi et al. defined the SRE model as a summation of the large-scale spatial variation, smooth small-scale spatial variation and the measurement error, where the unknown random variables are fixed in number, statistically independent, and coefficients of known but not-necessarily-orthogonal spatial basis function [18]. In a traffic network, the measured traffic flow $Z(s)$ at a finite number of locations $s = \{ \ s_1 \ \cdots \ s_N \ \}$ can be expressed by [19, 70]

$$Z(s) = X(s)^{\mathrm{T}} \beta + B(s)^{\mathrm{T}} \eta + \xi(s) + \varepsilon(s) \tag{3.3}$$

where the product of $X(s)^{\mathrm{T}}$ and β can be understood as the weighted sum of the average traffic flow from the L selected neighboring roads, the length of β represents the number of selected neighboring roads, $B(s)^{\mathrm{T}} \eta$ can be interpreted as the fluctuation caused by the varied traffic flow from the L selected neighboring roads, $\xi(s)$ can be understood as a fine-scale variability on s due to the nugget effect and flow generation or dissipation within some specific sections, and $\varepsilon(s)$ denotes the measurement error. In geostatistics, nugget effect represents the discontinuity at the beginning of semivariogram graphs, which is generally caused by inadequate sampling size [16].

3.3.1 Traffic Flow Analysis with a Closed Cut

Consider a two-dimensional traffic network with N_l single lane bi-directional roads (Fig. 3.1), which are composed of N_{m} roads with detectors and N_{un} roads without detectors, $N_l = N_{\text{m}} + N_{\text{un}}$.

The target road is defined as the road, on which the traffic flow should be estimated. The flow on the target road toward one direction is caused by the flow on its neighboring roads and

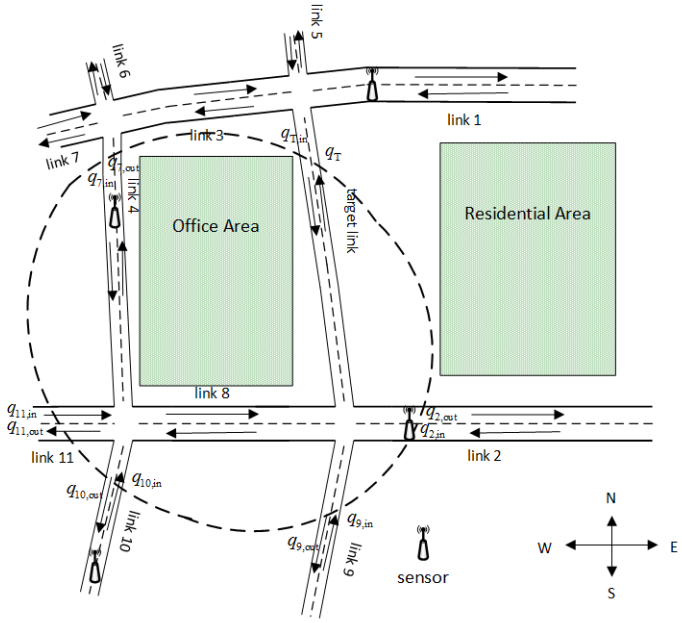

Figure 3.1: Illustration of a two-dimensional urban network with sensors

the flow directly injected to sinks or dissipated from sources within some specific sections. For instance, the flow on the target road toward north consists of portion of eastbound flow from road 8, westbound flow from road 2, northbound flow from road 9 and the missing flow, where the flow from road 2 can be acquired from the detector while the road 8 and road 9 show a lack of data. To investigate the flow relation between the target road and its neighboring roads, we propose Theorem 3.1, which gives the mathematical expression for the continuous flow relation between two roads.

Theorem 3.1. *Consider a simple case of two directional roads and define* $q_1(\tau)$, $q_2(\tau)$, $\mu_1(\tau)$ *and* $\mu_2(\tau)$ *as traffic flow and traffic density on the two roads, respectively. Further assume that traffic flow on road 2 all comes from road 1, the case that traffic from road 1 may divert to other roads is allowed. Assume that* $\mu_1(\tau)$ *and* $\mu_2(\tau)$ *are smaller than* μ_m *defined by the limiting density [34]. Then the relation between* $q_1(\tau)$ *and* $q_2(\tau)$ *can be approximated by*

$$\begin{aligned} q_2(\tau) &\approx q_1(\tau) \otimes h_{12,m}(\tau,t) \\ &= \int_0^{\tau} q_1(t)\, h_{12,m}(\tau - t, t)\, dt \end{aligned} \tag{3.4}$$

where "$\otimes$*" is two-dimensional convolution operation and* $h_{12,m}(\tau,t)$ *is a time-varied correlation coefficient function between two links.*

Proof. Partition the input flow stream on road 1 into N cells, each of which is with a short distance δx. Thus, the length of input flow stream L_1 can be expressed by $L_1 = N\delta x$ and there are $\mu_1(\tau)\delta x$ vehicles traveling through road 1 at τ-th time slot. For an arbitrarily selected cell, its length may expand to $\delta x'$ when the cell arrives at the sensor 2 because of the speed variation and difference of the vehicles within the cell. Here we only consider the case that $\delta x' > \delta x$, because shrink of the partial stream length can be viewed as overlap of two adjacent expanded cells. The density within the short distance δx and $\delta x'$ are assumed to be constant because δx and $\delta x'$ are small enough. Thus, there are $\mu_1(\tau)\delta x$ and $\mu_2(\tau)\delta x'$ vehicles traveling through the road 1 and 2 at τ-th time slot, respectively. Then, the number of vehicles traveling through road 2 at τ-th time slot can be expressed by

$$\mu_2(\tau)\delta x' = \sum_{i=1\cdots N} \mu_1(\tau_i)\delta x p_{i,12,\mathrm{m}}(\tau) \tag{3.5}$$

where τ_i is the time slot in which the i-th cell travels through road 1, $p_{i,12,\mathrm{m}}(\tau)$ is a function that represents the fraction of vehicles expanding over time zone for the i-th cell arriving at road 2, and then for each cell, it has

$$\begin{cases} p_{i,12,\mathrm{m}}(\tau) = 0 & \text{if the } i\text{-th cell is diverted to other links} \\ \int_0^\infty p_{i,12,\mathrm{m}}(\tau) = 1 & \text{if the } i\text{-th cell arrives at the link 2} \end{cases}$$

When $\delta x, \delta x' \to 0$, (3.5) can be transformed as

$$\begin{aligned} \mu_2(\tau) &= lim_{\delta x,\delta x'\to 0}\left(\sum_{i=1\cdots N} \mu_1(\tau_i)\frac{\delta x}{\delta x'} p_{i,12,\mathrm{m}}(\tau)\right) \\ &= \int_0^\tau \mu_1(t)\, h_{12,\mathrm{m}}(\tau - t, t)\,\mathrm{d}t \end{aligned} \tag{3.6}$$

where $h_{12,\mathrm{m}}(\tau,t)$ is a time-varied correlation coefficient function between road 1 and 2. If $\mu_1(\tau)$ and $\mu_2(\tau)$ are smaller than the limiting density, the mean speed will be unaffected and flow-density curve is closed to linearity [34]. Thus, (3.4) can be obtained via multiplying both sides of (3.6) with the mean speed at the input and output streams, respectively. □

Remark 3.2. It is worth noting that Theorem 3.1 is also valid for the case that the two roads are not adjacent to each other. In such case, all road segments between the two roads traveled through by the input stream can be virtualized as an intersection with a delay function. Any congestion occurring between roads 1 and 2 has no impact on the validity of Theorem 3.1. In that case, the time delay caused by congestion occurring between the two roads is given by

the time-varying correlation coefficient function between two roads $h_{12,\mathrm{m}}(\tau,t)$. In the case that there is congestion on road 1 or road 2, Theorem 3.1 can be slightly modified to express a relation in terms of traffic volume between each road, instead of flow:

$$\begin{aligned}\int q_2(\tau)\,\mathrm{d}\tau &= \int\int_0^{\tau} q_1(t)\,h_{12,\mathrm{m}}(\tau-t,t)\,\mathrm{d}t\mathrm{d}\tau \\ &= \int q_1(t)\int_0^{\tau} h_{12,\mathrm{m}}(\tau-t,t)\,\mathrm{d}\tau\mathrm{d}t \\ &= g_{12,\mathrm{m}}\int q_1(\tau)\,\mathrm{d}\tau\end{aligned}$$

where $g_{12,\mathrm{m}}$ is the correlation coefficient in terms of traffic volume between two roads, and can be expressed by

$$g_{12,\mathrm{m}} = \frac{\int q_1(t)\int_0^{\tau} h_{12,\mathrm{m}}(\tau-t,t)\,\mathrm{d}\tau\mathrm{d}t}{\int q_1(t)\,\mathrm{d}t}$$

The assumption in Theorem 3.1 is fulfilled because our employed data shows that no congestion occurs on the inspected roads. Suppose the number of selected feeding sources of the target road is K, which consists of K_{m} measured roads and K_{un} unmeasured roads. From Theorem 3.1, the flow on the target road can be expressed by

$$q_{\mathrm{T}}(\tau) \quad \approx \quad \bar{q}_{\mathrm{T}}(\tau) \quad + \quad C_{T,i}(\tau,t)\,C_{i,i}^{-1}(t) \quad \otimes_V \quad (q_i(\tau)-\bar{q}_i(\tau)) \quad + \quad w_{\mathrm{T}} \tag{3.7}$$

where $\bar{q}_{\mathrm{T}}(\tau)$ is the average flow on the target road at τ-th time slot, w_{T} is the missing flow and can be modeled as a stationary non-zero mean Gaussian variable, $q_i(\tau)$ and $\bar{q}_i(\tau)$ are $K\times 1$ vectors containing the instantaneous and average flow on the K feeding sources, $C_{T,i}(\tau,t)$ and $C_{i,i}(t)$ are a time-varied $1\times K$ Cross Correlation Matrix (CCM) between $q_{\mathrm{T}}(\tau)$ and $q_i(\tau)$, and a time-varied $K\times K$ Auto Correlation Matrix (ACM) of $q_i(\tau)$, respectively. By the definition, $C_{T,i}(\tau,t)$ and $C_{i,i}(t)$ can be expressed by

$$\begin{aligned}C_{T,i}(\tau,t) &= E\left((q_{\mathrm{T}}(\tau)-\bar{q}_{\mathrm{T}}(\tau))\,(q_i(t)-\bar{q}_i(t))^{\mathrm{T}}\right) \\ C_{i,i}(t) &= E\left((q_i(t)-\bar{q}_i(t))\,(q_i(t)-\bar{q}_i(t))^{\mathrm{T}}\right)\end{aligned} \tag{3.8}$$

Note that "$\otimes_V$" is vector convolution operation. For example, given two vectors a and b with the elements $a_{i,i=1\cdots N}(t)$ and $b_{i,i=1\cdots N}(t)$, respectively. Then the convolution of two vectors

can be expressed by

$$a^{\mathrm{T}}b=\left[\begin{array}{cccc} a_1(t) & a_2(t) & \cdots & a_N(t) \end{array}\right]\otimes_V\left[\begin{array}{c} b_1(t) \\ b_2(t) \\ \vdots \\ b_N(t) \end{array}\right]=\sum_{i=1\cdots N} a_i(t)\otimes b_i(t)$$

Remark 3.3. Note that (3.7) is valid no matter whether the corresponding road traffic is statistically independent or correlated. Equation (3.7) gives a general form to express the relation between the target flow and its neighboring flows for both statistically independent and correlated cases. In the case of correlated traffic, $C_{i,i}^{-1}(t)$ is a $K\times K$ auto correlation matrix, $C_{T,i}(\tau,t)$ is a $1\times K$ vector. It follows that $C_{T,i}(\tau,t)C_{i,i}^{-1}(t)$ is a $1\times K$ vector, each element of which can be represented as $h_{\mathrm{T}l}(\tau,t)$. In the case of independent traffic, $C_{i,i}(t)$ becomes a diagonal matrix and thus $h_{\mathrm{T}l}(\tau,t)$ becomes the correlation coefficient.

Then (3.7) can be rewritten as

$$q_{\mathrm{T}}(\tau)\approx\bar{q}_{\mathrm{T}}(\tau)+\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}(\tau)\otimes h_{\mathrm{T}l,\mathrm{m}}(\tau,t)+\sum_{j=1}^{K_{\mathrm{un}}}\tilde{q}_{j,\mathrm{un}}(\tau)\otimes h_{\mathrm{T}j,\mathrm{un}}(\tau,t)+w_{\mathrm{T}}-\mu_w \tag{3.9}$$

where $\tilde{q}_{l,\mathrm{m}}(\tau)$, $\tilde{q}_{j,\mathrm{un}}(\tau)$, $h_{\mathrm{T}l,\mathrm{m}}(\tau,t)$ and $h_{\mathrm{T}j,\mathrm{un}}(\tau,t)$ are the flow variation on the l-th measured and j-th unmeasured roads, and the time-varied correlation coefficient function between the target and the l-th measured and j-th unmeasured roads, respectively. The flow variation $\tilde{q}_{l,\mathrm{m}}(\tau)$ and $\tilde{q}_{j,\mathrm{un}}(\tau)$ denote the gap between the instantaneous flow and the average flow on the l-th and j-th roads, w_{T} is a non-zero mean Gaussian variable and $w_{\mathrm{T}}\sim N(\mu_w,\sigma_w)$. Note that $q_{l,\mathrm{m}}(\tau)$ and $q_{j,\mathrm{un}}(\tau)$ are stochastic process, which varies day to day at the same time slot. By the assumption of $q_{l,\mathrm{m}}(\tau)$ and $q_{j,\mathrm{un}}(\tau)$ keeping constant over a short time interval, (3.9) can be rewritten as

$$\begin{aligned} q_{\mathrm{T}}(\tau)\approx\bar{q}_{\mathrm{T}}(\tau)&+\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}\int_0^{\tau}h_{\mathrm{T}l,\mathrm{m}}(\tau-t,t)\,\mathrm{d}t \\ &+\sum_{j=1}^{K_{\mathrm{un}}}\tilde{q}_{j,\mathrm{un}}\int_0^{\tau}h_{\mathrm{T}j,\mathrm{un}}(\tau-t,t)\,\mathrm{d}t+w_{\mathrm{T}}-\mu_w \end{aligned} \tag{3.10}$$

where $\bar{q}_{\mathrm{T}}(\tau)$ corresponds to the element of $X(s)^{\mathrm{T}}\beta$ in (3.3), the sum of the second and the third terms can be transformed to the element of $B(s)^{\mathrm{T}}\eta$ via Karhunen-Loeve expansion. Note

that $\varepsilon(s)$ in (3.3) is neglected because the measurement process is assumed to be identical to the hidden process. For simplicity, we define $g_{\mathrm{T}l,\mathrm{m}}(\tau) = \int_0^\tau h_{\mathrm{T}l,\mathrm{m}}(\tau - t, t)\,\mathrm{d}t$ and $g_{\mathrm{T}j,\mathrm{un}}(\tau) = \int_0^\tau h_{\mathrm{T}j,\mathrm{un}}(\tau - t, t)\,\mathrm{d}t$. Then (3.10) can be rewritten as

$$q_{\mathrm{T}}(\tau) \approx \bar{q}_{\mathrm{T}}(\tau) + \sum_{l=1}^{K_{\mathrm{m}}} \tilde{q}_{l,\mathrm{m}} g_{\mathrm{T}l,\mathrm{m}}(\tau) + \sum_{j=1}^{K_{\mathrm{un}}} \tilde{q}_{j,\mathrm{un}} g_{\mathrm{T}j,\mathrm{un}}(\tau) + w_{\mathrm{T}} - \mu_w \tag{3.11}$$

where the average target flow $\bar{q}_{\mathrm{T}}(\tau)$ and $g_{\mathrm{T}l,\mathrm{m}}(\tau)$ can be obtained from the empirical data, $\tilde{q}_{l,\mathrm{m}}$ can be acquired by sensor measurements, $\tilde{q}_{j,\mathrm{un}}$, $g_{\mathrm{T}j,\mathrm{un}}(\tau)$ and w_{T} are unknown parameters. By modeling sum of $\sum_{j=1}^{K_{\mathrm{un}}} \tilde{q}_{j,\mathrm{un}} g_{\mathrm{T}j,\mathrm{un}}(\tau)$ and $w_{\mathrm{T}} - \mu_w$ as a zero-mean Gaussian variable with variance of σ_η^2, the probability function of $q_{\mathrm{T}}(\tau)$ by given $\bar{q}_{\mathrm{T}}(\tau)$ and $\sum_{l=1}^{K_{\mathrm{m}}} \tilde{q}_{l,\mathrm{m}} g_{\mathrm{T}l,\mathrm{m}}(\tau)$ can be expressed by

$$\Pr\left(q_{\mathrm{T}}(\tau) \,|\bar{q}_{\mathrm{T}}(\tau), \sum_{l=1}^{K_{\mathrm{m}}} \tilde{q}_{l,\mathrm{m}} g_{\mathrm{T}l,\mathrm{m}}(\tau)\right) = \frac{1}{\sqrt{2\pi}\sigma_\eta} \exp\left(-\frac{\begin{pmatrix} q_{\mathrm{T}}(\tau) - \bar{q}_{\mathrm{T}}(\tau) \\ -\sum_{l=1}^{K_{\mathrm{m}}} \tilde{q}_{l,\mathrm{m}} g_{\mathrm{T}l,\mathrm{m}}(\tau) \end{pmatrix}^2}{2\sigma_\eta^2}\right) \tag{3.12}$$

To maximize the conditional probability, the neighboring measured roads should be selected to minimize $E\left(\left(q_{\mathrm{T}}(\tau) - \bar{q}_{\mathrm{T}}(\tau) - \sum_{l=1}^{K_{\mathrm{m}}} \tilde{q}_{l,\mathrm{m}} g_{\mathrm{T}l,\mathrm{m}}(\tau)\right)^2\right)$, i.e., ESEE (Expected Squared Estimation Error), and meanwhile the neighboring unmeasured roads should be selected to minimize σ_η. However, σ_η is an unknown parameter. To tackle the problem, flow conservation law is utilized to find optimum set of neighboring unmeasured roads. In graph theory, the sum of flows entering the vertex (or a closed curve) is equal to the sum of flows leaving the vertex (the closed curve) if the vertex is neither (the closed curve does not contain) a source nor a sink [20]. Next, the flow conservation law will be applied to the road network depicted in Fig. 3.1.

Fig. 3.1 is modeled as a directed graph $G(V,E)$ with a source set S and a link set T, where V is the set of vertices and $E \in V \times V$ is the set of edges (road segments), respectively. A vertex $v_i \in V$ models a road intersection or an end of a road. An edge $e\left(v_i, v_j\right)$, which connects two vertices, represents a directed network segment. The size of source set and sink set are K_{S} and K_{T}, respectively. Then, Theorem 3.4 can be obtained as follows:

Theorem 3.4. *Let us create an arbitrary closed cut in $G(V,E)$ and define the flow on each intersected edge (road segment) as $f_i, i = 1 \cdots K_C$, where K_C is the total number of edges*

intersected by the closed cut. Furthermore, let $K_{T,in}$ and $K_{S,in}$ be the number of sinks and sources located inside of the closed cut, respectively, and let $f(v,t_k)$ and $f(s_l,u)$ be a flow from the vertex v to a sink t_k and a flow from a source s_l to the vertex u, respectively. Then, the following equality should hold:

$$\sum_{k=1}^{K_{C,in}} f_k - \sum_{l=1}^{K_{C,out}} f_l = \sum_{k=1}^{K_{T,in}} \sum_{v\in V} f(v,t_k) - \sum_{l=1}^{K_{S,in}} \sum_{u\in V} f(s_l,u) \tag{3.13}$$

where $K_C = K_{C,in} + K_{C,out}$ and $K_{C,in}$, $K_{C,out}$ are the number of flows entering the closed cut and leaving the closed cut, respectively.

Proof. The closed cut $\mathbb{C}$ partitions the graph into two disjoint vertex sets, denoted by V_1 and V_2. By defining a flow function between two sets of vertices X and Y as $f(X,Y) = \sum_{x\in X}\sum_{y\in Y} f(x,y)$, the left side of (3.13) can be expressed by $f(V_1,V_2)$. Reference [20] shows that for all $X \in V$, $f(X,X) = 0$, and for all $X,Y,Z \in V$ with $X \bigcap Y = \phi$, $f(X\bigcup Y,Z) = f(X,Z) + f(Y,Z)$ and $f(Z,X\bigcup Y) = f(Z,X) + f(Z,Y)$. Hence, it has

$$\begin{aligned} f(V_1,V_2) &= f(V_1,V) - f(V_1,V_1) \\ &= f(V_1,V) \\ &= f(S_{\text{in}},V) + f(T_{\text{in}},V) + f(V_1\setminus\{S_{\text{in}},T_{\text{in}}\},V) \end{aligned} \tag{3.14}$$

From the flow conservation law, $f(V_1\setminus\{S_{\text{in}},T_{\text{in}}\},V) = 0$. Hence, (3.13) can be rewritten as

$$\begin{aligned} f(V_1,V_2) &= f(S_{\text{in}},V) + f(T_{\text{in}},V) \\ &= \sum_{k=1}^{K_{\text{T,in}}} \sum_{v\in V} f(v,t_k) - \sum_{l=1}^{K_{\text{S,in}}} \sum_{u\in V} f(s_l,u) \end{aligned} \tag{3.15}$$

Theorem 3.4 is proved. □

Remark 3.5. It is worth noting that Theorem 3.4 ignored the storage capacity of roads, i.e., vehicles stored in the road segments enclosed by the closed cut. Therefore, strictly speaking, the relationship depicted in Theorem 3.4 only applies to *long-term* traffic flows where the storage capacity is of negligible impact. When Theorem 3.4 is applied to short-term traffic flows, the equality no longer holds strictly. Moreover, the loop detectors may also create some uncertainties about the number of the passing vehicles. The mismatch between incoming and outgoing traffic flows caused by storage capacity and measurement uncertainties caused by loop detectors can be captured by an error term or can be modeled by a source/sink inside the closed cut.

A closed cut (dotted line) is shown in Fig. 3.1 as a dotted line, which intersects the target road, road 7, road 11, road 10, road 9 and road 2. Applying Theorem 3.4, it has

$$\begin{aligned} q_{\mathrm{T}} + q_{7,\mathrm{out}} + q_{11,\mathrm{out}} + q_{10,\mathrm{out}} + q_{9,\mathrm{out}} + q_{2,\mathrm{out}} = \\ = q_{\mathrm{T,in}} + q_{7,\mathrm{in}} + q_{11,\mathrm{in}} + q_{10,\mathrm{in}} + q_{9,\mathrm{in}} + q_{2,\mathrm{in}} + w_s + w_t \end{aligned} \tag{3.16}$$

where q_{T} is the missing target flow and w_s and w_t are the (total) source flow and sink flow within the closed cut, respectively. Note that each vehicle will spend a different amount of time traveling from an entrance to an exit of the closed cut. Then for each arbitrary closed cut crossing the target road, the following relation can be obtained:

$$\begin{aligned} q_{\mathrm{T}} + \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{out}} + \sum_{l=1}^{K_{\mathrm{un}}} q_{l,\mathrm{out}} &= \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{in}} + \sum_{l=1}^{K_{\mathrm{un}}} q_{l,\mathrm{in}} \\ &+ w_s + w_t \end{aligned} \tag{3.17}$$

The above equation can be further rewritten in the following form:

$$\begin{aligned} q_{\mathrm{T}} + \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{out}} - \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{in}} &= \sum_{l=1}^{K_{\mathrm{un}}} q_{l,\mathrm{in}} - \sum_{l=1}^{K_{\mathrm{un}}} q_{l,\mathrm{out}} \\ &+ w_s + w_t \end{aligned} \tag{3.18}$$

where the right side of (3.18) is unknown and it has strong impact on σ_η^2 given by (3.12). Combined with (3.12), the optimization objective function to find the OCC can be expressed by

$$\begin{aligned} \mathrm{C_{occ}}\left(\mathrm{T}\right) &= \underset{\mathrm{C(T)}}{\operatorname{argmin}} \left\{ E \left[\begin{array}{c} \left(q_{\mathrm{T}} - \bar{q}_{\mathrm{T}} - \sum_{k=1}^{K_{\mathrm{m}}} \tilde{q}_{k,\mathrm{m}} g_{\mathrm{T}k,\mathrm{m}}\right)^2 + \\ + \left(q_{\mathrm{T}} + \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{out}} - \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{in}}\right)^2 \end{array} \right] \right\} \\ &= \underset{\mathrm{C(T)}}{\operatorname{argmin}} \left\{ \begin{array}{c} -2\sum_{k=1}^{K_{\mathrm{m}}} g_{\mathrm{T}k,\mathrm{m}} r_{\mathrm{T}k,\mathrm{m}} + g_{\mathrm{T,m}}^{\mathrm{H}} R_{\mathrm{C(T)}} g_{\mathrm{T,m}} \\ + \sum_{k=1}^{K_{\mathrm{m}}} r_{k,\mathrm{out}} + \sum_{k=1}^{K_{\mathrm{m}}} \sum_{l=1,l\neq k}^{K_{\mathrm{m}}} r_{kl,\mathrm{out}} \\ + 2\sum_{k=1}^{K_{\mathrm{m}}} r_{\mathrm{T}k,\mathrm{out}} - 2\sum_{k=1}^{K_{\mathrm{m}}} r_{\mathrm{T}k,\mathrm{in}} - \\ - 2\sum_{k=1}^{K_{\mathrm{m}}} \sum_{l=1}^{K_{\mathrm{m}}} r_{kl,\mathrm{in,out}} + \sum_{k=1}^{K_{\mathrm{m}}} r_{k,\mathrm{in}} \\ + \sum_{k=1}^{K_{\mathrm{m}}} \sum_{l=1,l\neq k}^{K_{\mathrm{m}}} r_{kl,\mathrm{m}} \end{array} \right\} \end{aligned} \tag{3.19}$$

where the first part of (3.19) is to maximize the conditional probability and the second part is to minimize the expected squared unknown metrics, $r_{\mathrm{T}k,\mathrm{m}}$, $R_{\mathrm{C(T)}}$, $r_{k,\mathrm{out}}$, $r_{kl,\mathrm{out}}$, $r_{\mathrm{T}k,\mathrm{out}}$, $r_{\mathrm{T}k,\mathrm{in}}$, $r_{kl,\mathrm{in,out}}$, $r_{k,\mathrm{in}}$ and $r_{kl,\mathrm{m}}$ are co-variance between the target flow and the k-th inflow, co-variance matrix of the K_{m} input flows, variance of the k-th outflow, co-variance between the k-th and l-th outflow, co-variance between the target flow and the k-th outflow, co-variance between the target flow and the k-th inflow, co-variance between the k-th inflow and the l-th outflow, variance of the k-th inflow and co-variance between the k-th inflow and the l-th inflow, respectively, and can be straightforwardly obtained by the empirical data. Equation (3.19) will be used in the next sub-chapter to find the OCC.

3.3.2 Novel OCC Search Algorithm

It shows in (3.19) that the optimization procedure is a minimum cut finding problem. Stoer–Wagner algorithm is a classical recursive algorithm, which can find the minimum cut in an un-directed graph [53]. Unfortunately, the algorithm cannot be applied in our scenario because we need find a minimum weighted closed cut in a directed graph. The closed cut should start and end at the target edge. Brute-force solution is to check all possible neighboring edges and select the one minimizing (3.19). However, the search complexity will increase exponentially with the number of edges. To tackle the problem, we propose an iterative searching strategy, which is a Modified version of Viterbi Algorithm (MVA). VA is a recursive optimal solution to the problem of estimating the state sequence of a discrete time finite-state Markov process observed in memory-less noise [41]. The finite-states and transition probabilities in VA are deterministic while MVA has non-deterministic states and transition probabilities for each iteration. Our computer validation shows that the OCC can be efficiently captured for each target link.

To describe MVA more clearly, I firstly start from the scenario that all edges are equipped with detectors and the empirical data are available for all roads. Then, I will extend MVA to the scenario of low density of detectors. For the former scenario, it only aims at minimizing ESEE given by the first term in (3.19). MVA can be interpreted as an iterative searching solution initiating from the target edge and try to find the optimal detector at each iteration, which can minimize ESEE. Let us define the finite-states at the i-th iteration as S_i, which contains N_i neighboring edges of l_{i-1} defined by the selected edge at the i-1-th iteration. The transition probability from the selected edge l_{i-1} to the s_i-th state is defined by $\pi_{l_{i-1}s_i}$, which is $1/N_i$. The ESEE at the i-th iteration for the n_i-th selected neighbor is represented by V_{i,n_i}. From (3.19), V_{i,n_i} can be determined by

$$
\begin{aligned}
V_{1,T} &= 0 \\
V_{i,n_i} &= \min_{l_{i-1}} \begin{pmatrix} V_{i-1,l_{i-1}} + g_{\mathrm{T}n_i,\mathrm{m}} r^{\mathrm{H}}_{l_{i-1}n_i} g_{\mathrm{T}l_{i-1},\mathrm{m}} + \\ g^{\mathrm{H}}_{\mathrm{T}l_{i-1},\mathrm{m}} r_{l_{i-1}n_i} g_{\mathrm{T}n_i,\mathrm{m}} + \\ g_{\mathrm{T}n_i,\mathrm{m}} r_{n_i} g^{\mathrm{H}}_{\mathrm{T}n_i,\mathrm{m}} - 2 r_{\mathrm{T}n_i} g_{\mathrm{T}n_i,\mathrm{m}} \end{pmatrix} \\
R_{n_i} &= \begin{bmatrix} R_{l_{i-1}} & r_{l_{i-1}n_i} \\ r^{\mathrm{H}}_{l_{i-1}n_i} & r_{n_i} \end{bmatrix} \\
g^{\mathrm{H}}_{\mathrm{T}n_{i,},\mathrm{m}} &= \begin{bmatrix} g^{\mathrm{H}}_{\mathrm{T}l_{i-1},\mathrm{m}} & g_{\mathrm{T}n_i,\mathrm{m}} \end{bmatrix}
\end{aligned} \tag{3.20}
$$

where l_{i-1} contains all selected roads at the i-1 -th iteration, $r_{l_{i-1}n_i}$, r_{n_i} and $r_{\mathrm{T}n_i}$ are the co-variance vector between l_{i-1} and n_i, variance of n_i and co-variance between T and n_i. Then, MVA can be described by Algorithm 3.1. In the line 3, a queue is created to store the ESEE and the selected road at the initial iteration, the "while" loop from the line 13 to line 23 selects each crossed edge by minimizing the ESEE at each iteration, and the OCC can finally be determined by tracing each edge back to its parent, which is defined by the selected edge in the last iteration. For the latter scenario, a number of unmeasured roads appear in the network. The flow conservation given by (3.19) need to be utilized to improve the estimation performance because of the lace of sensors. However, Algorithm 3.1 cannot be directly applied to the latter scenario because the second term in (3.19) cannot be determined unless a cut is pre-given. It makes impossible for Algorithm 3.1 to iteratively incorporate the flow conservation during the search procedure. To tackle the problem, an approximation is made for (3.19) aiming at minimizing the number of crossed unmeasured roads while minimizing the first term in (3.19). At each iteration, the edge is to be selected to be able to provide the maximum average Variance of the Hypothetical Means (VHM), which is interpreted as the difference between the variance and the conditional variance, then divided by the number of crossed roads. From the law of total variance, the variance of the target flow can be expressed by

$$
\begin{aligned}
\mathrm{var}\left(q_{\mathrm{T}}\right) &= E_{q_m}\left(var_{q_{\mathrm{T}}}\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)\right) + var_{q_m}\left(E_{q_{\mathrm{T}}}\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)\right) && (3.21) \\
&= ESEE + VHM && (3.22)
\end{aligned}
$$

The dual problem of ESEE minimization is to maximize the VHM at each iteration. The modified algorithm can be found in Algorithm 3.2.

Remark 3.6. Note that the number of cuts being found for the given area is determined by the empirical data and topological information. Therefore, the number of determined cuts

varies with different target road segment. For each target road segment, there is only one optimal cut, which is used to estimate the missing data. For an example, for the target roads "Snowy Mountains Highway" and "Monaro Highway", there are 4 and 18 closed cuts being determined via the proposed algorithm, respectively.

Algorithm 3.1 MVA for the former scenario

Input: target road T, graph G
Output: optimum closed cut OCC
Initialize an empty OCC, empty edge array $Ecell$ and a new queue Q with the initial information about T,$V_{1,T} = 0$;
While Q is not empty **do**
Get the ESEE V_{i-1} and the crossed edges l_{i-1} at $i-1$-th iteration, pop Q
  For $n_l = 1, 2, ..., L_{i-1}$ **do**
  Find the neighboring edges S_i for l_{i-1,n_l}
    For $n_i = 1, 2, ...,$ *number of neighboring edges* **do**
    If S_{i,n_i} approaches the target road
    Then remove S_{i,n_i} and continue
    Else update $V_{i-1,l_{i-1}}$ to V_{i,n_i} via (3.20)
    Search the minimum ESEE $V_{i,n_i,\min}$ through Q
      If $V_{i,n_i} = V_{i,n_i,\min}$
      Then $V_i = [V_{i-1} \quad V_{i,n_i}]$,$l_i = [l_{i-1} \quad S_{i,n_i}]$
      Else remove S_{i,n_i} and continue
    If S_{i,n_i} is a new edge
    Then add the new edge to $Ecell$
    End For
    If l_i is not empty
    Then construct a new element *info* with V_i and l_i
    push *info* into Q
  End For
End While
OCC is the concatenation of the *parent* field of each $Ecell$ element

3.3.3 OCC Based Novel Estimator

This sub-chapter proposes the OCC based Kriging estimator and the OCC based novel estimator, which incorporates the flow conservation law. After that OCC is determined by Algorithm 3.2, the missing data at the target road can be estimated via a Kriging estimator [16]

$$\hat{q}_{\mathrm{T,kriging}} = \bar{q}_{\mathrm{T}} + g_{\mathrm{T,m}}^{\mathrm{H}} \left(q_{\mathrm{m}} - \bar{q}_{\mathrm{m}} \right) \tag{3.23}$$

Algorithm 3.2 MVA for the latter scenario

The following modifications should be made:
line 11: update $V_{i-1,l_{i-1}}$ to V_{i,n_i}
line 12: Search the maximum expected variance improvement $V_{i,n_i,\max}$ through Q
line 13: **if** $V_{i,n_i} = V_{i,n_i,\max}$
The rest lines are identical to those in algorithm 3.1.

where q_{m} and $\bar{q}_{\mathrm{m}}$ contains the instantaneous flow and average flow for each crossed sensors, the vector $g_{\mathrm{T,m}}$ consists of K_{m} scaling factors $g_{\mathrm{T}k,\mathrm{m};k=1\cdots K_{\mathrm{m}}}$ between the target flow and the K_{m} measured flows and can be straightforwardly obtained by the empirical data. Then the conditional expectation $E(\hat{q}_{\mathrm{T}}|q_{\mathrm{m}})$ and the conditional variance $\mathrm{var}(\hat{q}_{\mathrm{T}}|q_{\mathrm{m}})$ can be expressed by

$$E(q_{\mathrm{T}}|q_{\mathrm{m}}) = \hat{q}_{\mathrm{T,kriging}} \tag{3.24}$$

and

$$\mathrm{var}(q_{\mathrm{T}}|q_{\mathrm{m}}) = \mathrm{var}(q_{\mathrm{T}}) - g_{\mathrm{T,m}}^{\mathrm{H}} R_{q_{\mathrm{m}}} g_{\mathrm{T,m}} \tag{3.25}$$

To our known, the conditional PDF (Probability Distribution Function) $\mathrm{P}(q_{\mathrm{T}}|q_{\mathrm{m}})$ is a Gaussian function [28]. Thus, the OCC based Kriging estimation can be formulated as

$$q_{\mathrm{T}}|q_{\mathrm{m}} = \hat{q}_{\mathrm{T,kriging}} + \zeta \tag{3.26}$$

where $\zeta \sim N(0, \mathrm{var}(q_{\mathrm{T}}|q_{\mathrm{m}}))$. To further reduce the uncertainty, the flow conservation law is incorporated. Let us define the OCC for the target road T as $C(\mathrm{T})$. From (3.18), q_{T} can be written as

$$\begin{aligned} q_{\mathrm{T}} &= \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{in}} - \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{out}} + \sum_{l=1}^{K_{\mathrm{un}}} q_{l,\mathrm{in}} - \sum_{l=1}^{K_{\mathrm{un}}} q_{l,\mathrm{out}} + w_s + w_t \\ &= \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{in}} - \sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{out}} + \gamma \end{aligned} \tag{3.27}$$

where $\gamma \sim N\begin{pmatrix} \sum_{l=1}^{K_{\mathrm{un}}} \bar{q}_{l,\mathrm{in}} - \sum_{l=1}^{K_{\mathrm{un}}} \bar{q}_{l,\mathrm{out}} \\ \sum_{l=1}^{K_{\mathrm{un}}} r_{l,\mathrm{in}} + \sum_{l=1}^{K_{\mathrm{un}}} r_{l,\mathrm{out}} + r_s + r_t \end{pmatrix}$, $r_{l,\mathrm{in}}$, $r_{l,\mathrm{out}}$, r_s and r_t are the flow variance for the in and out direction at the $\underline{l}$-th crossed edge, and the variance of generation flow and dissipation flow within the closed cut, respectively. Then the conditional expectation $E(q_{\mathrm{T}}|C(\mathrm{T}))$

and the conditional variance $\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)$ can be expressed by

$$E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)=\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}+\sum_{l=1}^{K_{\mathrm{un}}}\bar{q}_{l,\mathrm{in}}-\sum_{l=1}^{K_{\mathrm{un}}}\bar{q}_{l,\mathrm{out}} \tag{3.28}$$

and

$$\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)=\sum_{l=1}^{K_{\mathrm{un}}}r_{l,\mathrm{in}}+\sum_{l=1}^{K_{\mathrm{un}}}r_{l,\mathrm{out}}+r_s+r_t \tag{3.29}$$

The better estimation can be obtained by maximizing the joint probability function $\Pr\left(q_{\mathrm{T}}|q_{\mathrm{m}},C\left(\mathrm{T}\right)\right)$. With the Bayesian theorem, the objective function can be expressed by

$$\begin{aligned}
\hat{q}_{\mathrm{T,ML}} &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\Pr\left(q_{\mathrm{T}}|q_{\mathrm{m}},C\left(\mathrm{T}\right)\right)\right)\\
&= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\frac{\Pr\left(q_{\mathrm{m}},C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)\Pr\left(q_{\mathrm{T}}\right)}{\Pr\left(q_{\mathrm{m}},C\left(\mathrm{T}\right)\right)}\right)\\
&= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\Pr\left(q_{\mathrm{m}}|q_{\mathrm{T}}\right)\Pr\left(C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)\Pr\left(q_{\mathrm{T}}\right)\right)\\
&= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\Pr\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)/\Pr\left(q_{\mathrm{T}}\right)\right)\\
&= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(-\frac{\left(q_{\mathrm{T}}-E\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)\right)^2}{2\mathrm{var}\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)}+\frac{\left(q_{\mathrm{T}}-\bar{q}_{\mathrm{T}}\right)^2}{2\mathrm{var}\left(q_{\mathrm{T}}\right)}-\frac{\left(q_{\mathrm{T}}-E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)\right)^2}{2\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}\right)
\end{aligned} \tag{3.30}$$

By setting the first derivative of (3.30) with regard to q_{T} to zero, $\hat{q}_{\mathrm{T,ML}}$ can be expressed by

$$\hat{q}_{\mathrm{T,ML}}=\frac{\frac{\bar{q}_{\mathrm{T}}}{\mathrm{var}\left(q_{\mathrm{T}}\right)}-\frac{E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}{\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}-\frac{E\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)}{\mathrm{var}\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)}}{\frac{1}{\mathrm{var}\left(q_{\mathrm{T}}\right)}-\frac{1}{\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}-\frac{1}{\mathrm{var}\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)}} \tag{3.31}$$

where $E\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)$, $E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)$, $\mathrm{var}\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)$ and $\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)$ are given by (3.24), (3.28), (3.25) and (3.29), respectively. and they can be straightforwardly obtained by the empirical data.

Note that (3.30) and (3.31) are based on the assumption that the traffic flow on the crossed measured roads and the unmeasured roads are independent i.e. $\Pr\left(q_{\mathrm{m}},C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)=\Pr\left(q_{\mathrm{m}}|q_{\mathrm{T}}\right)\Pr\left(C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)$. In real scenario, however, it can depict the dependence between them. Thus, the conditional probability should be rewritten as

$$\Pr\left(q_{\mathrm{m}},C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)=\Pr\left(q_{\mathrm{m}}|q_{\mathrm{T}}\right)\Pr\left(C\left(\mathrm{T}\right)|q_{\mathrm{m}},q_{\mathrm{T}}\right) \tag{3.32}$$

where the latter term represents the conditional probability of sum of the unmeasured flow and the missing flow based on input flow and the target flow, while the former term stands for

the conditional probability of the input flow by giving the target flow. Because of the causal relationship between flows [2], the crossed flows on the OCC can be classified into causal flow and effect flow defined by the input and output of the OCC. Recall that γ is a set of unobserved data and can be modeled as a Gaussian variable, which is sum of the unmeasured input flow, the unmeasured output flow and the missing flow. The causal relation between q_{m} and γ can be utilized to obtain a more accurate PDF. Then, the second term of (3.32) can be transformed to

$$\begin{aligned}\Pr\left(C\left(\mathrm{T}\right)|q_{\mathrm{m}},q_{\mathrm{T}}\right) &= \frac{\Pr\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right),q_{\mathrm{m}}\right)\Pr\left(C\left(\mathrm{T}\right),q_{\mathrm{m}}\right)}{\Pr\left(q_{\mathrm{m}},q_{\mathrm{T}}\right)} \\ &= \frac{\Pr\left(C\left(\mathrm{T}\right),q_{\mathrm{m}}\right)\exp\left(-\frac{\left(\ q_{\mathrm{T}}-\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}+\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-E\left(C\left(\mathrm{T}\right)|q_{\mathrm{m}}\right)\ \right)^2}{2\mathrm{var}(C(\mathrm{T})|q_{\mathrm{m}})}\right)}{\Pr\left(q_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}\right)\sqrt{2\pi\mathrm{var}\left(C\left(\mathrm{T}\right)|q_{\mathrm{m}}\right)}}\end{aligned} \tag{3.33}$$

The reason that we write $\Pr\left(q_{\mathrm{m}},q_{\mathrm{T}}\right)=\Pr\left(q_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}\right)$ in (3.33) is the dependence between q_{m} and q_{T} has been taken into account in the first term of (3.32), and (3.33) only considers the causal relation between q_{m} and γ, q_{T} and γ. Hence, (3.30) can be improved as

$$\begin{aligned}\hat{q}_{\mathrm{T,ML,improved}} &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\Pr\left(q_{\mathrm{T}}|q_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right),q_{\mathrm{m}}\right)/\Pr\left(q_{\mathrm{T}}\right)\right) \\ &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\begin{array}{c}\frac{(q_{\mathrm{T}}-\bar{q}_{\mathrm{T}})^2}{2\mathrm{var}(q_{\mathrm{T}})}-\frac{(q_{\mathrm{T}}-E(q_{\mathrm{T}}|q_{\mathrm{m}}))^2}{2\mathrm{var}(q_{\mathrm{T}}|q_{\mathrm{m}})}- \\ \frac{\left(q_{\mathrm{T}}-\left(\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-\sum_{k=1,k\neq\mathrm{T}}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}\right)-E(C(\mathrm{T})|q_{\mathrm{m}})\right)^2}{2\mathrm{var}(C(\mathrm{T})|q_{\mathrm{m}})}\end{array}\right)\end{aligned} \tag{3.34}$$

Then, $\hat{q}_{\mathrm{T,ML,improved}}$ can be expressed by

$$\hat{q}_{\mathrm{T,ML,improved}} = \frac{\frac{\bar{q}_{\mathrm{T}}}{\mathrm{var}(q_{\mathrm{T}})}-\frac{\left(\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-\sum_{k=1,k\neq\mathrm{T}}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}\right)+E(C(\mathrm{T})|q_{\mathrm{m}})}{\mathrm{var}(C(\mathrm{T})|q_{\mathrm{m}})}-\frac{E(q_{\mathrm{T}}|q_{\mathrm{m}})}{\mathrm{var}(q_{\mathrm{T}}|q_{\mathrm{m}})}}{\frac{1}{\mathrm{var}(q_{\mathrm{T}})}-\frac{1}{\mathrm{var}(C(\mathrm{T})|q_{\mathrm{m}})}-\frac{1}{\mathrm{var}(q_{\mathrm{T}}|q_{\mathrm{m}})}} \tag{3.35}$$

and

$$\begin{aligned} E\left(C\left(\mathrm{T}\right)|q_{\mathrm{m}}\right) &= \bar{\gamma}+\sigma_{\gamma,q_{\mathrm{m}}}\sigma_{q_{\mathrm{m}},q_{\mathrm{m}}}^{-1}\left(q_{\mathrm{m}}-\bar{q}_{\mathrm{m}}\right) \\ \mathrm{var}\left(C\left(\mathrm{T}\right)|q_{\mathrm{m}}\right) &= \mathrm{var}\left(\gamma\right)-g_{\gamma,\mathrm{m}}^{\mathrm{H}}R_{q_{\mathrm{m}}}g_{\gamma,\mathrm{m}} \end{aligned} \tag{3.36}$$

where $\sigma_{\gamma,q_{\mathrm{m}}}$ and $g_{\gamma,\mathrm{m}}$ are the co-variance matrix and conditional correlation coefficients between γ and q_{m}, respectively, and can be easily obtained by the empirical data

3.4 Experimental Results with Real Traffic Data

In this sub-chapter, the proposed estimation strategies will be evaluated by comparing to other two imputation methods: NHA (Nearest Historical Average) and kNN. Various missing type and missing ratio are observed in performance comparison among the three imputation methods by MAPE and RMSE.

3.4.1 Data Description

Traffic flow data for novel missing data estimation strategies was provided by Sydney RMS (Roads and Maritime Services). The selected data was collected by loop detectors on the arterial roads located in Sydney south area over 198 days (Fig. 3.2).

Figure 3.2: Selected map for experiment: Sydney South Area

All measured roads are single-lance bi-directional roads. Each sensor provides the flow

data of 1-hour interval from 00:00 - 23:00h on each day. Each green node in Fig. 3.2 represents a sensor. The data can be represented as a tensor $T \in \mathbb{R}^{N_m \times M \times K}$, where $N_m = 10$, $M = 24$ and $K = 181$. The number of historical days K is chosen to be 181 instead of 198 because the weekend and public days are removed due to the different traffic pattern.

3.4.2 Generation of Missing Data

To evaluate the estimation performance of each methods, the missing data are intentionally generated with different missing ratio that ranges from 20% to 50% at every 10% increment as usual in the research field ([77, 78]). To verify the robustness of the proposed methods, three types of missing data in this thesis are considered: 1) Missing Completely at Random manner (MCR), where the missing points are independently and uniformly distributed over the spatio-temporal domain. This may occur due to temporary from power or communication failures [15]. 2) Missing Group Randomly in the Temporal domain (MGRT), where the missing points appear as a group of fixed length sequential points lost at one road, and the group is independently and uniformly distributed over the temporal domain. This may occur due to a prolonged physical damage, malfunction of communication device or temporary sensor deployment. 3) Not Missing Randomly (NMR), where the occurrences of missing data are scattered and simultaneous over different roads. NMR is often caused by a long time malfunction of the loop detectors [39].

3.4.3 Imputation Techniques for Comparison Analysis

NHA is the most common method in the data imputation because it shows a stable performance regardless of the missing data size with easy implementation [56]. NHA replaces the missing data by arithmetic average or weighted average of the nearest historical data [15]. NHA does not incorporate the information from neighboring roads at the same day and is based on the assumption that traffic pattern at the same sensor at the same time is similar from day to day. In this study, the missing data is filled with the arithmetic average data of the same time over 10 historical days.

The second comparison method is kNN method, which has been discussed in [8, 56]. The original kNN method is to fill the the missing data with arithmetic or weighted average of data on k neighboring roads. The k neighboring roads are selected by searching for the data with close physical distances with the target road. In [8], the authors proposed the improved kNN, which replaced the physical distance with the equivalent distance, which is related to the physical distance among roads h, connective grade of a road g and correlation coefficient between the historical time series of two road r. The k neighboring roads are selected by a

given suitable threshold of the equivalent distance. Then, the missing data on the target road is estimated by the arithmetic average of the data on the k neighboring roads.

3.4.4 Results and Discussion

In this section, estimation performance of three novel approaches will be examined: OCC based Kriging (3.23), OCC based ML (3.31) and improved ML (3.34), and they will be compared to correlative kNN [15] and NHA [8] in terms of MAPE and RMSE over different missing ratios and three missing patterns. The performance of the proposed approaches was evaluated with 198 days of the historical data. Missing data in testing were intentionally produced from the available data sheet and compared with the actual value for the performance evaluation.

Fig. 3.3 and Fig. 3.4 depict MAPE and RMSE of the three novel estimation methods for MCR pattern, respectively. The accuracy results of estimation represented by MAPE show that the three novel estimation methods dominantly outperform the correlative kNN and NHA over the missing ratio from 0.25 to 0.5. By incorporating the flow conservation law introduced in Theorem 3.4, the estimation performance can be further improved via OCC based ML and improved ML. Fig. 3.5, 3.6 and Fig. 3.7, 3.8 show the estimation performance for MGRT and NMR patterns, respectively. As shown in Fig. 3.5, OCC based ML and improved ML depict better estimation performance than correlative kNN and NHA over almost whole scale of missing ratios while OCC based Kriging is more appropriate for the missing ratio being lower than 0.35. Beyond 0.35, OCC based Kriging shows a worse performance than the comparing methods. For NMR pattern, the three novel approaches slightly outperforms the comparing methods over almost whole scale of missing ratio.

Comparing the three novel approaches, the improved ML shows the best estimation performance for three missing patterns because it incorporates the flow conservation law and takes into account the dependence between the measured and unmeasured roads.

Comparing the three missing patterns, three novel methods depict the best estimation performance for MCR pattern while the worst performance for NMR.

It is observed in Fig. 3.5 that the performance of OCC based Kriging for MGRT pattern becomes worse than NHA and correlative kNN when missing ratio is larger than 0.35. This is mainly because for group missing pattern, the number of the measured neighboring roads captured by OCC decreases quickly with the increase of the missing ratio and thus the correlation between the measured neighboring and the target roads plays less role compared to the flow conservation for the missing data estimation.

To summarize, the experimental results show that the three novel approaches perform better than the comparing methods in almost all missing pattern, with exception for MGRT for which the OCC based Kriging performs worse than the two comparing methods for the mis-

sing ratio between larger than 0.35. The improved ML outperforms all other methods for all missing patterns and missing ratios.

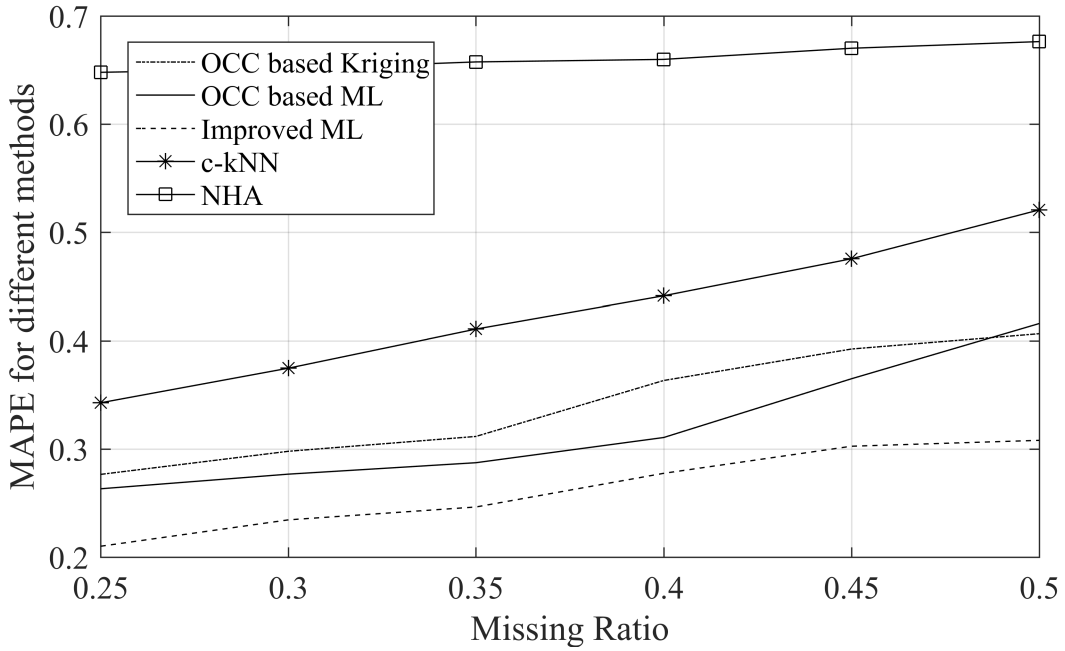

Figure 3.3: MAPE of five different methods in MCR pattern

Remark 3.7. Although in this thesis, only the non-congested case is considered. Theoretically, congestion and non-recurrent events pose no impact on the performance of our methods as long as the sampling period is much larger than the travel time. Because in this case, almost all traffic flow measured by the target detector originates from the flow measured by its upstream detectors during the same time slot. In the case that the travel time is much larger than the sampling period due to congestion or non-recurrent events, the time lag should be considered to improve the performance.

3.5 Summary

In this chapter, OCC based estimation strategies have been proposed for traffic flow incompleteness in urban network. Based on the determination of optimum sensors for imputation via novel OCC finding algorithm, three different estimators: OCC based Kriging, OCC based ML (Maximum Likelihood) estimator and improve ML estimator are compared in terms of MAPE for three missing patterns: MCR, MGRT and NMR. In addition, the novel methods are compared to NHA and correlative kNN. From the experimental results, the following conclusions can be drawn:

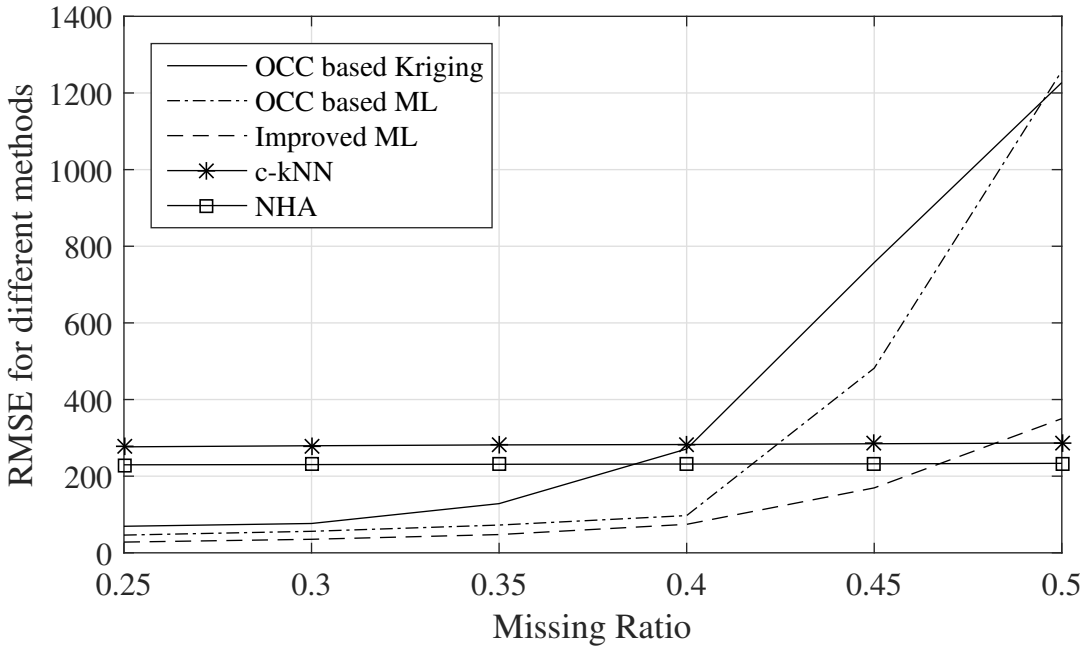

Figure 3.4: RMSE of five different methods in MCR pattern

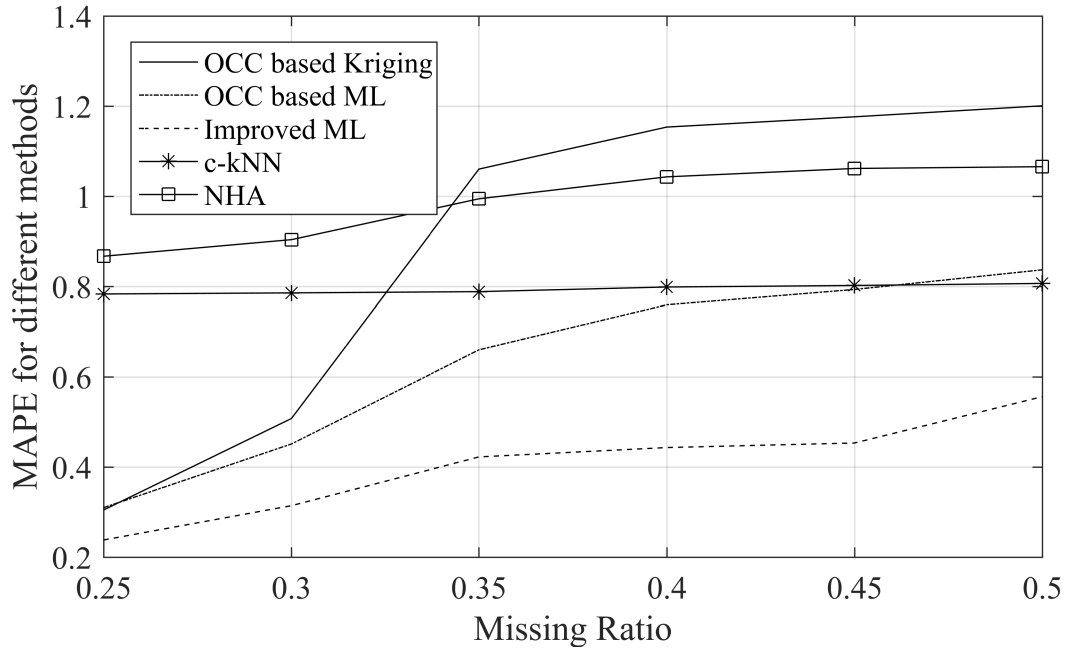

Figure 3.5: MAPE of five different methods in MGRT pattern

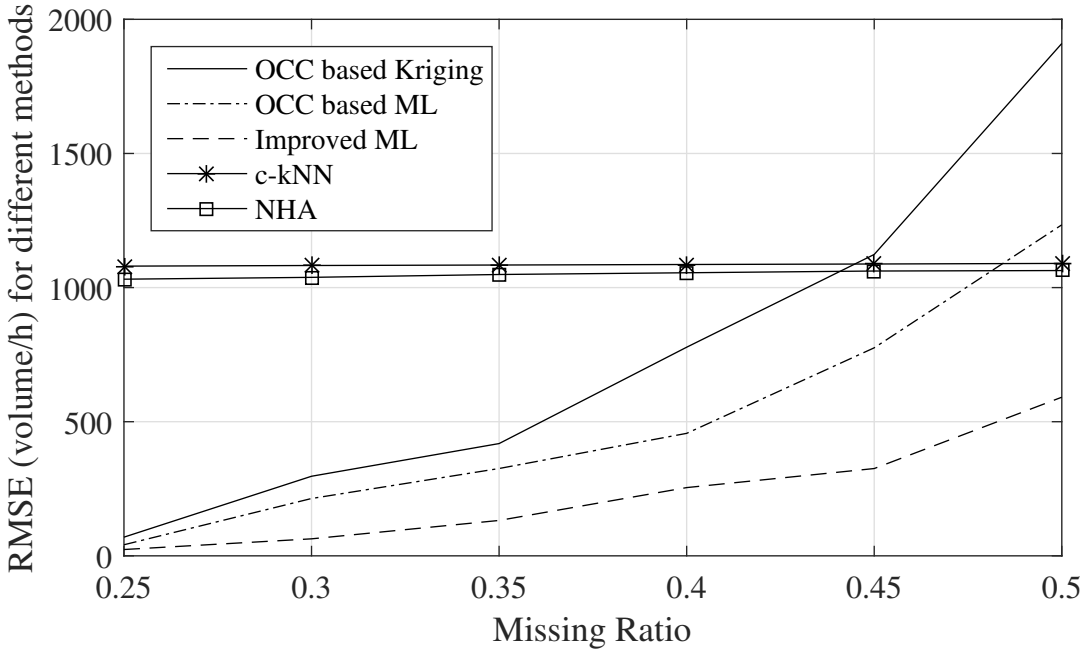

Figure 3.6: RMSE of five different methods in MGRT pattern

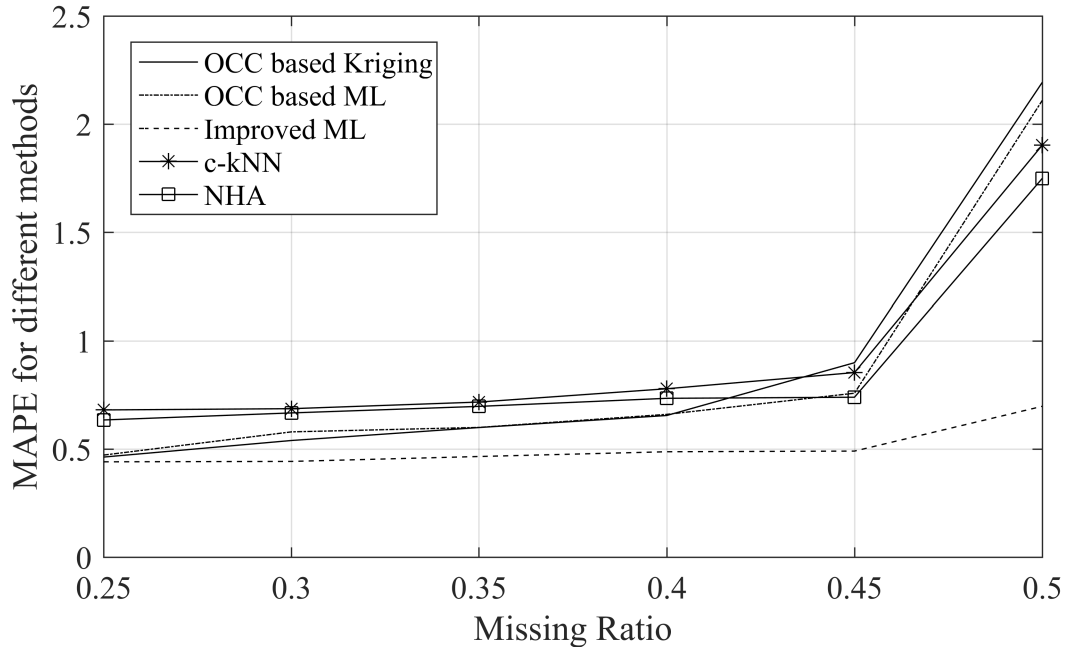

Figure 3.7: MAPE of five different methods in NMR pattern

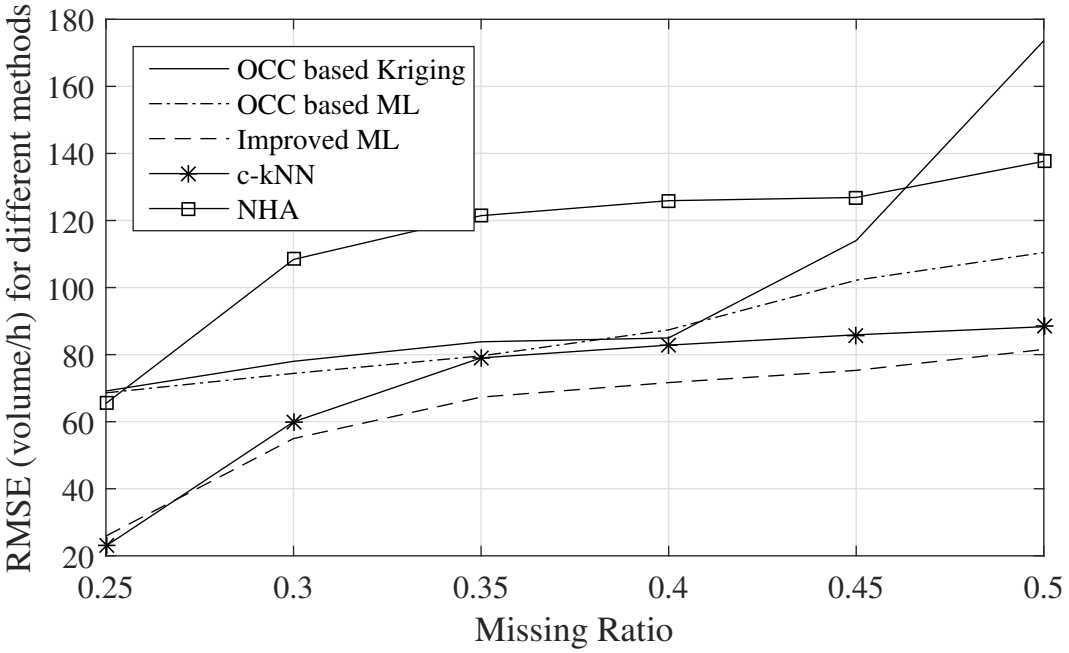

Figure 3.8: RMSE of five different methods in NMR pattern

1. The three novel methods outperforms NHA and kNN for three missing patterns over almost all missing ratios because the topological information was utilized and a sophisticated OCC finding algorithm was designed to determine the optimum sensors before imputation.

2. The two ML estimators can deliver a better estimation performance than OCC based Kriging because the flow conservation law has been incorporated.

3. By consideration of the dependence between the measured and unmeasured roads, the estimation accuracy can be further improved. Therefore, the improved ML estimator is the most appropriate imputation scheme for all missing patterns.

Chapter 4

Fundamental Limits of Missing Traffic Data Estimation

4.1 Overview

Chapter 3 proposed novel missing traffic data estimation strategies, which have been verified to be superior over two existing methods: NHA (Near Historical Average) and correlative kNN because the proposed strategies incorporates topological information. So far, the existing research on missing data estimation has mostly focused on using data-driven or model-driven models to estimate the missing data, there is a lack of study on the achievable estimation accuracy and the conditions to achieve accurate missing data estimation.

This chapter investigates the fundamental limits of missing traffic data estimation accuracy in urban networks using the spatio-temporal random effects (STRE) model. Firstly, the Squared Flow Error Bound (SFEB) is derived for the cases of the Fisher matrix being a singular and non-singular matrix, respectively. Then, the sufficient and necessary conditions of the existence of an unbiased estimator are shown that the number of missing points is less than or equal to the rank of the Fisher matrix. For the case that no unbiased estimator can be found, this chapter derives an inequality for the SFEB and shows that SFEB is readily determined by the co-variance matrix of the unknown (missing) parameter vector, flow correlation between the unknown and the available data, and the sensor locations. Finally, an optimal spatio-temporal Kriging estimator is developed.

4.2 System Model and Error Bound on Missing Data Estimation

This sub-chapter will firstly propose the urban traffic network model and the spatio-temporal random effects (STRE) model. Based on the proposed urban traffic network model, it will derive the SFEB of parameter vector estimation and show the relation among SFEB, the co-variance matrix of the real flow vector and the co-variance matrix of the observed flow vector. Furthermore, the relation between rank of the Fisher Information Matrix (FIM) and the number of missing points will be presented, and the sufficient and necessary conditions of existence of unbiased estimator will be given.

4.2.1 Network Model

Consider a two-dimensional traffic network with N_l roads, which are composed of N_m roads with sensors and N_u roads without sensors, $N_l = N_m + N_u$. At a time instant t, the traffic flow measured by the well-functioned sensors is represented as $q_{1,\mathrm{w},t}, q_{2,\mathrm{w},t} \cdots q_{M,\mathrm{w},t}$, while those flow over the rest $N_m - M$ roads with malfunctioned sensors and N_u unmeasured roads are represented as $q_{1,\mathrm{m,t}}, q_{2,\mathrm{m},t} \cdots q_{N_m-M,\mathrm{m},t}$ and $q_{1,\mathrm{u},t}, q_{2,\mathrm{u},t} \cdots q_{N_u,\mathrm{u},t}$, respectively. The letters "w", "m" and "u" denote well-functioned, malfunctioned and unmeasured roads, respectively. The data structure can be more clearly viewed by formulating the traffic flow on each road over consecutive time instants into a spatio-temporal matrix (see (4.3)). The task is to estimate the missing traffic data on the malfunctioned roads and unmeasured roads at each missing time instant. The set of missing data on all malfunctioned roads and unmeasured roads at all time instants is denoted by $\mathscr{N}_\mathrm{D} = \left\{ \begin{array}{cccc} 1, & 2, & \cdots & N_\mathrm{D} \end{array} \right\} \triangleq \mathscr{N}_\mathrm{M} \cup \mathscr{N}_\mathrm{U}$, where N_D is the total number of missing points on the malfunctioned roads and unmeasured roads, $\mathscr{N}_\mathrm{M}$ and $\mathscr{N}_\mathrm{U}$ denote the set of missing points on the malfunctioned roads and the unmeasured roads, respectively. In this chapter, the traffic process in urban networks is modeled with the spatio-temporal random effects (STRE) model, by which the observations can be represented as a sum of a deterministic function, small-scale variation, fine-scale variation caused by the nugget effect in geostatistics, and measurement error [45, 70], that is

$$Z(s;t) = \mu_t(\mathbf{s}) + S_t(s)^{'} \eta_t + \xi(s;t) + \varepsilon(s;t) \tag{4.1}$$

$$\eta_{t+1} = H_{t+1}\eta_t + \zeta_{t+1} \tag{4.2}$$

$$G = \left[\begin{array}{ccc|ccc|ccc} q_{1,\mathrm{w},1} & \cdots & q_{M,\mathrm{w},1} & q_{1,\mathrm{m},1} & \cdots & q_{N_m-M,\mathrm{m},1} & q_{1,\mathrm{u},1} & \cdots & q_{N_\mathrm{u},\mathrm{u},1} \\ q_{1,\mathrm{w},2} & \cdots & q_{M,\mathrm{w},2} & q_{1,\mathrm{m},2} & \cdots & q_{N_m-M,\mathrm{m},2} & q_{1,\mathrm{u},2} & \cdots & q_{N_\mathrm{u},\mathrm{u},2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q_{1,\mathrm{w},t} & \cdots & q_{M,\mathrm{w},t} & q_{1,\mathrm{m},t} & \cdots & q_{N_m-M,\mathrm{m},t} & q_{1,\mathrm{u},t} & \cdots & q_{N_\mathrm{u},\mathrm{u},t} \\ q_{1,\mathrm{w},t+1} & \cdots & q_{M,\mathrm{w},t+1} & q_{1,\mathrm{m},t+1} & \cdots & q_{N_m-M,\mathrm{m},t+1} & q_{1,\mathrm{u},t+1} & \cdots & q_{N_\mathrm{u},\mathrm{u},t+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q_{1,\mathrm{w},T} & \cdots & q_{M,\mathrm{w},T} & q_{1,\mathrm{m},T} & \cdots & q_{N_m-M,\mathrm{m},T} & q_{1,\mathrm{u},T} & \cdots & q_{N_\mathrm{u},\mathrm{u},T} \end{array}\right] \tag{4.3}$$

In a traffic network, $Z(s;t)$ represents the traffic flow on a finite number of road segments $s = \left\{ \begin{array}{ccc} s_1 & \cdots & s_N \end{array} \right\}'$ at the time instant of t, $\mu_t(s)$ represents the average flow on $\boldsymbol{s}$ at the t-th time instant, $S_t(s)'\eta_t$ can be interpreted as the flow fluctuation caused by the flow variation from L selected neighboring roads, $\xi(s;t)$ represents the variation on $\boldsymbol{s}$ at time instant t caused by nugget effect and flow generation or dissipation within some specific sections, $\varepsilon(s;t)$ is the measurement error on $\boldsymbol{s}$ at time instant t, $\xi(s;t)$ and $\varepsilon(s;t)$ can be modeled as independent white Gaussian process with mean zero and variances σ_ξ^2 and σ_ε^2, respectively, H_{t+1} is a first-order auto-regressive matrix and ζ_{t+1} is a innovation vector. In the next subsection, we will derive the *squared flow error bound* (SFEB) based on (4.1) and (4.2).

4.2.2 Error Bound on Missing Data Estimation

Based on (4.1), the traffic flow on each road segment can be expressed by a weighted sum of flows on its neighboring roads and independent random variations. Note that the neighboring roads not only include adjacent road segments, but also include the so-called l-th order, where l represents the spatial order of neighbors. For example, the first-order neighbors are those links that directly incident to the target road segments, while the second-order neighbors are indirectly connected to target road segments, via the first-order neighbors [22]. Yang et al. underlined that positive correlations exist among traffic collected at hundreds of sensors distributed on the entire road network sparsely, not just the neighborhood surrounding the target road segments [75].

Let us define an unknown parameter vector θ_1 that includes all missing datum on the $N_m - M$ malfunctioned roads, and an unknown parameter vector θ_2 that includes all unmeasured data points on N_u unmeasured roads, and a known parameter vector θ_3 that includes all measured data points on $N_m - M$ malfunctioned roads and M measured roads, i.e.,

$$\theta_1 = \left[\begin{array}{cccc} q_{1,\mathrm{m}} & q_{2,\mathrm{m}} & \cdots & q_{K_\mathrm{m},\mathrm{m}} \end{array} \right]^\mathrm{T}$$
$$\theta_2 = \left[\begin{array}{cccccccc} q_{1,\mathrm{u},1} & \cdots & q_{N_\mathrm{u},\mathrm{u},1} & \cdots & q_{1,\mathrm{u},\mathrm{T}} & \cdots & q_{N_\mathrm{u},\mathrm{u},\mathrm{T}} \end{array} \right]^\mathrm{T}$$
$$\theta_3 = \left[\begin{array}{cccccccc} q_{1,w,1} & \cdots & q_{M,w,1} & \cdots & q_{1,w,T} & \cdots & q_{M,w,T} & \\ q_{1,\mathrm{m},\mathrm{av}} & q_{2,\mathrm{m},\mathrm{av}} & \cdots & q_{K_\mathrm{av},\mathrm{m},\mathrm{av}} & & & & \end{array} \right]^\mathrm{T}$$

where K_m and K_av are the number of missing data points caused by temporary failure of detectors and available data points on the malfunctioned roads, respectively. Let us define R_1 and R_2 as the data missing ratio and the failure rate of detectors, respectively, then, the relation between R_1 and R_2, and K_m, K_av can be expressed by

$$\begin{aligned} R_1 &= \frac{N_m + (N_m - M)(1 - R_2)}{N_m + N_u} \\ K_\mathrm{m} &= (N_m - M) R_2 T \\ K_\mathrm{av} &= (N_m - M)(1 - R_2) T \end{aligned} \tag{4.4}$$

Then, the following theorem can be obtained:

Theorem 4.1. *An unbiased estimator of the parameter vector* $\theta = \left[\begin{array}{ccc} \theta_1^T & \theta_2^T & \theta_3^T \end{array} \right]^T$ *with finite variance exists iff (if only if)* $A^T R_\varepsilon^{-1} A + R_\theta^\dagger$ *is a full-rank matrix, where* A *is the scaling matrix and* R_ε *is the error co-variance matrix given by (4.8) and (4.9), respectively,* R_θ *is the co-variance matrix of the real flow vector,* $R_\theta^\dagger$ *is the Moore-Penrose pseudo-inverse of* R_θ*. Based on the given condition, the mean squared error (MSE) matrix of* $\hat{\theta}$ *by any estimators satisfies the following inequality*

$$\mathbb{E}_{q_o,\theta} \left\{ \left(\hat{\theta} - \theta\right) \left(\hat{\theta} - \theta\right)^H \right\} \succeq R_\theta - R_\theta A^T R_{q_o}^{-1} A R_\theta$$

where $M_1 \succeq M_2$ *means* $M_1 - M_2$ *is positive semi-definite,* R_{q_o} *is the co-variance matrix of the observed flow vector.*

Proof. For simplicity, we abbreviate the n-th road segment and the t-th time instant as the (n,t)-th point. Recall that equations (4.1) and (4.2) give us the information that in a traffic network, traffic flow on each (n,t)-th point can be represented as the summation of a mean value, variation and error term. The variation is caused by the variation on tis neighboring time instant and road segments, and evolves with time. By Kriging technique [17], the observed

traffic flow on the (n,t)-th point can be expressed by

$$\begin{aligned} q_{n,t} &= q_{n,t,\mathrm{r}} + \xi(n;t) + \varepsilon(n;t) \\ &= \bar{q}_{n,t,\mathrm{r}} + r_{n,t}^{\mathrm{T}} G_{n,t}^{-1} \left(\theta_{n,t,\mathrm{r}} - \bar{\theta}_{n,t,\mathrm{r}}\right) + \zeta(n;t) + \xi(n;t) + \varepsilon(n;t) \end{aligned} \tag{4.5}$$

where $q_{n,t,\mathrm{r}}$ and $\bar{q}_{n,t,\mathrm{r}}$ are the (n,t)-th instantaneous flow and average flow, respectively, $r_{n,t}$ is a $[(N_m+N_u)T-1]\times 1$ column vector containing the empirical cross correlation between the (n,t)-th data point and the rest spatial-temporal data points $(n',t') = (1\cdots N_m+N_u, 1\cdots T), (n',t') \neq (n,t)$, $G_{n,t}$ is a $[(N_m+N_u)T-1]\times[(N_m+N_u)T-1]$ empirical co-variance matrix consisting of co-variance between each pair of spatial-temporal data points in the rest spatial-temporal data set, $\theta_{n,t,\mathrm{r}}$ is a $[(N_m+N_u)T-1]\times 1$ column vector containing instantaneous flow at all unmeasured data points and measured data points, $\bar{\theta}_{n,t,\mathrm{r}}$ is the $[(N_m+N_u)T-1]\times 1$ average real flow vector, $\zeta(n;t)$ is the flow variation introduced by measurement error on rest spatial-temporal data points and flow generation or dissipation within some specific sections, $\xi(n;t)$ is the flow variation caused by nugget effect and $\varepsilon(n;t)$ is the measurement error for the (n,t)-th observation. For simplicity, we define a new $[MT+K_{\mathrm{av}}-1]\times 1$ column vector $\theta_3^{(n,t)}$ obtained by removing the (n,t)-th element from θ_3. Then, $r_{n,t}$, $G_{n,t}$ and $\theta_{n,t}$ can be expressed by

$$\theta_{n,t} = \begin{bmatrix} \theta_1^{\mathrm{T}} & \theta_2^{\mathrm{T}} & \theta_3^{(n,t)\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$$

$$\begin{aligned} r_{n,t} &= \mathbb{E}\left[(q_{n,t}-\bar{q}_{n,t})\left(\theta_{n,t}-\bar{\theta}_{n,t}\right)\right] = \mathbb{E}\left[(\tilde{q}_{n,t}+\varepsilon_{n,t})\begin{pmatrix} \tilde{\theta}_1+\xi_1+\varepsilon_1 \\ \tilde{\theta}_2+\xi_2+\varepsilon_2 \\ \tilde{\theta}_3^{(n,t)}+\xi_3+\varepsilon_3^{(n,t)} \end{pmatrix}\right] \\ &= \begin{bmatrix} r_{n,t,\theta_1}^{\mathrm{T}} & r_{n,t,\theta_2}^{\mathrm{T}} & r_{n,t,\theta_3^{(n,t)}}^{\mathrm{T}} \end{bmatrix}^{\mathbf{T}} \end{aligned}$$

$$\begin{aligned} G_{n,t} &= \mathbb{E}\left[\left(\theta_{n,t}-\bar{\theta}_{n,t}\right)\left(\theta_{n,t}-\bar{\theta}_{n,t}\right)^{\mathrm{T}}\right] = \mathbb{E}\left[\begin{pmatrix} \tilde{\theta}_1+\xi_1+\varepsilon_1 \\ \tilde{\theta}_2+\xi_2+\varepsilon_2 \\ \tilde{\theta}_3^{(n,t)}+\xi_3+\varepsilon_3^{(n,t)} \end{pmatrix}\begin{pmatrix} \tilde{\theta}_1+\xi_1+\varepsilon_1 \\ \tilde{\theta}_2+\xi_2+\varepsilon_2 \\ \tilde{\theta}_3^{(n,t)}+\xi_3+\varepsilon_3^{(n,t)} \end{pmatrix}^{\mathrm{T}}\right] \\ &= R_{\theta_{n,t,r}} + \left(\sigma_\xi^2+\sigma_\varepsilon^2\right)I \end{aligned}$$

where $r_{n,t,\theta}$ is the flow correlation between the (n,t)-th data point and θ, $R_{\theta_k\theta_l}$ is the co-

variance matrix between θ_k and θ_l, σ_ξ^2 is variance of nugget effect and flow generation or dissipation, σ_ε^2 is variance of the measurement error, $\tilde{q}_{n,t}$ is the flow variation at the (n,t)-th data point, $\tilde{\theta}_1$,$\tilde{\theta}_2$ and $\tilde{\theta}_3^{(n,t)}$ are the flow variation vectors for θ_1,θ_2 and $\theta_3^{(n,t)}$, respectively.

Note that the co-variance with the data vector on the malfunctioned road segments can be determined by historical data because of the temporary failure of detectors. The co-variance with the data vector on the unmeasured road segments can be determined by historical data measured by temporarily installed detectors on those road segments. In the case that some unmeasured road segments have no temporary detectors installed previously, θ_2 can be considered as a latent parameter vector, and $\theta_{n,t}$ becomes $\left[\begin{array}{cc} \theta_1^{\mathrm{T}} & \theta_3^{(n,t)\mathrm{T}} \end{array}\right]^{\mathrm{T}}$.

The variance of the flow variation $\zeta(n;t)$ can be determined by

$$\begin{aligned} var(\zeta(n;t)) &= var(q_{n,t,\mathrm{r}}) - r_{n,t}^{\mathrm{T}} G_{n,t}^{-1} R_{\theta_{n,t,r}} G_{n,t}^{-\mathrm{T}} r_{n,t} \\ &= var(q_{n,t,\mathrm{r}}) - r_{n,t}^{\mathrm{T}} U_{\theta_{n,t,r}} \Lambda_{\theta_{n,t,r}} U_{\theta_{n,t,r}}^{\mathrm{H}} r_{n,t} \end{aligned}$$

where $R_{\theta_{n,t,r}}$ is the co-variance matrix of the real instantaneous flow vector $\theta_{n,t}$, $\Lambda_{\theta_{n,t,r}}$contains non-zero Eigenvalues of $\frac{\lambda_k}{\lambda_k+\sigma_\varepsilon^2}, k = 1\cdots K_{n,t}$ in its diagonal, λ_k is the non-zero Eigenvalues of $R_{\theta_{n,t,r}}$, $U_{\theta_{n,t,r}}$ is the Eigenspace of $R_{\theta_{n,t,r}}$ dedicated to $\lambda_k, k = 1\cdots K_{n,t}$. Recall that $\xi(n;t)$ and $\varepsilon(n;t)$ are independent white Gaussian process in space and time with mean zero and variance σ_ξ^2 and σ_ε^2, then (4.5) can be rewritten as

$$\begin{aligned} q_{n,t} &= \bar{q}_{n,t,\mathrm{r}} + r_{n,t}^{\mathrm{T}} G_{n,t}^{-1} \left(\theta_{n,t,\mathrm{r}} - \bar{\theta}_{n,t,\mathrm{r}}\right) + \varepsilon_{n,t} \\ &= r_{n,t}^{\mathrm{T}} G_{n,t}^{-1} \theta_{n,t,\mathrm{r}} + \bar{q}_{n,t,\mathrm{r}} - r_{n,t}^{\mathrm{T}} G_{n,t}^{-1} \bar{\theta}_{n,t,\mathrm{r}} + \varepsilon_{n,t} \end{aligned} \tag{4.6}$$

where $\varepsilon_{n,t}$ is an i.i.d zero-mean Gaussian process with variance of $var(q_{n,\mathrm{r}}) - r_{n,t}^{\mathrm{T}} U_{\theta_{n,t,r}} \Lambda_{\theta_{n,t,r}} U_{\theta_{n,t,r}}^{\mathrm{H}} r_{n,t} + \sigma_\xi^2 + \sigma_\varepsilon^2$. Formulating $MT + K_{\mathrm{av}}$ observations into a system of linear equations (SLE), which can be expressed by

$$q_o = A\theta + u + \varepsilon \tag{4.7}$$

where q_o is a $[MT + K_{\mathrm{av}}] \times 1$ observation flow vector, θ is a $(N_m + N_u)T \times 1$ instantaneous real traffic flow vector containing flow on each road segment at T time instants, A is a $[MT + K_{\mathrm{av}}] \times (N_m + N_u)T$ scaling matrix obtained by $r_{n,t}^{\mathrm{T}} G_{n,t}^{-1}, n = 1\cdots N_m + N_u, t = 1\cdots T$, u is a $[MT + K_{\mathrm{av}}] \times 1$ bias vector and ε is the measurement error vector, that is

$$A = \begin{bmatrix} A_i & A_j & A_{\ell 1} \end{bmatrix}$$

$$A_i = \left[\underbrace{\left[G_1^{-\mathrm{T}}\right]_1^{K_\mathrm{m}} r_1 \quad \cdots \quad \left[G_{MT}^{-\mathrm{T}}\right]_1^{K_\mathrm{m}} r_{MT}}_{MT} \cdots \underbrace{\begin{matrix}\left[G_{MT+1}^{-\mathrm{T}}\right]_1^{K_\mathrm{m}} \\ \times r_{MT+1}\end{matrix} \quad \cdots \quad \begin{matrix}\left[G_{MT+K_\mathrm{av}}^{-\mathrm{T}}\right]_1^{K_\mathrm{m}} \\ \times r_{MT+K_\mathrm{av}}\end{matrix}}_{K_\mathrm{av}} \right]^{\mathrm{T}}$$

$$A_j = \left[\underbrace{\begin{matrix}\left[G_1^{-\mathrm{T}}\right]_{K_\mathrm{m}+1}^{K_\mathrm{m}+N_\mathrm{u}T} \\ \times r_1\end{matrix} \quad \cdots \quad \begin{matrix}\left[G_{MT}^{-\mathrm{T}}\right]_{K_\mathrm{m}+1}^{K_\mathrm{m}+N_\mathrm{u}T} \\ \times r_{MT}\end{matrix}}_{MT} \cdots \underbrace{\begin{matrix}\left[G_{MT+1}^{-\mathrm{T}}\right]_{K_\mathrm{m}+1}^{K_\mathrm{m}+N_\mathrm{u}T} \\ \times r_{MT+1}\end{matrix} \quad \cdots \quad \begin{matrix}\left[G_{MT+K_\mathrm{av}}^{-\mathrm{T}}\right]_{K_\mathrm{m}+1}^{K_\mathrm{m}+N_\mathrm{u}T} \\ \times r_{MT+K_\mathrm{av}}\end{matrix}}_{K_\mathrm{av}} \right]^{\mathrm{T}}$$

$$[A_{\ell 1}]_{k=1}^{MT+K_\mathrm{av}} = \begin{bmatrix} \left[G_k^{-\mathrm{T}}\right]_{K_\mathrm{m}+N_\mathrm{u}T+1}^{K_\mathrm{m}+N_\mathrm{u}T+k-1} r_k \\ 0 \\ \left[G_k^{-\mathrm{T}}\right]_{K_\mathrm{m}+N_u T+k}^{K_\mathrm{m}+N_\mathrm{u}T+MT+K_\mathrm{av}-1} r_k \end{bmatrix}^{\mathrm{T}} \tag{4.8}$$

$$\begin{aligned} \theta &= \begin{bmatrix} \theta_{1,\mathrm{r}}^{\mathrm{T}} & \theta_{2,\mathrm{r}}^{\mathrm{T}} & \theta_{3,\mathrm{r}}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, A^{\mathrm{T}} R_\varepsilon^{-1} A + R_\theta^{\dagger} u = \bar{q}_\mathrm{r} - A\bar{\theta} \\ \varepsilon &= \begin{bmatrix} \varepsilon_1 & \cdots & \varepsilon_{MT+K_\mathrm{av}} \end{bmatrix}^{\mathrm{T}}, \bar{q}_\mathrm{r} = \begin{bmatrix} \bar{q}_{1,\mathrm{r}} & \cdots & \bar{q}_{MT+K_\mathrm{av},\mathrm{r}} \end{bmatrix}^{\mathrm{T}} \end{aligned} \tag{4.9}$$

where $[G]_{k1}^{k2}$ denotes the sub-matrix that extracts $[k1 \sim k2]$-th rows of G, $\bar{\theta}$ is the real average traffic flow vector, R_ε is $[MT + K_\mathrm{av}] \times [MT + K_\mathrm{av}]$ diagonal matrix with the n-th element on its diagonal $var\left(q_{n,t,\mathrm{r}}\right) - r_{n,t}^{\mathrm{T}} U_{\theta_{n,t,r}} \Lambda_{\theta_{n,t,r}} U_{\theta_{n,t,r}}^{\mathrm{H}} r_{n,t} + \sigma_\xi^2 + \sigma_\varepsilon^2$.

Let $\hat{\theta}$ denote an estimate of the vector θ based on observation q_o. Then, the mean squared error (MSE) matrix of $\hat{\theta}$ satisfies the information inequality [61]

$$\mathbb{E}_{q_o,\theta}\left\{\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)^{\mathrm{H}}\right\} \succeq \mathrm{J}_{\theta,\mathrm{r}}^{-1} \tag{4.10}$$

$$\mathbb{E}_{q_o,\theta}\left\{\parallel \hat{\theta}-\theta \parallel^2\right\} \geq \mathrm{tr}\left(\mathrm{J}_{\theta,\mathrm{r}}^{-1}\right) \tag{4.11}$$

where $\mathrm{J}_{\theta,\mathrm{r}}$ is the Fisher Information Matrix (FIM) for the parameter vector θ, $A \succeq B$ denotes that the matrix $A - B$ is a positive semi-definite matrix, $\parallel \cdot \parallel$ is the Euclidean norm of its argument, $\mathrm{tr}\{\cdot\}$is the trace operation. Note that the parameter vector θ is a random parameter and $\mathrm{J}_{\theta,\mathrm{r}}$ is a Bayesian information matrix that does not depend on any particular θ, but requires

an average over all possible θ.

Initially, we will give the FIM for deterministic parameter vector θ, then the PDF of parameters is incorporated to derive the FIM for random parameter vector. Let us define $f(q_o|\theta)$ as the likelihood ratio of the observation vector q_o conditioned on θ, then the FIM for a deterministic parameter vector θ is given by [61]

$$\mathrm{J}_{\theta,\mathrm{d}} \triangleq \mathbb{E}_{q_o}\left\{\left[\frac{\partial}{\partial\theta}\ln f(q_o|\theta)\right]\left[\frac{\partial}{\partial\theta}\ln f(q_o|\theta)\right]^{\mathrm{T}}\right\} \tag{4.12}$$

From (4.7) to (4.9), the likelihood function can be expressed by

$$\begin{aligned} f(q_o|\theta) &= \frac{\exp\left\{-\frac{1}{2}(q_o - A\theta - u)^{\mathrm{T}} R_\varepsilon^{-1}(q_o - A\theta - u)\right\}}{(2\pi)^{\frac{(MT+K_{\mathrm{av}})}{2}}\det^{\frac{1}{2}}(R_\varepsilon)} \\ &\propto \exp\left\{-\frac{1}{2}\left(q_o^{\mathrm{T}} - \theta^{\mathrm{T}}A^{\mathrm{T}} - u^{\mathrm{T}}\right)R_\varepsilon^{-1}(q_o - A\theta - u)\right\} \end{aligned} \tag{4.13}$$

Substituting (4.13) in (4.6), the FIM $\mathrm{J}_{\theta,\mathrm{d}}$ can be expressed by

$$\begin{aligned} \mathrm{J}_{\theta,\mathrm{d}} &= -\mathbb{E}_{q_o}\left\{-\frac{\partial^2\left\{\begin{array}{c}(q_o^{\mathrm{T}} - \theta^{\mathrm{T}}A^{\mathrm{T}} - u^{\mathrm{T}}) \\ \times R_\varepsilon^{-1}(q_o - A\theta - u)\end{array}\right\}}{2\partial\theta^2}\right\} \\ &= A^{\mathrm{T}}R_\varepsilon^{-1}A = \begin{bmatrix} A_i^{\mathrm{T}}R_\varepsilon^{-1}A_i & A_i^{\mathrm{T}}R_\varepsilon^{-1}A_j & A_i^{\mathrm{T}}R_\varepsilon^{-1}A_{\ell 1} \\ A_j^{\mathrm{T}}R_\varepsilon^{-1}A_i & A_j^{\mathrm{T}}R_\varepsilon^{-1}A_j & A_j^{\mathrm{T}}R_\varepsilon^{-1}A_{\ell 1} \\ A_{\ell 1}^{\mathrm{T}}R_\varepsilon^{-1}A_i & A_{\ell 1}^{\mathrm{T}}R_\varepsilon^{-1}A_j & A_{\ell 1}^{\mathrm{T}}R_\varepsilon^{-1}A_{\ell 1} \end{bmatrix} \end{aligned} \tag{4.14}$$

Then, a lower bound of co-variance matrix of $\hat{\theta}$ can be straightly obtained by (4.10) and (4.11) when $A^{\mathrm{T}}R_\varepsilon^{-1}A$ is a full-rank matrix, that is $MT + K_{\mathrm{av}} = (N_m + N_u)T$. Unfortunately, $A^{\mathrm{T}}R_\varepsilon^{-1}A$ is a singular matrix because A is a fat matrix. Thus, an unbiased estimation of the entire parameter vector is impossible to be found because some components in the parameter vector may have infinite variance [54]. In such scenario, (4.10) can be rewritten as

$$\mathbb{E}_{q_o,\theta}\left\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^{\mathrm{H}}\right\} \succeq \mathrm{J}_{\theta,\mathrm{d}}^{\dagger}$$

where $\mathrm{J}_{\theta,\mathrm{d}}^{\dagger}$ is the Moore-Penrose pseudoinverse of $\mathrm{J}_{\theta,\mathrm{d}}$. There are three approaches coping with the case of the FIM being a singular matrix, that is, incorporating the PDF of parame-

ter vector to covert the original deterministic parameter estimation problem into a Bayesian estimation problem; posing deterministic constrains on the parameter vector to reduce the parameter dimension; using a biased estimator [54]. This thesis focuses on incorporating the a priori knowledge of the parameter vector.

Let us define $f(\theta)$ as the PDF of parameter vector θ, then the joint PDF of observation and parameter vector can be expressed by

$$f(q_o, \theta) = f(q_o|\theta) f(\theta)$$

where $f(q_o|\theta)$ is given by (4.13), and thus the FIM becomes

$$\mathrm{J}_{\theta,\mathrm{r}} = \mathbb{E}_\theta \left(\mathrm{J}_{\theta,\mathrm{d}} + \mathrm{J}_\theta \right)$$

where J_θ is the FIM for the a priori knowledge of the parameter vector and can be expressed by

$$\begin{aligned} \mathrm{J}_\theta &= \mathbb{E}_\theta \left\{ \left[\frac{\partial}{\partial \theta} \ln f(\theta) \right] \left[\frac{\partial}{\partial \theta} \ln f(\theta) \right]^{\mathrm{T}} \right\} \\ &= -\mathbb{E}_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \ln f(\theta) \right\} \end{aligned}$$

Let us define R_θ as the co-variance matrix of the real flow vector, then the PDF of θ can be expressed by

$$f(\theta) \propto \exp \left\{ -\frac{1}{2} (\theta - \bar{\theta}) R_\theta^\dagger (\theta - \bar{\theta})^{\mathrm{T}} \right\}$$

where $\bar{\theta}$ has the same definition as (4.9), $R_\theta^\dagger$ is the Moore-Penrose pseudoinverse of R_θ. Note that R_θ has a possibility of being a singular matrix because of the spatial-temporal correlation between traffic at different road segments and time instants. The rank of R_θ intuitively represents the number of "sources" in the traffic network. Hence, the FIM for the a priori knowledge of the parameter vector can be obtained by

$$\mathrm{J}_\theta = R_\theta^\dagger$$

and the FIM for the joint PDF of observation and parameter vector can be expressed by

$$\mathrm{J}_{\theta,\mathrm{r}} = A^{\mathrm{T}} R_\varepsilon^{-1} A + R_\theta^\dagger \tag{4.15}$$

where $A^{\mathrm{T}}R_{\varepsilon}^{-1}A$ and $R_{\theta}^{\dagger}$ are of the rank of $MT+K_{\mathrm{av}}$ and r, respectively. The rank of $\mathrm{J}_{\theta,\mathrm{r}}$ has the following relation

$$\begin{aligned}\mathrm{R}\left(\mathrm{J}_{\theta,\mathrm{r}}\right) &\leq \mathrm{R}\left(A^{\mathrm{T}}R_{\varepsilon}^{-1}A\right)+\mathrm{R}\left(R_{\theta}^{\dagger}\right)\\ &= MT+K_{\mathrm{av}}+r\\ &= MT+\left(N_m-M\right)\left(1-R_2\right)T+r\end{aligned}$$

Thus, the necessary condition that an unbiased estimator of the entire parameter vector with finite variance exists is that $MT+\left(N_m-M\right)\left(1-R_2\right)T+r\geq\left(N_m+N_u\right)T$. In the following context, the lower error bound will be derived by assuming that $\mathrm{J}_{\theta,\mathrm{r}}$ is a full-rank matrix and a singular matrix, respectively. The former means that there exists an unbiased estimator with finite variance for the entire parameter vector θ and the latter means that no unbiased estimator can be found and thus parameter transformation or biased estimator should be applied. Determining the error bound requires inverting the FIM $\mathrm{J}_{\theta,\mathrm{r}}$ in (4.15), which can be rewritten as

$$\mathrm{J}_{\theta,\mathrm{r}}=A^{\mathrm{T}}R_{\varepsilon}^{-1}A+R_{\theta}^{\dagger}$$

With the Woodbury matrix identity, the inverse of $\mathrm{J}_{\theta,\mathrm{r}}$ can be expressed by

$$\begin{aligned}\mathrm{J}_{\theta,\mathrm{r}}^{-1} &= \left(R_{\theta}^{\dagger}+A^{\mathrm{T}}R_{\varepsilon}^{-1}A\right)^{-1}\\ &= R_{\theta}-R_{\theta}A^{\mathrm{T}}\left(R_{\varepsilon}+AR_{\theta}A^{\mathrm{T}}\right)^{-1}AR_{\theta}\\ &= R_{\theta}-R_{\theta}A^{\mathrm{T}}R_{q_o}^{-1}AR_{\theta}\end{aligned}\tag{4.16}$$

From (4.11), the mean squared error can be obtained from the trace of $\mathrm{J}_{\theta,\mathrm{r}}^{-1}$, that is

$$\mathbb{E}_{q_o,\theta}\left\{\parallel\hat{\theta}-\theta\parallel^2\right\}\geq\mathrm{tr}\left(R_{\theta}\right)-\mathrm{tr}\left(R_{\theta}A^{\mathrm{T}}R_{q_o}^{-1}AR_{\theta}\right)\tag{4.17}$$

Note that A in (4.11) is given by (4.8), which is determined by the flow correlation between available data and the rest parameter vector (partial available data and entire unknown data) $r_{n,t}$, and the co-variance matrix of the rest parameter vector $G_{n,t}$. In the case of appropriate selection of sensor locations so that there is no causal relationship between each pair of available data points, A can be directly expressed by a product of cross correlation matrix between the available data and the unknown parameter vector, and inverse matrix of co-variance matrix

of the unknown parameter vector. Then, (4.17) can be rewritten as

$$\mathbb{E}_{q_o,\theta}\left\{\| \hat{\theta}_u - \theta_u \|^2\right\} \geq \mathrm{tr}\left(R_{\theta_u}\right) - \mathrm{tr}\left(R_{\theta_u,q_o} R_{q_o}^{-1} R_{\theta_u,q_o}^{\mathrm{H}}\right) \tag{4.18}$$

where R_{θ_u} and R_{θ_u,q_o} are the co-variance matrix of the unknown parameter vector and cross correlation matrix between the available data and the unknown parameter vector, respectively. Then, Theorem 4.1 is proved. □

When $\mathrm{J}_{\theta,\mathrm{r}}$ is a singular matrix, the following corollary can be obtained:

Corollary 4.2. *In the case that $J_{\theta,r}$ is a singular matrix, an unbiased estimator of the unknown parameter vector θ_u exists iff (if only if) the number of missing points is less than or equal to the rank of $J_{\theta,r}$ i.e. $(N_m+N_u)T - K_{av} - MT \leq r_J$, where r_J is the rank of $J_{\theta,r}$. Based on the given condition, the variance of $\hat{\theta}_u$ satisfies the following inequality*

$$var_\theta\left(\hat{\theta}_u\right) \succeq \begin{bmatrix} I & 0 \end{bmatrix} U_1 \Lambda_J^{-1} U_1^H \begin{bmatrix} I \\ 0 \end{bmatrix}$$

where I is the identity matrix, Λ_J and U_1 are the non-zero eigenvalues and the corresponding eigenspace, respectively. When $(N_m+N_u)T - K_{av} - MT > r_J$, no unbiased estimator can be found and the variance of $\hat{\theta}_u$ satisfies the following inequality

$$H_u var_\theta\left(\hat{\theta}_u\right) H_u^H \succeq H U_1 \Lambda_J^{-1} U_1^H H^{\boldsymbol{H}} + \left(\sigma_\xi^2 + \sigma_\varepsilon^2\right)\left(H_a H_a^H - 2HA^\dagger A^{\dagger H} H^{\boldsymbol{H}}\right)$$

where H is an arbitrary scaling matrix whose row space is orthogonal to the eigenspace dedicated to the zero eigenvalues of $J_{\theta,r}$, H_u and H_a are the sub-matrices of H corresponding to θ_u and θ_a, respectively.

Proof. From [28], the general form of CRLB can be expressed by

$$\begin{aligned}\mathbb{E}_{q_o,\theta}\left\{\begin{array}{c}(\hat{\alpha}-E(\hat{\alpha}))\times \\ (\hat{\alpha}-E(\hat{\alpha}))^{\mathrm{H}}\end{array}\right\} &\succeq \frac{\partial E(\hat{\alpha})}{\partial\theta} \mathrm{J}_{\theta,\mathrm{r}}^{\dagger} \left(\frac{\partial E(\hat{\alpha})}{\partial\theta}\right)^{\mathrm{H}} \\ &= \frac{\partial E(\hat{\alpha})}{\partial\theta}\left(R_\theta - R_\theta A^{\mathrm{T}} R_{q_o}^{-1} A R_\theta\right)\left(\frac{\partial E(\hat{\alpha})}{\partial\theta}\right)^{\mathrm{H}}\end{aligned} \tag{4.19}$$

where $\alpha = h(\theta)$ and $\frac{\partial E(\hat{\alpha})}{\partial\theta}$ is a $(N_m+N_u)T \times (N_m+N_u)T$ matrix. To guarantee the right side of (4.19) has no unbounded eigenvalue on its diagonal, the row space of $\frac{\partial E(\hat{\alpha})}{\partial\theta}$ must be orthogonal to the eigenspace dedicated to the zero eigenvalues of $\mathrm{J}_{\theta,\mathrm{r}}$. Assume that h is a

linear function and α can be expressed by $\alpha = H\theta$. By unbiased estimation of α, (4.19) can be rewritten as

$$\mathbb{E}_{q_o,\theta}\left\{[\hat{\alpha} - E(\hat{\alpha})][\hat{\alpha} - E(\hat{\alpha})]^{\mathrm{H}}\right\} \succeq H\mathrm{J}^{\dagger}_{\theta,\mathrm{r}}H^{\mathrm{H}} \tag{4.20}$$

Let U_2 be the eigenspace dedicated to zero eigenvalues of $\mathrm{J}_{\theta,\mathrm{r}}$, then a sufficient condition that $\hat{\alpha}$ is a variance-limited unbiased estimator is $HU_2 = 0$. By utilizing the eigenvector/eigenvalue representation of $\mathrm{J}_{\theta,\mathrm{r}}$

$$\mathrm{J}_{\theta,\mathrm{r}} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^{\mathrm{H}} \\ U_2^{\mathrm{H}} \end{bmatrix}$$

the right side of (4.19) can be rewritten as $HU_1\Lambda_1^{-1}U_1^{\mathrm{H}}H^{\mathbf{H}}$. By an unbiased estimator $\hat{\alpha}$, it has

$$var_{\theta}(\hat{\alpha}) \succeq HU_1\Lambda_1^{-1}U_1^{\mathrm{H}}H^{\mathbf{H}}$$

and

$$\begin{bmatrix} H_u & H_a \end{bmatrix} \begin{bmatrix} \hat{\theta}_u \\ \hat{\theta}_a \end{bmatrix} = \hat{\alpha} \tag{4.21}$$

where $\hat{\theta}_u$ and $\hat{\theta}_a$ are the $(N_m + N_u)T - K_{av} - MT \times 1$ and $K_{av} + MT \times 1$ vectors representing the estimated unknown and available parameters, respectively. Let us define the rank of $\mathrm{J}_{\theta,\mathrm{r}}$ as r_J. In the case that $(N_m + N_u)T - K_{av} - MT \leq r_J$, the unbiased estimation of $\hat{\theta}_u$ can be obtained by

$$\hat{\theta}_u = \left(H_u^{\mathrm{H}}H_u\right)^{-1}H_u^{\mathrm{H}}[\hat{\alpha} - H_a(\theta_a + \xi + \varepsilon)]$$

Then,

$$var_{\theta}\left(\hat{\theta}_u\right) \succeq \begin{bmatrix} I & 0 \end{bmatrix} U_1\Lambda_1^{-1}U_1^{\mathrm{H}} \begin{bmatrix} I \\ 0 \end{bmatrix} = \left(H_u^{\mathrm{H}}H_u\right)^{-1}H_u^{\mathrm{H}}HU_1\Lambda_1^{-1}U_1^{\mathrm{H}}H^{\mathbf{H}}H_u\left(H_u^{\mathrm{H}}H_u\right)^{-1} \tag{4.22}$$

In the case that $(N_m + N_u)T - K_{av} - MT > r_J$, (4.22) does not hold because H_u is no longer full-rank. Equation (4.22) can be transformed into

$$H_u\hat{\theta}_u = \hat{\alpha} - H_a(\theta_a + \xi + \varepsilon)$$

Then,

$$H_u var_\theta\left(\hat{\theta}_u\right) H_u^{\mathrm{H}} \succeq H U_1 \Lambda_1^{-1} U_1^{\mathrm{H}} H^{\mathbf{H}} + \left(\sigma_\xi^2 + \sigma_\varepsilon^2\right) H_a H_a^{\mathrm{H}} - 2\mathrm{cov}[\hat{\alpha}, H_a(\xi + \varepsilon)] \tag{4.23}$$

Note that $\mathrm{cov}(\hat{\alpha}, H_a(\xi+\varepsilon))$ depends on H_a and $\hat{\alpha}$, i.e., the applied estimator. It is well-known that a maximum likelihood (ML) estimator is asymptotically efficient [28]. By introducing $\alpha = H\theta$ to (4.7), it can be transformed to

$$q_o = AH^\dagger \alpha + u + \varepsilon$$

Then,

$$\begin{aligned}\hat{\alpha}_{\mathrm{ml}} &= \left(AH^\dagger\right)^\dagger (q_o - u) \\ &= \alpha + HA^\dagger \varepsilon\end{aligned}$$

and (4.23) can be rewritten as

$$H_u var_\theta\left(\hat{\theta}_u\right) H_u^{\mathrm{H}} \succeq H U_1 \Lambda_1^{-1} U_1^{\mathrm{H}} H^{\mathbf{H}} + \left(\sigma_\xi^2 + \sigma_\varepsilon^2\right)\left(H_a H_a^{\mathrm{H}} - 2HA^\dagger A^{\dagger\mathrm{H}} H^{\mathbf{H}}\right) \tag{4.24}$$

Then, Corollary 4.2 is proved. □

Note that Theorem 4.1 and Corollary 4.2 indicate that the minimum mean squared error can be readily determined by covariance of the unknown parameter vector, flow correlation between the unknown parameter vector and the available data, and the sensor locations. Thus, Theorem 4.1 and Corollary 4.2 can also be used to optimize the sensor locations in a large traffic network when the number of available sensors is limited.

4.3 Optimal and Iterated Spatial-Temporal Kriging and Performance Analysis

In this sub-chapter, an optimal spatio-temporal Kriging estimator is proposed and its error analysis is conducted based on the data structure in (4.3). For simplicity of expression, we formulate the unknown parameters θ_1 and θ_2 into a new parameter vector θ_u, and the observed parameters into q_o. It is often the case that the mean flow value is not constant over space and

time in a traffic network. Thus, a simple spatio-temproal Kriging estimator proposed in [17] is employed, which can be expressed by

$$\begin{aligned}\hat{\theta}_u &= \mathbb{E}\left(\theta_u | q_o\right) \\ &= \bar{\theta}_u + R_{\theta_u,q_o} R_{q_o}^{-1}\left(q_o - \bar{\theta}_o\right) \qquad (4.25) \\ &= \bar{\theta}_u + R_{\theta_u,q_o}\left[\ R_{\theta_o} + \left(\sigma_\xi^2 + \sigma_\varepsilon^2\right) I\ \right]^{-1}\left[\ \theta_o - \bar{\theta}_o + \xi + \varepsilon\ \right] \qquad (4.26)\end{aligned}$$

where $\bar{\theta}_u$ and $\bar{\theta}_o$ are the $(N_m + N_u)T - K_{av} - MT \times 1$ and $K_{av} + MT \times 1$ average flow vectors for the unknown parameters and the observed parameters, respectively, R_{θ_u,q_o} and R_{q_o} are the co-variance matrix between θ_u and q_o, and auto correlation matrix of q_o, respectively.

Remark 4.3. It is assumed that the number of samples is sufficiently large so that the perturbations can be neglected. In practice, $\bar{\theta}_u$, R_{θ_u,q_o}, R_{q_o} and $\bar{\theta}_o$ are perturbed by small drifts due to the possibility of insufficient statistics of the empirical data.

Equation (4.25) shows that spatial-temporal Kriging estimator is biased towards $\bar{\theta}_u$, rather than θ_u if the dimension of q_o is not sufficiently large. The mean squared error (MSE) of (4.25) is given by

$$\begin{aligned}mse\left(\hat{\theta}_u\right) &= \mathrm{tr}\left\{\mathbb{E}\left[\left(\theta_u - \hat{\theta}_u\right)\left(\theta_u - \hat{\theta}_u\right)^{\mathrm{H}}\right]\right\} \\ &= \mathrm{tr}\left\{R_{\theta_u} - R_{\theta_u,q_o}\left[\ R_{\theta_o} + \left(\sigma_\xi^2 + \sigma_\varepsilon^2\right) I\ \right]^{-1} R_{\theta_u,q_o}^{\mathrm{H}}\right\} \\ &= \mathrm{tr}\left(R_{\theta_u} - R_{\theta_u,q_o} R_{q_o}^{-1} R_{\theta_u,q_o}^{\mathrm{H}}\right) \qquad (4.27)\end{aligned}$$

where R_{θ_u} is the auto correlation matrix of θ_u. Since $R_{\theta_u,q_o}\left[\ R_{\theta_o} + \left(\sigma_\xi^2 + \sigma_\varepsilon^2\right) I\ \right]^{-1} R_{\theta_u,q_o}^{\mathrm{H}}$ is a semi-positive definite matrix, then $mse\left(\hat{\theta}_u\right) \leq \mathrm{tr}\left(R_{\theta_u}\right)$.

Remark 4.4. Note that in (4.26) and (4.27), Moore-Penrose pseudoinverse should be applied if $R_{\theta_o} + \left(\sigma_\xi^2 + \sigma_\varepsilon^2\right) I$ is a singular matrix.

Then, the following Corollary can be obtained:

Corollary 4.5. *Spatial-temporal estimator proposed by (4.26) is efficient for both cases where causal relationship among available data points exists or does not exist.*

Proof. In the case that there is no causal relationship between each pair of available data points, it is straightforward to show that (4.26) is efficient because (4.27) achieves the CRLB

proposed by (4.18). In the case that the causal relationship among available points exists and $\mathrm{J}_{\theta,\mathrm{r}}$ is not a singular matrix, the lower bound of MSE of $\hat{\theta}_u$ can be expressed by

$$\begin{aligned}\mathbb{E}_{q_o,\theta}\left\{\| \hat{\theta}_u - \theta_u \|^2\right\} &\geq \mathrm{tr}\left\{\left[\begin{array}{cc} I & 0 \end{array}\right]\left(R_\theta - R_\theta A^{\mathbf{H}} R_{q_o}^{-1} A R_\theta \right)\left[\begin{array}{c} I \\ 0 \end{array}\right]\right\} \\ &= \mathrm{tr}\left\{R_{\theta_u} - \left[\begin{array}{cc} I & 0 \end{array}\right] R_\theta A^{\mathbf{H}} R_{q_o}^{-1} A R_\theta \left[\begin{array}{c} I \\ 0 \end{array}\right]\right\} \end{aligned} \tag{4.28}$$

Compared to the MSE in (4.27), it has

$$\begin{aligned}\mathrm{tr}\left(\begin{array}{c} R_{\theta_\mathrm{u}} - R_{\theta_u,q_o} R_{q_\mathrm{o}}^{-1} R_{\theta_u,q_o}^{\mathrm{H}} - R_{\theta_\mathrm{u}} \\ + \left[\begin{array}{cc} I & 0 \end{array}\right] R_\theta A^{\mathbf{H}} R_{q_o}^{-1} A R_\theta \left[\begin{array}{c} I \\ 0 \end{array}\right] \end{array}\right) &= \mathrm{tr}\left(\begin{array}{c} \left[\begin{array}{cc} I & 0 \end{array}\right] R_\theta A^{\mathbf{H}} R_{q_o}^{-1} A R_\theta \left[\begin{array}{c} I \\ 0 \end{array}\right] \\ -R_{\theta_u,q_o} R_{q_\mathrm{o}}^{-1} R_{\theta_u,q_o}^{\mathrm{H}} \end{array}\right) \\ &= \mathrm{tr}\left[\left(\begin{array}{c} A R_\theta \left[\begin{array}{c} I \\ 0 \end{array}\right]\left[\begin{array}{cc} I & 0 \end{array}\right] R_\theta A^{\mathbf{H}} \\ -R_{\theta_u,q_o}^{\mathrm{H}} R_{\theta_u,q_o} \end{array}\right) R_{q_\mathrm{o}}^{-1}\right] \end{aligned} \tag{4.29}$$

Equation (4.29) indicates that the spatio-temporal estimator can be shown to be efficient if $A R_\theta \left[\begin{array}{c} I \\ 0 \end{array}\right]\left[\begin{array}{cc} I & 0 \end{array}\right] R_\theta A^{\mathbf{H}} = R_{\theta_u,q_o}^{\mathrm{H}} R_{\theta_u,q_o}$ can be proved. Recall that A is a $MT + K_{\mathrm{av}} \times (N_m + N_u)T$ scaling matrix, each row of which is actually a product of a flow correlation vector, an inverse matrix of a cofactor matrix of R_θ and an insertion matrix. Thus, the row vector of A can be rewritten as

$$A_l = r_l^{\mathrm{T}} R_{\theta,l,l}^{*-1} B_l \tag{4.30}$$

where A_l stands for the l-th row of A, $R_{\theta,l,l}^{*}$ is the (l,l)-th co-factor matrix of R_θ and B_l is the l-th insertion matrix inserting the $(l + (N_m + N_u)T - MT + K_{\mathrm{av}})$-th column of $(N_m + N_u)T \times (N_m + N_u)T$ identity matrix with zero vector. By substituting (4.30) into (4.29), it is not hard to see that $r_l^{\mathrm{T}} R_{\theta,l,l}^{*-1} B_l R_\theta \left[\begin{array}{c} I \\ 0 \end{array}\right]$ is the flow correlation vector between the l-th available point and the unknown parameter vector, and thus $A R_\theta \left[\begin{array}{c} I \\ 0 \end{array}\right] \triangleq R_{\theta_u,q_o}^{\mathrm{H}}$ and Corollary 4.5 is proved. □

Corollary 4.5 reveals that (4.26) is an optimal estimator. Equation (4.27) shows that the

optimal spatial-temporal estimator requires an inverse operation of the auto correlation matrix of the observed parameter vector, which leads to high computational effort with a large size of the observed parameter vector. To tackle the problem, I propose a sub-optimal iterative multiple-points spatial-temporal Kriging technique, which iteratively imputes the missing points via the most appropriate available data. At each iteration, the most appropriate available data being used for imputation is selected such that the flow correlation with the missing points is maximum while the co-variance matrix of the available data is not ill-conditioned. Note that the available data in the i-th iteration may be the estimated value in the $i-1$-th iteration.

Let us define k_{av} and k_{im} as the constant number of the selected available data and the missing points being imputed at each iteration, respectively. Let e^{χ} be MSE of the iterated spatial-temporal Kriging when a specific imputation order $\chi \in \Phi$ is used. The cardinality of imputation order set Φ is determined by the number of total missing points $K_m + N_{\mathrm{u}}T$ and the missing points being imputed at each iteration k_{im}, where

$$|\Phi| = \prod_{i=1}^{\lfloor (K_m+N_{\mathrm{u}}T)/k_{\mathrm{im}} \rfloor} \begin{pmatrix} K_m + N_{\mathrm{u}}T - ik_{\mathrm{im}} \\ k_{\mathrm{im}} \end{pmatrix}$$

Then, e^{χ} can be expressed by

$$\begin{aligned} e^{\chi} &= \mathbb{E}_{q_o,\theta} \left\{ \sum_{i=1}^{I} \parallel \hat{\theta}_u^{i,\chi} - \theta_u^{i,\chi} \parallel^2 \right\} \\ &= \sum_{i=1}^{L} \mathbb{E}_{q_o,\theta} \left\{ \parallel \hat{\theta}_u^{i,\chi} - \theta_u^{i,\chi} \parallel^2 \right\} \qquad (4.31) \\ &= \sum_{i=1}^{L} e^{i,\chi} \qquad (4.32) \end{aligned}$$

where L is the number of iterations, $\hat{\theta}_u^{i,\chi}$ and $\theta_u^{i,\chi}$ are the estimated and real unknown parameter vector of the i-th iteration for the order χ. In what follows, the closed form of e^{χ} is given and the relation between e^{χ} and χ, k_{av} and k_{im} is investigated.

By (4.27), at the i-th iteration for a specific order χ, the $e^{i,\chi}$ can be expressed by

$$\begin{aligned} e^{i,\chi} &= \mathrm{tr}\left(R_{\theta_u}^{i,\chi}\right) - \mathrm{tr}\left(R_{\theta_u,q_{av}}^{i,\chi} R_{q_{av}}^{-1,i,\chi} R_{\theta_u,q_{av}}^{\mathrm{H},i,\chi}\right) \\ q_{av}^{i,\chi} &= \underset{q_{av} \in \left\{ q_{av}^{i-1,\chi} \cup \hat{\theta}_u^{i-1,\chi} \right\}}{\mathrm{argmax}} \left[\mathrm{tr}\left(R_{\theta_u,q_{av}}^{i,\chi} R_{q_{av}}^{-1,i,\chi} R_{\theta_u,q_{av}}^{\mathrm{H},i,\chi}\right) \right] \qquad (4.33) \end{aligned}$$

where $R^{i,\chi}_{\theta_u}$, $R^{i,\chi}_{q_{av}}$ and $R^{i,\chi}_{\theta_u,q_{av}}$ are the $k_{\text{im}} \times k_{\text{im}}$, $k_{\text{av}} \times k_{\text{av}}$ and $k_{\text{im}} \times k_{\text{av}}$ auto correlation matrix of the selected unknown parameter vector, available parameter vector and co-variance matrix between the selected unknown parameter vector and the available parameter vector, respectively. Note that $q^{i,\chi}_{av}$ in (4.33) is the most appropriate available data at the i-th iteration. Substituting (4.33) into (4.32), e^{χ} can be rewritten by

$$e^{\chi} = \text{tr}\left(R_{\theta_u}\right) - \sum_{i=1}^{I} \text{tr}\left(R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{av}} R^{\text{H},i,\chi}_{\theta_u,q_{av}}\right) \tag{4.34}$$

Remark 4.6. Note that at the i-th iteration, where $i > 1$, $e^{i,\chi}$ may be slightly different from that given by (3.34), because there is possibility that some elements of $\hat{\theta}_u^{i-1,\chi}$ are selected to be the available data. In such case, $\hat{\theta}_u^{i,\chi}$ should be rewritten by

$$\begin{aligned}\hat{\theta}_u^{i,\chi} &= \bar{\theta}_u^{i,\chi} + R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{av}} \left(q^{i,\chi}_{av} - \bar{\theta}^{i,\chi}_{av}\right) \\ &= \bar{\theta}_u^{i,\chi} + R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{av}} \left(\theta^{i,\chi}_{av} - \bar{\theta}^{i,\chi}_{av} + \varepsilon^{i,\chi} \right)\end{aligned} \tag{4.35}$$

and

$$\begin{aligned}e^{i,\chi} &= \text{tr}\left\{\mathbb{E}\left[\left(\theta_u^{i,\chi} - \hat{\theta}_u^{i,\chi}\right)\left(\theta_u^{i,\chi} - \hat{\theta}_u^{i,\chi}\right)^{\text{H}}\right]\right\} \\ &= \text{tr}\left\{ \mathbb{E}\left[\begin{array}{c} \theta_u^{i,\chi} - \bar{\theta}_u^{i,\chi} - \\ R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{av}} \left(\theta^{i,\chi}_{av} - \bar{\theta}^{i,\chi}_{av} + \varepsilon^{i,\chi}\right) \end{array} \right] \left[\begin{array}{c} \theta_u^{i,\chi} - \bar{\theta}_u^{i,\chi} - \\ R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{av}} \left(\theta^{i,\chi}_{av} - \bar{\theta}^{i,\chi}_{av} + \varepsilon^{i,\chi}\right) \end{array} \right]^{\text{H}} \right\} \\ &= \text{tr}\left[\begin{array}{c} R^{i,\chi}_{\theta_u} - 2R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{av}} R^{\text{H},i,\chi}_{\theta_u,q_{av}} - R^{i,\chi}_{\theta_u,\varepsilon} R^{-1,i,\chi}_{q_{\text{av}}} R^{\text{H},i,\chi}_{\theta_u,q_{av}} - R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{\text{av}}} R^{\text{H},i,\chi}_{\theta_u,\varepsilon} + \\ R^{i,\chi}_{\theta_u,q_{av}} R^{-1,i,\chi}_{q_{\text{av}}} \left(\begin{array}{c} R^{i,\chi}_{\theta_{av}} + \\ \Lambda^{i,\chi}_{\varepsilon} \end{array} \right) R^{-1,i,\chi}_{q_{\text{av}}} R^{\text{H},i,\chi}_{\theta_u,q_{av}} \end{array} \right]\end{aligned} \tag{4.36}$$

where $\varepsilon^{i,\chi}$ is the error term of the available data at the i-th iteration for the order χ, $\Lambda^{i,\chi}_{\varepsilon}$ is the co-variance matrix of $\varepsilon^{i,\chi}$, $R^{i,\chi}_{\theta_u,\varepsilon}$ is the co-variance matrix between $\theta_u^{i,\chi}$ and $\varepsilon^{i,\chi}$.

Remark 4.6 shows equation (4.36) reveals the same result as (4.33) when $\Lambda^{i,\chi}_{\varepsilon} = \left(\sigma^2_{\xi} + \sigma^2_{\varepsilon}\right) I$, i.e. all selected available data is observed data, rather than containing the available data estimated from the $i-1$-th iteration, where I is the identity matrix. For $i = 1$, it is clear that $\varepsilon^{i,\chi}$ only consists of the measurement error $\xi + \varepsilon$ from the observed data. But for $i > 1$, $\varepsilon^{i,\chi}$ may contain the estimation error delivered in the $i-1$-th iteration due to the available data selection in the current iteration. The estimation error can be traced back to the observed data and can

be classified into two types: measurement error propagation and error caused by insufficient statistics. The measurement error of the observed data can propagate with the evolution if the estimated value at the previous iteration is used as the available data at the current iteration. Error caused by insufficient statistics is attributable to that for some missing points at specific iterations, the selected available data may not cover all available information relating to the missing points. For example, in a specific iteration, there are $k_{\mathrm{av}}^{'}+u$ observed data having spatial-temporal correlation with the missing points. However, the k_{av} selected available data can only be traced back to $k_{\mathrm{av}}^{'}$ related observed data. The rest u observed data is not utilized and thus leads to additional estimation error.

From (4.36), e^{χ} can be expressed by

$$\begin{aligned} e^{\chi} &= \mathrm{tr}\left(R_{\theta_u}\right) - \sum_{i=1}^{I} \mathrm{tr}\left(R_{\theta_u,q_{av}}^{i,\chi} R_{q_{av}}^{-1,i,\chi} R_{\theta_u,q_{av}}^{\mathrm{H},i,\chi}\right) \\ &\quad - 2\sum_{i=1}^{I} \mathrm{tr}\left(R_{\theta_u,\varepsilon}^{i,\chi} R_{q_{av}}^{-1,i,\chi} R_{\theta_u,q_{av}}^{\mathrm{H},i,\chi}\right) + \sum_{i=1}^{I} \mathrm{tr}\left\{ R_{\theta_u,q_{av}}^{i,\chi} R_{q_{av}}^{-1,i,\chi} \left[\begin{array}{c} \Lambda_{\varepsilon}^{i,\chi} - \\ \left(\sigma_{\xi}^2+\sigma_{\varepsilon}^2\right) I \end{array} \right] R_{q_{av}}^{-1,i,\chi} R_{\theta_u,q_{av}}^{\mathrm{H},i,\chi} \right\} \end{aligned} \tag{4.37}$$

Then, the following Lemma can be derived:

Lemma 4.7. *Iterated spatial-temporal kriging estimator can achieve the CRLB proposed by (4.17) iff (if only if) the selected available data contain all spatial-temporal information relating to the missing points at each iteration. In such case, the performance of iterated spatial-temporal kriging estimator is irrelevant to a specific imputation order* χ.

Proof. When the selected available data contain all spatial-temporal information relating to the missing points at each iteration, $R_{\theta_u,\varepsilon}^{i,\chi}$ becomes zero, $\sum_{i=1}^{I} \mathrm{tr}\left(R_{\theta_u,q_{av}}^{i,\chi} R_{q_{av}}^{-1,i,\chi} R_{\theta_u,q_{av}}^{\mathrm{H},i,\chi}\right)$ achieves $\mathrm{tr}\left(R_{\theta_u,q_o} R_{q_o}^{-1} R_{\theta_u,q_o}^{\mathrm{H}}\right)$, and $\Lambda_{\varepsilon}^{i,\chi}$ only contains the measurement error propagated from the previous iterations. If the measurement error comes totally from the observed data, it is clear that $\Lambda_{\varepsilon}^{i,\chi} = \left(\sigma_{\xi}^2+\sigma_{\varepsilon}^2\right) I$ for each iteration and thus the last term in (4.37) disappears. For the case that the measurement error does not come directly from the observed data, but fully or partially from the data estimated in the previous iteration, $\Lambda_{\varepsilon}^{i,\chi}$ can be expressed by

$$\Lambda_{\varepsilon}^{i,\chi} = \mathbb{E}\left(\begin{bmatrix} \xi_1 + \varepsilon_1 \\ \vdots \\ \xi_{k_o} + \varepsilon_{k_o} \\ \varepsilon_1^{i-1,\chi} \\ \vdots \\ \varepsilon_{k_{av}^i}^{i-1,\chi} \end{bmatrix} \begin{bmatrix} \xi_1 + \varepsilon_1 \\ \vdots \\ \xi_{k_o} + \varepsilon_{k_o} \\ \varepsilon_1^{i-1,\chi} \\ \vdots \\ \varepsilon_{k_{av}^i}^{i-1,\chi} \end{bmatrix}^{\mathbf{H}} \right) \tag{4.38}$$

where k_{av}^i is the number of available data estimated in the previous iteration, the rest $k_{av} - k_{av}^i$ available data are from the observed data, when $k_{av}^i = k_{av}$, the available data are fully from the estimated data. Equation (4.38) can be further derived as

$$\Lambda_{\varepsilon}^{i,\chi} = \begin{bmatrix} \left(\sigma_{\xi}^2 + \sigma_{\varepsilon}^2\right) I & 0 \\ 0 & R_{\varepsilon}^{i,\chi} \end{bmatrix}$$

where $R_{\varepsilon}^{i,\chi}$ is not necessarily a diagonal matrix because of possibility of cross correlation among the estimation error at different iterations. To evaluate $\Lambda_{\varepsilon}^{i,\chi}$, we need to investigate how $\varepsilon^{i,\chi}$ propagates with i. Let us define $\varepsilon_n^{i,\chi}$ as the estimation error at the n-th missing point at the i-th iteration for an arbitrary imputation order, then it can be expressed by

$$\varepsilon_n^{i,\chi} = R_{\theta_n,q_{av}}^{i,\chi} R_{q_{av}}^{-1,i,\chi} \varepsilon^{i-1,\chi} = \sum_{k=1}^{k_{av}^i} w_k^i \varepsilon_k^{i-1,\chi} \tag{4.39}$$

where $w_k^i, k = 1 \cdots k_{av}^i$ is the weight vector revealing correlation between $\varepsilon^{i-1,\chi}$ is the error vector estimated at the i-1-th iteration. The k-th error term at the $i-1$-th iteration $\varepsilon_k^{i-1,\chi}$ can be further traced back to the observed data with the same procedure as (4.39), Thus, $\varepsilon_n^{i,\chi}$ can be further expressed by

$$\varepsilon_n^{i,\chi} = \sum_{k=1}^{k_o} \left(\prod_{\ell=1}^{i} w_{k,n}^{\ell} \right) (\xi_k + \varepsilon_k)$$

where $\prod_{\ell=1}^{i} w_k^{\ell}$ is the correlation between the k-th observed data and the n-th missing point at the i-th iteration, and thus it is the (n,k)-th element of the matrix $R_{\theta_n,q_o}^{i,\chi} R_{q_o}^{-1,i,\chi}$. Thus, the last term in (4.37) can be straightforwardly transformed to

$$\sum_{i=1}^{I} \mathrm{tr}\left\{ R_{\theta_u,q_o}^{i,\chi} R_{q_o}^{-1,i,\chi} \left[\left(\sigma_\xi^2+\sigma_\varepsilon^2\right) I - \left(\sigma_\xi^2+\sigma_\varepsilon^2\right) I \right] R_{q_o}^{-1,i,\chi} R_{\theta_u,q_o}^{\mathrm{H},i,\chi} \right\} = 0$$

Then, Lemma 4.7 is proved. □

Remark 4.8. Because the performance is irrelevant to the imputation order when the selected available data cover all related information, Lemma 4.7 is specially useful when we only want to estimate partial unknown data. In such case, the benefits of iterated spatial-kriging technique becomes more dominant because it can greatly lower the computational complexity without performance degradation. In a traffic network, the available data that cover all information relating to the missing points is not hard to be determined because of road topology.

Lemma 4.7 also suggests that there is possibility to simplify a certain form of matrix multiplication by reducing matrix dimension. From Lemma 4.7, the following corollary can be easily derived:

Corollary 4.9. *For any $K' \times N'$ matrix A and $N' \times N'$ non-singular Hermitian matrix M, where A and M can be written in their sub-matrices forms*

$$A = \begin{bmatrix} A_1 & A_2 & \cdots & A_N \end{bmatrix} \tag{4.40}$$

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1N} \\ M_{12}^H & M_{22} & \cdots & M_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1N}^H & M_{2N}^H & \cdots & M_{NN} \end{bmatrix} \tag{4.41}$$

, A_n and M_{mn}^H,$m \neq n$ are with the size of $K' \times N_n^{'}$ and $N_n^{'} \times N_m^{'}$, respectively, if there exists matrices λ_n,$n = 1 \cdots N$ that fulfill

$$A_m = \sum_{n=1}^{N} \lambda_n M_{nm}^H, m = 1 \cdots N \tag{4.42}$$

then the following equation can be obtained

$$AM^{-1}A^{\boldsymbol{H}} = \sum_{m,n=1}^{N} \lambda_m M_{mn} \lambda_n^H \tag{4.43}$$

Proof. Suppose that we use N' observed data to impute K' missing points. The N' observed data can be arbitrarily classified into N groups. The matrix M is the co-variance matrix of the N' observed data, which can be expressed by the form of (4.41). Each sub-matrix M_{mn} can

be considered as cross-correlation between the m-th and n-th group. We assume that the K' missing points is actually a linear combination of N groups of the observed data with the form:

$$\theta_{\mathrm{u}} = \sum_{n=1}^{N} \lambda_n \theta_{\mathrm{o}}^{(n)}$$

where $\theta_{\mathrm{o}}^{(n)}$ is the n-th group of the observed data and the definition of λ_n is the same as that is given in (4.42). From (4.27), the MSE of estimation error of the K' missing points can be determined by

$$mse\left(\theta_{\mathrm{u}}|q_o\right) = R_{\theta_{\mathrm{u}}} - AM^{-1}A^{\mathbf{H}} \tag{4.44}$$

where A is the cross-correlation matrix between θ_{u} and θ_{o}, and can be expressed by (4.42). Suppose that there exists K' available data with the same correlation to the observed data and covers all information about the missing points, then the MSE of estimation error of the K' missing points based on the K' available data can be expressed by

$$mse\left(\theta_{\mathrm{u}}|q_{av}\right) = R_{\theta_{\mathrm{u}}} - \sum_{m,n=1}^{N} \lambda_m M_{mn} \lambda_n^{\mathrm{H}} \tag{4.45}$$

From lemma 4.7, it can be known that (4.44) must be identical to (4.45). Then, Corollary 4.9 is proved. □

Remark 4.10. Note that Corollary 4.9 can also be verified by matrix operation. It is specially useful for the large dimension of M, where inverse operation becomes cumbersome. Corollary 4.9 can also be used to support Lemma 4.7.

4.4 Validation using Real Traffic Data

Based on the theoretical analysis proposed in sub-chapters 4.2.2 and 4.3, in this sub-chapter, the real traffic data provided by the transportation department of Kaohsiung, Taiwan is utilized to validate our theoretical findings.

The selected flow data was collected by 30 loop detectors on the road segments located in an urban area of Kaohsiung, Taiwan, over the period of 01/04/2013 to 30/04/2013 (Fig. 4.1). The black nodes in Fig. represent the deployed sensors. The horizontal and vertical axis represent the longitude and latitude, respectively. Each sensor provides the flow data of 10-minute interval from 00:00 to 23:59 for each day. To validate the theoretical findings in this chapter, the malfunctioned sensor and the unmeasured spatio-temporal points are intentionally generated with different missing ratios that ranges from 25% to 45% at every 5% increment. We

consider the Missing Completely at Random manner (MCR), where the malfunctioned sensors and the unmeasured spatio-temporal points are independently and uniformly distributed over the spatio-temporal domain.

Firstly, the SFEB (Squared Flow Error Bound) is calculated, which is determined by the co-variance matrix of the real flow vector, scaling matrix and the co-variance matrix of the observed flow vector. Then, the flow values on the malfunctioned sensors and the unmeasured spatio-temporal points are estimated by using the optimal spatio-temporal Kriging estimator, Nearest Historical Average (NHA) and K Nearest Neighbor (KNN) methods, and their performance is compared to each other in terms of RMSE (Root Mean Squared Error). The optimal and iterated spatio-temporal Kriging estimators were proposed in sub-chapter 4.3. NHA is the most commonly used method in the data estimation because it shows a stable performance regardless of the missing data size with an easy implementation [56]. NHA replaces the missing data by the arithmetic average or weighted average of the nearest historical data [15]. NHA does not incorporate the information from neighboring roads at the same day and is based on the assumption that traffic pattern at the same sensor at the same time instant is similar from day to day. KNN methods is to fill the missing data with the arithmetic or weighted average of data on K neighboring sensors, which are selected by searching for the data with close physical distances or equivalent distances [8].

The experimental results are shown in Fig. 4.2 to Fig. 4.5, which compare the squared root of SFEB, RMSE of optimal Kriging estimator, KNN and NHA for SNR (Signal to Noise Ratio) is equal to 10dB, 20dB, 30dB and 40dB, respectively. SNR is this thesis is defined by the ratio between the variance of real traffic flow data to the variance of measurement noise. To validate the impact on SFEB by different measurement noise power, measurement noise is intentionally created with different SNR. As can be seen from Fig. 4.2 to Fig. 4.5, theoretical RMSE of the Optimal Kriging Estimator (OKE) given by Equation (4.27) coincides with the squared root of SFEB given by Theorem 4.1 for all missing ratio and all SNRs. The experimental RMSE of the OKE depicts a gap with the squared root of SFEB for SNR = 10dB while generally coincides with the squared root of SFEB for higher SNR values. The experimental RMSE of the OKE is obtained by Equation (4.26), which requires to determine the co-variance matrix between the unmeasured data and the observed data, and the auto-correlation matrix of the observed data. With larger measurement noise, the determined co-variance matrix and auto-correlation matrix are subject to perturbation due to insufficient statistics, and thus the gap is created. RMSE delivered by KNN and NHA is dominantly larger than the squared root for SFEB and RMSE of OKE for all missing ratios and the all SNR. The results show that performance of any missing points estimators is lowered by SFEB, and the OKE is an efficient estimator, and thus Theorem 4.1 and Corollary 4.5 are validated. It can be seen that NHA

indeed shows a general stable performance for different missing ratio and the performance of KNN degrades with an increasing missing ratio. SFEB increases with an increasing variance of the measurement noise.

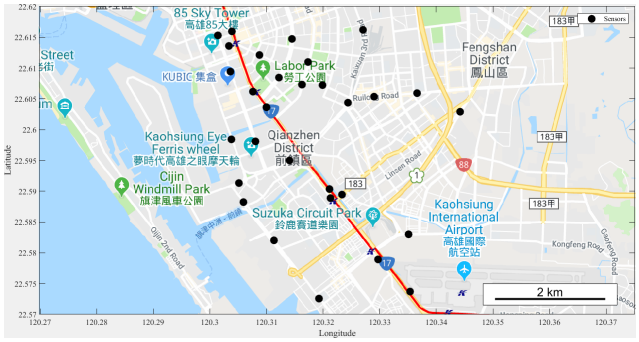

Figure 4.1: Selected map for experiment: urban area of Kaohsiung, Taiwan

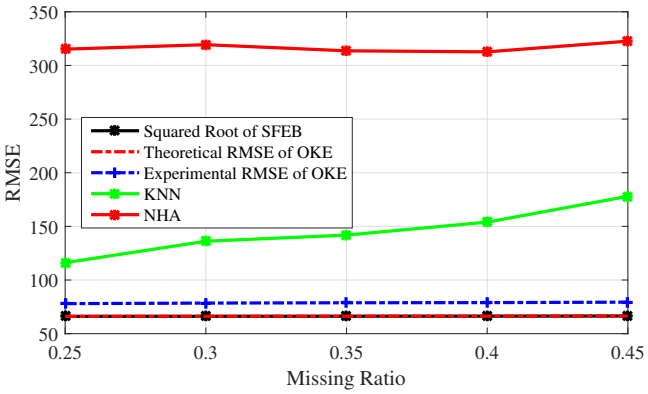

Figure 4.2: Comparison of the squared root of SFEB, RMSE of the optimal Kriging estimator, KNN and NHA for SNR = 10dB

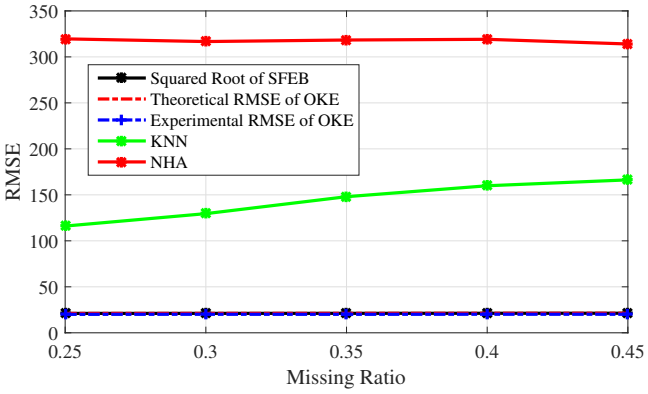

Figure 4.3: Comparison of the squared root of SFEB, RMSE of the optimal Kriging estimator, KNN and NHA for SNR = 20dB

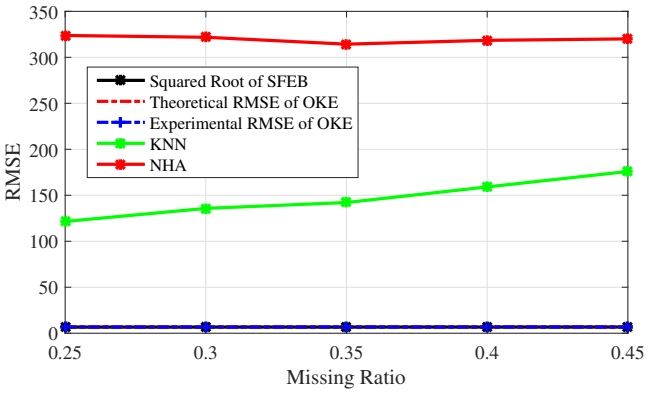

Figure 4.4: Comparison of the squared root of SFEB, RMSE of the optimal Kriging estimator, KNN and NHA for SNR = 30dB

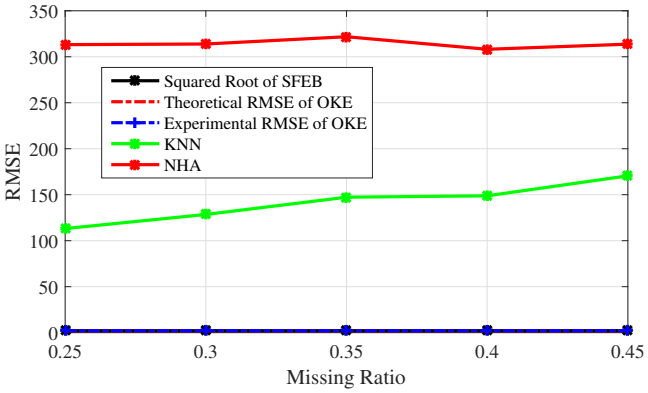

Figure 4.5: Comparison of the squared root of SFEB, RMSE of the optimal Kriging estimator, KNN and NHA for SNR = 40dB

4.5 Summary

In this chapter, I analyzed the CRLB of missing traffic data estimation and the sufficient and necessary condition for the existence of an unbiased estimator for both the cases that Fisher matrix is a non-singular or a singular matrix, respectively. It showed that the CRLB strongly relies on the three parameters, viz. co-variance of the unknown parameter vector, flow correlation between the unknown parameter vector and the available data, and the sensor locations. It showed the importance of relationship between the number of missing points and the ran of the Fisher matrix, i.e., and unbiased estimator of the unknown parameter vector can be found if only if the number of missing points is less than or equal to the rank of the Fisher matrix. I also derived an inequality on the variance of estimation error for the case that the above mentioned condition is not fulfilled.

I proved that the spatio-temporal estimator proposed in [17] is optimal for both cases that causal relationship between available data points exists or not. To overcome high computational complexity with a large size of the observed parameter vector, I proposed a sub-optimal iterated multiple-points spatio-temporal Kriging technique . I derived the variance of estimator error and showed the sufficient and necessary condition with which the iterated multiple-points spatio-temporal Kriging estimator can achieve the CRLB, i.e. the selected available data covers all information relating to the missing points at each iteration, in such case, estimation order becomes irrelevant to the performance. I further derived an useful corollary that can simplify a certain form of matrix multiplication by reducing matrix dimension. This corollary becomes more significant with a large matrix dimension where inverse operation becomes cumbersome.

The proposed results in this chapter can be used as a benchmark to evaluate the performance of missing data estimators in future work. The performance of any missing data estimators is lower bounded by the proposed SFEB. Corollary 4.2 can be utilized to optimize the sensor locations and the number of deployed sensors in the network. An application example is to deploy the sensors on the locations so that FIM is a full-rank matrix. Lemma 4.7 can be utilized to reduce computational complexity with a large size of the observed parameter vector. As an example, the missing points can be estimated iteratively with optimal selection of available data according to Lemma 4.7 at each iteration, rather than using the optimal Kriging estimator.

Chapter 5

Optimum Traffic Control in Sensor-Equipped Road Networks

5.1 Overview

Chapter 3 and Chapter 4 investigated how to improve estimation performance using limited number of sensors and the fundamental limits of traffic data estimation, respectively. High estimation performance of traffic data estimation is vital for ATMS (Advance Traffic Management Systems) to deploy traffic control polices to economically utilize the spatio-temporal resources of roads and traffic signal settings to maximize road network capacity.

As mentioned in Chapter 1, almost all existing research efforts proposed simple control policies to optimize throughput by only considering drivers' route choice or traffic signal control, and without considering the flow divergence among different OD pairs.

This chapter investigates a hybrid dynamical system, which incorporates flow swap process, green-time proportion swap process and flow divergence for a general network with multiple OD pairs and multiple routes, where flow swap process is specified in which traffic swaps from more costly to less costly input links, green-time proportion swap process is specified in which green time at each intersection swaps from less pressurized stages to more pressurized stages. flow may diverge at each intersection from on OD pair to other OD pairs. Unlike the dynamical system model considered in [49] where bottleneck delays need to be intentionally constructed to yield the equilibrium flow vector and green-time proportion vector, I propose a novel control policy to fill that gap by only adjusting the green-time proportion vector. A sufficient condition for the existence of equilibrium of the dynamical system is derived under the mild constraints that (1) the travel cost function and stage pressure function should be continuous functions; (2) the flow and green-time proportion swap process project

all flow and green-time proportion vectors on the boundary of the feasible region onto itself. I derive the condition of unique equilibrium for fixed green-time proportion vector and show that with varying green-time proportion vector, the set of of equilibrium is a compact, non-convex set, and with the same partial derivative of travel cost function with respect to the flow and green-time proportion vectors. Finally, the stability of the proposed dynamical system is proved by using Lyapunov stability analysis.

5.2 Novel Traffic Assignment Model in a Realistic Network

For simplicity, let us firstly introduce the novel traffic assignment model in a simple network with only two OD pairs, each of which has two routes. Then, the traffic assignment model will be extended to a generalized form with multiple OD pairs, each of which has M routes with $M > 2$. We use the definition of capacity maximization given in [49], which can be described as

Definition 5.1. Suppose P is a control policy. "The control policy P maximizes network capacity" means that: for a network with a rigid or steady demand and traffic signals; if there is a route-flow vector $\boldsymbol{X}$, which meets the given inelastic demand and is within the capacity region of the given network, then there exist a route-flow vector X^*, delay vector b^*, green-time proportion vector G^*, which meet the given inelastic demand, and is within the capacity limitations of the given network. Furthermore, X^*, b^* and G^* fulfill

1. X^* is a Wardrop equilibrium when the delay vector, green-time vector is $\begin{bmatrix} b^* & G^* \end{bmatrix}$ and
2. G^* satisfies the control policy P when the delay vector, route-flow vector is $\begin{bmatrix} b^* & X^* \end{bmatrix}$.

We give the definition of reserve capacity at equilibrium by

Definition 5.2. The reserve capacity at equilibrium is defined by the maximum possible increase in traffic demand that can be accommodated in the network considering (optimum) traffic signal control at each intersection and drivers' route choice behavior [68].

5.2.1 Equilibrium in a Simple Network without Signal Control

Consider a simple network with two OD pairs with two routes each in a steady quasi-dynamic state with vertical queue (Fig. 5.1). The vertical queue assumes an idealized scenario for analytic purpose that vehicles on a roadway stack up upon one another at the point where

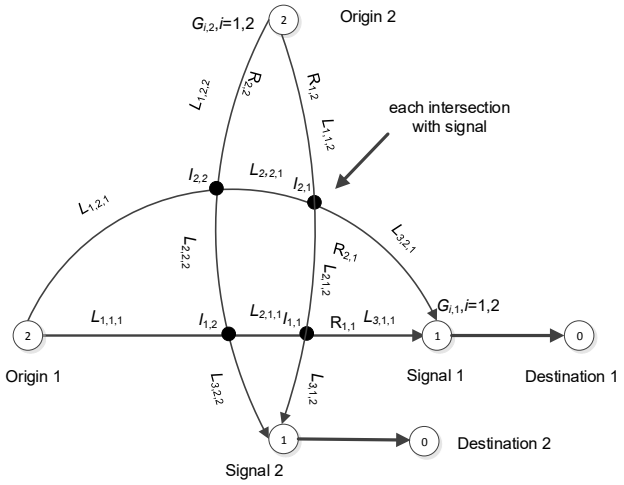

Figure 5.1: A signalized network with two OD pairs with two routes each. There is traffic signal at each intersection.

congestion begins, rather than backing up along the length of roadway. Vertical queue was used in [36, 49, 51]. The definition of "steady quasi-dynamic" can be found in [49]. The symbol $R_{n,m}$ in Fig. 5.1 represents the n-th route of the m-th OD pair, where $n,m = 1,2$.

Let us define $C_{n,m}$, $L_{n,m}$, $s_{n,m}$, $X_{n,m}$, $b_{n,m}$, $G_{n,m}$, $I_{n,n'}$ as the free-flow travel cost via $R_{n,m}$, the length of $R_{n,m}$, the saturation flow at $R_{n,m}$, the flow on $R_{n,m}$, the bottleneck delay of $R_{n,m}$ at the merging point (Signal 1 and 2 in Fig. 5.1), the green-time proportion assigned to $R_{n,m}$ at the merging point and the intersection between the n-th route of OD pair 1 and the n'-th route of OD pair 2, respectively. Let $G_{l,I_{n,n'}}$ and $b_{l,I_{n,n'}}$ be the green-time proportion and the bottleneck delay assigned to the l-th stage of the intersection $I_{n,n'}$, respectively. The unit of $C_{n,m}$, $b_{n,m}$ and $b_{l,I_{n,n'}}$ are seconds and the unit of $s_{n,m}$ and $X_{n,m}$ are vehicles per second. The saturation flow at $R_{n,m}$ denotes the capacity of $R_{n,m}$ at the critical density ρ_c if the green-time proportion of that stage if one for the traffic signals at all intersections and the merging point. The symbol $L_{l,n,m}, l = 1,2,3; n,m = 1,2$ is defined by the l-th link for the n-th route of the m-th OP pair, $X_{l,n,m}$ is defined by the flow on the link $L_{l,n,m}$, $p_{(l,n,m),(l',n',m')}$ is defined by turning rate from $L_{l,n,m}$ to $L_{l',n',m'}$.

Let us define F, D and S by the feasible, demand and supply sets, respectively, then it has

$$F = \left\{ \begin{pmatrix} G_{k,m} \\ G_{l,I_{n,n'}} \\ n,n' = 1,2 \\ k,l = 1,2 \\ m = 1,2 \end{pmatrix} : \begin{matrix} G_{1,m} + G_{2,m} = 1, \\ G_{1,I_{n,n'}} + G_{2,I_{n,n'}} = 1, \\ G_{k,m} \geq 0, G_{l,I_{n,n'}} \geq 0 \end{matrix} \right\} \tag{5.1}$$

Suppose for convenience that $s_{2,m} > s_{1,m}$ and $C_{2,m} > C_{1,m}$ for $m = 1,2$; the other scenarios can be readily accommodated by our analysis. Let us define T_1 and T_2 as fixed demands at the origin 1 and 2, respectively. It assumes that there is no flow divergence at each intersection and no traffic signal coordination, then, $p_{(l,n,m),(l',n',m')} = 1$ for $l' = l+1, (n,m) = (n',m')$ and $p_{(l,n,m),(l',n',m')} = 0$ for $(n,m) \neq (n',m')$.

Then, D and S can be obtained by

$$D = \left\{ \begin{pmatrix} X_{n,m} \\ m,n = 1,2 \end{pmatrix} : \begin{matrix} X_{1,m} + X_{2,m} = T_m \\ X_{n,m} \geq 0 \end{matrix} \right\} \tag{5.2}$$

and

$$S = \left\{ \begin{pmatrix} X_{k,m} \\ G_{k,m} \\ G_{l,I_{n,n'}} \\ n,n' = 1,2 \\ l,m = 1,2 \\ k = 1,2 \end{pmatrix} : \begin{matrix} X_{k,m} \leq \\ \min \begin{pmatrix} s_{k,m} G_{k,m} \\ s_{k,m} G_{m,I_{k,n'}} \end{pmatrix} \end{matrix} \right\} \tag{5.3}$$

From (5.2) and (5.3), the feasible demand set for T_m with $m = 1,2$ can be determined by

$$\begin{aligned} T_m &\leq \min \begin{pmatrix} s_{1,m} G_{1,m} \\ s_{1,m} G_{m,I_{1,n'}} \end{pmatrix} + \min \begin{pmatrix} s_{2,m} G_{2,m} \\ s_{2,m} G_{m,I_{2,n'}} \end{pmatrix} \\ &\leq s_{1,m} G_{1,m} + s_{2,m} G_{2,m} \leq s_{2,m} \end{aligned} \tag{5.4}$$

and

$$T_1 + T_2 \leq \min \begin{bmatrix} s_{2,1} + s_{2,2} \\ \max(s_{1,1}, s_{1,2}) + \max(s_{2,1}, s_{2,2}) \\ \max(s_{1,1}, s_{2,2}) + \max(s_{2,1}, s_{1,2}) \end{bmatrix} \tag{5.5}$$

No green-time proportion in F can be found to satisfy the route flow in S if T_1 and T_2 are not in the feasible demand set given by (5.4) and (5.5).

By the definition of Wardrop equilibrium that no traveler can reduce his travel cost by unilaterally changing routes [42], equilibrium flow on each route is obtained by solving the following equilibrium problem:

$$\begin{aligned} \min_X \sum_a \int_0^{X_a} t_a(X,G)\,dX \\ st.(5.1),(5.2),(5.3) \end{aligned} \tag{5.6}$$

where $t_a(X,G)$ is a strictly monotonically increasing travel cost function on the a-th route by the given flow vector X and the green-time proportion vector G. Taking first derivative of the objective function and setting it to zero, the equilibrium flow can be obtained by

$$\begin{aligned} t_{1,m}(X^*,G) = t_{2,m}(X^*,G) \\ m = 1,2 \end{aligned} \tag{5.7}$$

which can be rewritten into the following equations:

$$\begin{aligned} C_{1,1} + b_{1,1} + b_{1,I_{1,1}} + b_{1,I_{1,2}} = C_{2,1} + b_{2,1} + b_{1,I_{2,1}} + b_{1,I_{2,2}} \\ C_{1,2} + b_{1,2} + b_{2,I_{1,1}} + b_{2,I_{1,2}} = C_{2,2} + b_{2,2} + b_{2,I_{2,1}} + b_{2,I_{2,2}} \end{aligned} \tag{5.8}$$

In [49], the bottleneck delays within a quasi-dynamic model are given by

$$b = Q/sG \tag{5.9}$$

where Q is the expected length of vertical queue at exit of the link and it is an explicit variable that is determined by the green time and the route flows. However, Q is hard to measure or estimate. The solution is to express the bottleneck delay as a function of the route flows, the green-time proportion and the cycle duration of traffic light.

The maximum bottleneck delay can be interpreted as the required time to *vacate* a queue caused by red signals. Then, in the simple network proposed in [49] with only one OD pair, the maximum bottleneck delay at one route fulfills the following relation [69]:

$$X(1-G)\Delta T + Xb = sb \tag{5.10}$$

It follows that the maximum bottleneck delay can be expressed by

$$b = \frac{X(1-G)\Delta T}{s-X} \tag{5.11}$$

where ΔT is the cycle duration. If $X = sG$, then $b = G\Delta T$, which is equal to the green time duration. An ever-increasing queue will be generated when $X > sG$. In the network proposed in Fig. 5.1, the maximum bottleneck delays at signal 1 and 2 can be expressed by

$$b_{n,m} = \frac{X_{n,m}(1-G_{n,m})\Delta T}{s_{n,m}-X_{n,m}}; m,n = 1,2 \tag{5.12}$$

The maximum bottleneck delays at each intersection can be expressed by

$$b_{l,I_{n,n'}} = \frac{X_{n,l}\left(1-G_{l,I_{n,n'}}\right)\Delta T}{s_{n',l}-X_{n,l}}; l,n,n' = 1,2 \tag{5.13}$$

Substituting (5.12) and (5.13) to (5.8), the equilibrium flow $X^*_{1,1}$, $X^*_{2,1}$, $X^*_{1,2}$ and $X^*_{2,2}$ can be determined by

$$C_{1,1} + \frac{X^*_{1,1}\left[3-\left(G_{1,1}+G_{1,I_{1,1}}+G_{1,I_{1,2}}\right)\right]\Delta T}{s_{1,1}-X^*_{1,1}} = C_{2,1} + \frac{X^*_{2,1}\left[2+\left(G_{1,1}-G_{1,I_{2,1}}-G_{1,I_{2,2}}\right)\right]\Delta T}{s_{2,1}-X^*_{2,1}}$$

$$C_{1,2} + \frac{X^*_{1,2}\left[1-\left(G_{1,2}-G_{1,I_{1,1}}-G_{1,I_{1,2}}\right)\right]\Delta T}{s_{1,2}-X^*_{1,2}} = C_{2,2} + \frac{X^*_{2,2}\left[G_{1,2}+G_{1,I_{2,1}}+G_{1,I_{2,2}}\right]\Delta T}{s_{2,2}-X^*_{2,2}} \tag{5.14}$$

Combining with (5.1), (5.2), (5.3), (5.12) and (5.13), the equilibrium flow and the bottleneck delays can be expressed by a function of $G_{n,m}$, $G_{l,I_{n,n'}}$, $C_{n,m}$, ΔT, T_m and $s_{n,m}$, where $C_{n,m}$, ΔT , T_m and $s_{n,m}$ are constants while $G_{n,m}$ and $G_{l,I_{n,n'}}$ are adjustable.

When flow divergence is allowed, $0 < p_{(l,n,m),(l',n',m')} < 1$ when $L_{l,n,m}$ and $L_{l',n',m'}$ are adjacent, while $p_{(l,n,m),(l',n',m')} = 0$ when $L_{l,n,m}$ and $L_{l',n',m'}$ are non-adjacent, and $\sum_{l' \in \mathcal{N}_{(l,n,m)}} p_{(l,n,m),l'} = 1$, where $\mathcal{N}_{(l,n,m)}$ is the set of all downstream neighboring links of $L_{l,n,m}$. The supply set S

should be modified as

$$S = \left\{ \begin{pmatrix} X_{l,n,m} \\ G_{n,m} \\ G_{l,I_{n,n'}} \end{pmatrix} : \begin{array}{c} X_{1,2,1} \le s_{2,1} G_{1,I_{2,2}} \\ X_{1,1,1} \le s_{1,1} G_{1,I_{1,2}} \\ \vdots \\ X_{3,2,2} \le s_{2,2} G_{2,I_{1,2}} \end{array} \right\} \tag{5.15}$$

where

$$X_{l,n,m} = \sum_{l',n',m'} X_{l',n',m'} p_{(l',n',m'),(l,n,m)} \tag{5.16}$$

Combining with (5.2), the demand set D can be rewritten by

$$D = \left\{ (X, T_{\mathrm{s}}) : \begin{array}{c} P'X \le S', X \ge 0 \\ 1^{\mathrm{T}} X = T_{\mathrm{s}} \end{array} \right\} \tag{5.17}$$

where $X = \begin{bmatrix} X_{1,1,1} & X_{1,2,1} & X_{1,1,2} & X_{1,2,2} \end{bmatrix}^{\mathrm{T}}$, 1 is all one vector, T_{s} is the sum of demand, S' contains the maximum saturation flow at each intersection

$$S' = \begin{bmatrix} s_{2,1} & s_{2,2} & \max(s_{2,1}, s_{2,2}) & \cdots & \max(s_{1,1}, s_{1,2}) \end{bmatrix}^{\mathrm{T}} \tag{5.18}$$

P' is the turning rate matrix that contains the sum of turning rate from each input flow to the links, which intersect at an intersection, and thus the k-th row and n-th column element of P' can be expressed by

$$P'_{k,n} = \sum_{L_l \in I_k^-} \prod_{m = L_n \to L_l} p_{m,m+1} \tag{5.19}$$

where I_k^- is the set of incident links at the k-th intersection, L_n is the n-th input link, $L_n \to L_l$ represents the path from the n-th input link to the l-th incident link, $p_{m,m+1}$ is the turning rate from the m-th link to the $m+1$-th downstream neighboring link.

Remark 5.3. Note that the demand set given by (5.17) assumes that there is no flow generated or dissipated at each intersection or link. When considering the flow generation or dissipation, S' should be replaced by $S' + W$, where W is considered as an error term.

Recall that the maximum bottleneck delay at each exit of a link (intersection) is determined by the link flow, the saturation flow, the green-time proportion and the cycle duration of traffic light. We assume that the turning rates, i.e., the proportion of vehicles turn left, right or go straight etc., at each intersection can be estimated from empirical data. By taking into account

the turning rates, (5.10) should be reformulated as

$$pX(1-G)\Delta T + pXb = psb \tag{5.20}$$

where the turning rate p appears both on the left and on the right side of (5.20), and thus it can be neglected. Then, by a given green-time proportion vector G, the maximum bottleneck delays at signal 1 and 2 can be determined by

$$b_{n,m} = \frac{X_{3,n,m}(1-G_{n,m})\Delta T}{s_{n,m} - X_{3,n,m}} \tag{5.21}$$

The maximum bottleneck delay at each intersection can be expressed by

$$b_{l,I_{n,n'}} = \frac{X_{m,k,l}\left(1-G_{l,I_{n,n'}}\right)\Delta T}{s_{k,l} - X_{m,k,l}} \tag{5.22}$$

where $k = n$ or n', m = 1, 2 or 3. Then, the equilibrium flow $X^*_{l,n,m}, l = 1,2,3; n,m = 1,2$ can be determined by combining (5.8), (5.16), (5.21) and (5.22).

The equilibrium flow solved by (5.8) has a unique solution because $t_a(X,G)$ is a non-decreasing function of X for a given G. When no solution can be found within the set S, that means, at least one route flow for each OD pair has approached the boundary of the set S for the given G, then an ever-increasing queue will be generated on that route because more vehicles prefer to choose that cheaper route.

Note that S, D and F given by (5.15), (5.17) and (5.1) can be extended to a general network with multiple OD pairs. I will show in the next sub-chapter how this may be done.

5.2.2 Extension to A General Network

Consider a general network with M OD pairs, each of which has N_m routes. Let us define $R_{n,m}$ as the n-th route of the m-th OD pair, $N_{I_{n,m}}$ as the number of intersections for the n-th route of the m-th OD pair, $X_{l,n,m}$ as the link flow at the l-th link of the n-th route of the m-th OD pair, $b_{l,n,m}$ as the maximum bottleneck delay at the l-th link of the n-th route of the m-th OD pair. Then, the number of intersections in this general network is

$$N_I = \sum_{m=1}^{M}\sum_{n=1}^{N_m} N_{I_{n,m}}$$

The demand set D of route flow vectors with non-negative components meeting all origin-destination demands is given by

$$D = \left\{ (X_{\mathrm{I}}, T_{\mathrm{s}}) : \begin{array}{c} P'X_I \leq S', X_{\mathrm{I}} \geq 0 \\ 1^{\mathrm{T}} X_{\mathrm{I}} = T_{\mathrm{s}} \end{array} \right\} \tag{5.23}$$

where X_{I} is the input flow vector with $X_{\mathrm{I}} = \left[\begin{array}{cccc} X_{1,1,1} & X_{1,2,1} & \cdots & X_{1,N_M,M} \end{array} \right]^{\mathrm{T}}$, 1 is an all-one vector, T_{s} is the sum of demands, S' contains the maximum saturated incident flow at N_I intersections and M exits of OD pairs with $S' = \left[\begin{array}{cccc} s_{1,\mathrm{m}} & s_{2,\mathrm{m}} & \cdots & s_{N_I+M,\mathrm{m}} \end{array} \right]^{\mathrm{T}}$, P' is the turning rate matrix given by (5.19). The feasible set F contains all feasible green-time proportion vectors, where the green time proportions at each intersection sum to one, are non-negative and given by

$$F = \left\{ \left(\begin{array}{c} G_{n,m} \\ G_{m,I_{nm,n'm'}} \end{array} \right) : \begin{array}{c} \sum_{n=1}^{N_m} G_{n,m} = 1 \\ \sum_{m,m'} G_{m,I_{nm,n'm'}} = 1 \\ G_{n,m} \geq 0 \\ G_{m,I_{nm,n'm'}} \geq 0 \\ m = 1 \cdots M \\ n = 1 \cdots N_m \end{array} \right\} \tag{5.24}$$

The supply set S contains link flow and green time proportion, for which each link flow is no greater than the saturation flow multiplied by the link green time proportion, and it is given by

$$S = \left\{ \left(\begin{array}{c} X_{l,n,m} \\ G_{n,m} \\ G_{m,I_{nm,n'm'}} \end{array} \right) : \begin{array}{c} X_{l,n,m} \leq s_{n,m} G_{m,I_{nm,n'm'}} \\ X_{l,n,m} \geq 0 \\ m = 1 \cdots M \\ n = 1 \cdots N_m \end{array} \right\} \tag{5.25}$$

where the link flow $X_{l,n,m}$ is the sum of the flows at all incident links multiplied by the respective turning rate between the incident links and the current link.

By putting $X_{l,n,m}$ into the vector $X = \left[\begin{array}{cccc} X_{1,1,1} & X_{2,1,1} & \cdots & X_{L_{N_M},N_M,M} \end{array} \right]^{\mathrm{T}}$ and $b_{l,n,m}$ into the vector $b = \left[\begin{array}{cccc} b_{1,1,1} & b_{2,1,1} & \cdots & b_{L_{N_M},N_M,M} \end{array} \right]^{\mathrm{T}}$, the equilibrium flow on each link can be determined by finding a link-flow vector $X^* = \left[\begin{array}{cccc} X^*_{1,1,1} & X^*_{2,1,1} & \cdots & X^*_{L_{N_M},N_M,M} \end{array} \right]^{\mathrm{T}}$ and a green-time proportion vector $G^* = \left[\begin{array}{ccccccc} G^*_{1,1} & \cdots & G^*_{N_M,M} & G^*_{1,I_{11,12}} & \cdots & G^*_{M,I_{N_M M,N_{M'}M'}} \end{array} \right]^{\mathrm{T}}$, which minimize the sum of integrals of the link performance functions over the set of all $[X,G] \in D \times F \cap S$:

$$[X^*, G^*] = \min_{[X,G]} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{l=1}^{L_{n,m}} \int_0^{X_{l,n,m}} t_{l,n,m}(X,G)\, dX \tag{5.26}$$
$$st.\,(5.23),(5.24),(5.25)$$

Remark 5.4. Note that $D \times F$ represents the Cartesian product of the sets D and F. The symbol $\cap$ represents the intersection of two sets.

When vertical queue is considered and bottleneck delay is neglected, (5.26) is equivalent to the weighted total travel time minimization problem [42], which aims at finding a solution minimizing the sum of product of link flow and link cost:

$$\begin{aligned}[X^*, G^*] &= \min_{[X,G]} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{l=1}^{L_{n,m}} t_{l,n,m}(X,G) X_{l,n,m} \\ &= \min_{[X,G]} t^{\mathrm{T}}(X,G) X \end{aligned} \tag{5.27}$$

The problem (5.27) is equivalent to finding a $[X^*, G^*]$ such that $\left[-t^{\mathrm{T}}(X^*, G^*), 0\right]$ is normal at $[X^*, G^*]$ to the set $D \times F \cap S$ (See [49] for the definition of "normal").

The reserve capacity can be determined by finding a $[X^*, G^*]$ that maximizes the sum of input flow over the set of all $[X, G] \in D \times F \cap S$:

$$RC = \max_{X^* \in \mathscr{X}^*} \sum_{k=1}^{\Sigma_{m=1}^{M} N_m} X_{k,\mathrm{I}}^* - T_{\mathrm{c}} \tag{5.28}$$

where T_{c} is the current demand, $\mathscr{X}^*$ is the set of equilibrium flow vectors corresponding to the given green-time proportion vectors in the feasible set F:

$$\mathscr{X}^* = \left\{ \begin{bmatrix} X^* \\ G^* \end{bmatrix} : \begin{bmatrix} X^* \\ G^* \end{bmatrix} = \min_{[X,G] \in D_{T_{\mathrm{s}}} \times F \cap S} t^{\mathrm{T}}(X,G) X \right\} \tag{5.29}$$
$$D_{T_{\mathrm{s}}} = \left\{ X_{\mathrm{I}} : 1^{\mathrm{T}} X_{\mathrm{I}} = T_{\mathrm{s}} \right\}, T_{\mathrm{s}} \in D$$

Note that the set of equilibrium flows and green-time proportion vectors given by (5.29) are obtained by searching $t^{\mathrm{T}}(X,G)X$ over S and all X satisfying $1^{\mathrm{T}} X_{\mathrm{I}} = T_{\mathrm{s}}$, where $T_s \in D$, while (5.27) delivers a point by a given T_s. The problems (5.28) and (5.29) are equivalent to firstly determining $\mathscr{X}^*$, which consists of $[X^*, G^*]$ that lets $\left[-t^{\mathrm{T}}(X^*, G^*), 0\right]$ be normal at $[X^*, G^*]$ to the set $D_{T_{\mathrm{s}}} \times F \cap S$ for each $T_s \in D$, and then finding the X^* such that

$P = \left[\begin{array}{cc} 1^{\mathrm{T}}_{\sum_{m=1}^{M} N_m} & 0^{\mathrm{T}}_{N_L - \sum_{m=1}^{M} N_m} \end{array} \right]^{\mathrm{T}}$ is normal at X^* to the set $\mathscr{X}^*$, where 1_K and 0_K are $K \times 1$ all-one vector and all-zero vector, respectively.

Remark 5.5. Note that there is possibility that no $[X^*, G^*]$ can be found for some $T_s \in D$ due to a lack of intersection between D_{T_s} and $F \cap S$. It is also possible for some $[X^*, G^*]$ that they approach a point at the boundary of $D_{T_s} \times F \cap S$, at which $\left[-t^{\mathrm{T}}(X^*, G^*), 0\right]$ is not normal to the supporting hyperplane. In these cases, $[X^*, G^*]$ are removed from $\mathscr{X}^*$.

5.2.3 Applying an Optimum Control Policy to the General Network

Directly solving the equilibrium flow and green-time proportion vector from (5.27) requires knowing the sum of input flow T_s, which is hard to realize in practice. Smith has proved that with the P0 control policy and adjustable bottleneck delay, the capacity can be maximized in a simple network by re-routing and traffic signal assignment without knowing the demand information. Prior to applying the P0 control policy to the proposed general network, I firstly introduce the original P0 control policy proposed by Smith and then give the condition, under which the P0 control policy can be utilized to maximize the network capacity with a steady demand and the network reserve capacity.

Consider a simple network with just two routes, the original P0 control policy aims at equalizing the product of saturation flow and bottleneck delay by selecting appropriate green-time proportion vector, that is, $s_1 b_1 = s_2 b_2$[49]. If this is not possible, then, the policy should select green-time proportion vector to guarantee two numbers to be as equal as possible. For the proposed general network with multiple OD pairs, the P0 control policy aims at equalizing the stage pressure $P_{n,m} = \sum_{l=1}^{N_{n,m}} s_{l,n,m} b_{l,n,m} = \sum_{l=1}^{N_{n,m}} P_{l,n,m}, n = 1, 2 \cdots N_m$ for M OD pairs by selecting appropriate green-time proportion vectors. The stage pressure $P_{n,m}$ is defined by adding the terms $s_{l,n,m} b_{l,n,m}$ over the links on the n-th route of the m-th OD pair and $P_{l,n,m}$ is the pressure at the (l,n,m)-th link.

Because the target of the P0 control policy may not always be achieved or even achievable although the green-time proportion vector can be slowly adjusted to approach the target [49], the P0 control policy for the simple network in [49] was stated as

> *For any X and delay vector b, choose a stage green-time proportion vector G in F so that the stage pressure vector $P(b)$ is normal at G to F.*

Recall that in the proposed general network, there are $N_{I_{n,m}}$ traffic signals at all intersections for the n-th route of the m-th OD pair and one traffic signal at the merging point, and thus there are $N_{I_{n,m}} + 1$ adjustable green-time proportions for the n-th route of the m-th OD pair. Note that any changes of the green-time proportions at the $N_{I_{n,m}}$ intersections will influence

the equilibrium flow for other OD pairs because of F given by (5.24). Recall that it assumes that the cycle duration ΔT for each traffic signal is the same, thus, the P0 control policy for the general network can be restated as

For any X and delay vector b choose a stage green-time proportion vector $G = \begin{bmatrix} G_{1,1,1} & \cdots & G_{l,n,m} & \cdots & G_{L_{N_M},N_M,M} \end{bmatrix}^T$ in F to maximize $\sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{l=1}^{N_n}\begin{pmatrix} s_{l,n,m}b_{l,n,m} \\ \times G_{l,n,m} \end{pmatrix}$, that is, to let $P(b) = \begin{bmatrix} P_{1,1,1} & \cdots & P_{l,n,m} & \cdots & P_{L_{N_M},N_M,M} \end{bmatrix}^T$ be normal at G to F.

The green-time proportion is represented by $G_{l,n,m}$ instead of $G_{l,I_{n,n'}}$ occurring in F and S for the sake of simplicity of expression, $G_{l,n,m}$ stands for the green-time proportion assigned to the (l,n,m)-th link.

In what follows, I will give the condition that P0 control policy can maximize the general network capacity with vertical queue. Recall that the problem (5.27) is to find a $[X^*,G^*]$ such that $\left[-t^{\mathrm{T}}(X^*,G^*),0\right]$ is normal at $[X^*,G^*]$ to the set $D\times F\cap S$, the equilibrium travel cost $t^{\mathrm{T}}(X^*,G^*)$ varies with X^* and G^* because it consists of the free travel cost C (constant) and the bottleneck delay b (non-negative function of X^* and G^*). Suppose that there is a $[X^*,G^*]$ at which $[-C,0]$ is normal to the set $D\times F\cap S$. Then, a question arises: what is the condition that lets $\begin{bmatrix} -t^{\mathrm{T}}(X^*,G^*) & P(b) \end{bmatrix}$ be normal at $[X^*,G^*]$ to the set $D\times F\cap S$, where $t^{\mathrm{T}}(X^*,G^*) = C + b(X^*,G^*)$?

The supply set S is defined by a set of linear inequalities and thus it is the intersection of a finite number of closed half-spaces. The demand set D and the feasible set F are subset of an affine space. The three sets are compact and the set $D\times F\cap S$ is the intersection of D and S by given F and thus a subset of an affine space. Now suppose that $D\times F\cap S\neq\emptyset$ and let the half spaces of S be $\begin{matrix} H_{s1} & H_{s2} & \cdots & H_{sL} \end{matrix}$, where L is the number of links for the M OD pairs. Each half-space H_{sl} can be expressed by $H_{sl} = \{(X,G) : X_l \le s_l G_l\}$, where G_l is the green-time proportion assigned to the l-th link. The boundaries of S is perpendicular to each other, thus, a $2L\times 1$ vector $n_{sl} = \begin{bmatrix} 0^{\mathrm{T}} & 1 & 0^{\mathrm{T}} & -s_l & 0^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ can be defined, which is non-zero and normal to the half-space H_{sl}. The first 0^{T} is a $1\times l-1$ zero vector, the second $0^{\mathrm{T}}$is a $1\times L-1$ zero vector and the third $0^{\mathbf{T}}$ is a $1\times L-l$ zero vector. The vector n_{sl} is perpendicular to $n_{sl'}$ for $l\neq l'$. Let us define $n_{D\times F}$ as a non-zero vector normal to the set $D\times F$. Recall that $D\times F$ is a subset of an affine space, thus, $n_{D\times F}$ is normal to $D\times F$ at any elements at its boundary.

Because it assumes that $[-C,0]$ is normal at $[X^*,G^*]$ to the set $D\times F\cap S$, then $[-C,0]$ can be expressed by a non-negative combination of n_{sl} and $n_{D\times F}$, that is, $[-C,0] = \sum_{l=1}^{L} w_{sl}n_{sl} + w_{D\times F}n_{D\times F}$, where w_{sl} and $w_{D\times F}$ are non-negative weights. From Definition 5.1, the P0 control policy can maximize the general network capacity at $[X^*,G^*]$, which means that

$[-(C+b), P(b)]$ is normal at $[X^*, G^*]$ to $D \times F$, i.e., $-(C+b)$ is normal at X^* to D and $P(b)$ is normal at G^* to F. If $[-C, 0]$ is perpendicular to the supporting hyperplane of the set $D \times F \cap S$ at $[X^*, G^*]$, then, $[-C, 0]$ is scaled by $n_{D\times F}$ and thus $w_{sl} = 0$ for $l = 1 \cdots L$. Thus, the condition for the general network capacity maximization is that $H\, b^* = k_1 HC$ and $G^* = k_2 s \bullet b^*$, where k_1 and k_2 are the scaling factors, H is an $\sum_{m=1}^{M} N_m \times L$ matrix with $\begin{bmatrix} 0_{(r-1)L_{r-1}} & 1_{L_r} & 0 \end{bmatrix}$ on its r-th row, $s \bullet b^*$ represents the point-wise multiplication between s and b^*.

When $[-C, 0]$ is not perpendicular to the supporting hyperplane of the set $D \times F \cap S$ at $[X^*, G^*]$, $[X^*, G^*]$ must be located at the boundary of S and D because the set $D \times F \cap S$ is a subset of an affine space and $-C$ is a constant vector. Then, $w_{sl} \neq 0$ and $w_{D\times F} n_{D\times F} = [-C, 0] - \sum_{l=1}^{L} w_{sl} n_{sl}$. Let b_l^* be w_{sl}, $\sum_{l=1}^{L} w_{sl} n_{sl}$ becomes $\begin{bmatrix} b^{*\mathrm{T}} & -s^{\mathrm{T}} \bullet b^{*\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ and thus the condition for the general network capacity maximization is that $\begin{bmatrix} -(C+b^*)^{\mathrm{T}} & s^{\mathrm{T}} \bullet b^{*\mathrm{T}} \end{bmatrix}^{\mathrm{T}} = w_{D\times F} n_{D\times F}$. Note that b^* is not normal to S if $[X^*, G^*]$ is located at the interior of the half-spaces H_{sl} but b_l^* is not zero.

The two conditions indicate that by proper construction of b^*, the P0 control policy can maximize the general network capacity. Recall that the bottleneck delay is a non-linear function of equilibrium link flow and green-time proportion vector (5.11), b^* can only be constructed by adjusting the green-time proportion vector. Then, the following lemma can be obtained:

Lemma 5.6. *If the proposed two conditions are not fulfilled, no b^* can be found through adjusting the green-time proportion vector to maximize capacity of the general network using P0 control policy. By adjusting the green-time proportion vector, the equilibrium flow X^* can be maintained as long as $-[C + b(X^*, G^* + \Delta G)]^T = w_D n_D$ and X^* is within the set S for the new green-time vector $G^* + \Delta G$.*

Proof. The normality of the travel cost vector $-(C+b)^{\mathrm{T}}$ to the demand set D can be ensured by adjusting green-time proportion vector to change b, which is however unable to guarantee that $s \bullet b$ is normal to the feasible set F. That means, it is possible to find another G_{m} in F that lets $(s \bullet b)^{\mathbf{T}} G_{\mathrm{m}}$ be larger than $(s \bullet b)^{\mathbf{T}} (G^* + \Delta G)$ and thus $[X^*, G^* + \Delta G]$ may not be the solution to maximize capacity of the general network using P0 control policy. From the definition of D, S and capacity maximization, it is easy to get the condition to keep the equilibrium flow X^* unchanged by adjusting the green-time proportion vector. □

Lemma 5.6 suggests that network capacity cannot always be maximized with P0 control policy by adjusting green-time proportion vector. The prerequisite of capacity maximization is the capability to freely construct bottleneck delay b such as changing the length of queue at

each link exit, which is consistent with Smith's statement. To relieve the condition of capacity maximization by adjusting green-time proportion vector for any $[X^*, G^*]$, at which $[-C, 0]$ is normal to the set $D \times F \cap S$, I propose a novel control policy in this paper, which can be stated as

> *For any X and delay vector b, choose a stage green-time proportion vector G in F to maximize $\sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{l=1}^{N_n} k_{l,n,m} \left(C_{l,n,m} + b_{l,n,m}\right) G_{l,n,m}$, that is, to let $P(b) = k \bullet (C + b)$ be normal at G to F, where k is the scaling vector between the normal vector n_D to D at X and the normal vector n_F to F at G, i.e., $n_F = k \bullet n_D$.*

Because D and F are subset of affine spaces by the definition, n_D and n_F are constant for any X and G on D and F, respectively. Then, k is also constant. The following lemma can be obtained:

Lemma 5.7. *For any solution $[X^*, G^*]$, at which $[-C, 0]$ is normal to the set $D \times F \cap S$, a bottleneck delay $b^* = b(X^*, G^* + \Delta G)$ can be found by adjusting the green-time proportion vector G in F so that the solution $[X^*, G^* + \Delta G, b^*]$ is consistent with the novel control policy if X^* is within set S for the new green-time vector $G^* + \Delta G$ and $-[C + b(X^*, G^* + \Delta G)]^T = w_D n_D$.*

Proof. Let us define a $2L \times 1$ normal vector to the half spaces of S: $H_{s1} \quad H_{s2} \quad \cdots \quad H_{sL}$ by $n_{sl} = \begin{bmatrix} 0^{\mathrm{T}} & b_l & 0^{\mathrm{T}} & -k_l (C + b_l) & 0^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$. By the given condition, $[-C, 0]$ is normal to the set $D \times F \cap S$ at $[X^*, G^* + \Delta G]$ for any ΔG, then, $[-C, 0] = \sum_{l=1}^{L} w_{sl} n_{sl} + w_{D \times F} n_{D \times F}$, where $n_{D \times F}$ must be normal at $[X^*, G^* + \Delta G]$ to $D \times F$ and $\sum_{l=1}^{L} w_{sl} n_{sl}$ must be normal at $[X^*, G^* + \Delta G]$ to S. Then, it has $w_{D \times F} n_{D \times F} = [-C, 0] - \sum_{l=1}^{L} w_{sl} n_{sl}$. Let w_{sl} be one, then $w_{D \times F} n_{D \times F} = \begin{bmatrix} -(C + b(X^*, G^* + \Delta G))^{\mathrm{T}}, \\ k^{\mathrm{T}} \bullet (C + b(X^*, G^* + \Delta G))^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$. To guarantee that $n_{D \times F}$ is normal at $[X^*, G^* + \Delta G]$ to $D \times F$, $-(C + b(X^*, G^* + \Delta G))^{\mathrm{T}}$ should be normal at $[X^*, G^* + \Delta G]$ to D, and $k^{\mathrm{T}} \begin{pmatrix} C + \\ b(X^*, G^* + \Delta G) \end{pmatrix}^{\mathrm{T}}$ should be normal at $[X^*, G^* + \Delta G]$ to F. Because of $n_F = k \bullet n_D$ and the condition for flow maintenance proposed by Lemma 5.6, Lemma 5.7 is proved. □

Lemma 5.7 indicates that any bottleneck delay b, with which $-(C + b)^{\mathrm{T}}$ is normal to D, fulfills that $P(b)$ is also normal to F.

Now I have shown the superiority of the novel control policy over the P0 control policy because it relieves the condition for capacity maximization. Note that Lemma 5.7 states a sufficient condition for determining the equilibrium by adjusting the green-time proportion vector and fixing the equilibrium flow vector with the novel control policy. In next subsection,

I propose a dynamical system, which combines flow swap process, green-time proportion swap process and flow divergence. The proof of the existence, uniqueness and stability of equilibrium will be given.

5.3 A Hybrid Dynamical System

In this subsection, I firstly describe a dynamical system for the proposed general network and then given the proof of equilibrium existence, condition of uniqueness and stability.

5.3.1 Hybrid Dynamical System for the General Network

In this thesis, I apply the proportional switch adjustment process introduced by Smith to our proposed general network. The proportional switch adjustment process (PAP) can be utilized to change flow vector and green-time proportional vector at each time instant. For the case where there are several routes joining a single OD pair [49], the flow change on each route at each time instant can be expressed by

$$\Delta X_{\mathbf{SP}} = \sum_{\{(r:s):r<s\}} w \left\{ \begin{array}{c} X_r \left[C_r(X) - C_s(X)\right]_+ \Delta_{rs} + \\ X_s \left[C_s(X) - C_r(X)\right]_+ \Delta_{sr} \end{array} \right\} \tag{5.30}$$

where r and s represent the r-th and s-th routes for the OD pair, respectively, X_r and X_s represent the flow on the r-th and s-th route at the current time instant, respectively, X is the route flow vector, $C_r(X)$ and $C_s(X)$ represent the travel cost on the r-th and s-th routes with the flow vector X, respectively, $\left[C_r(X) - C_s(X)\right]_+ = \max\left\{\left[C_r(X) - C_s(X)\right], 0\right\}$, $\Delta_{rs} = \left[\begin{array}{ccccc} 0^{\mathrm{T}} & -1 & 0^{\mathrm{T}} & 1 & 0^{\mathrm{T}} \end{array}\right]^{\mathrm{T}}$, $\Delta_{sr} = \left[\begin{array}{ccccc} 0^{\mathrm{T}} & 1 & 0^{\mathrm{T}} & -1 & 0^{\mathrm{T}} \end{array}\right]^{\mathrm{T}}$, w is a positive scaling factor. Note that Δ_{rs} and Δ_{sr} are $N_m \times 1$ vectors, the first, second and third 0^{T} are $r-1 \times 1$ vector, $s-r-1 \times 1$ vector and $N_m - s \times 1$ vector, respectively. Equation (5.30) states that at each time instant, the flow swapping from a higher cost route to a lower cost route is proportional to the produce of flow on the higher cost route and the cost difference between the two routes, the flow swap vector $\Delta X_{\mathbf{SP}}$ is determined by summing the flow swaps between any pair of routes.

In the proposed general network, link flow on the links along the same route is different because of flow diverge at each link exit. The flow swaps between routes denote the input flow re-assignment between any pair of routes for each OD pair. Recall that Wardrop equilibrium $[X^*, G^*]$ can be determined by finding a $[X^*, G^*]$ such that $\left[-t^{\mathrm{T}}(X^*, G^*), 0\right]$ is normal at $[X^*, G^*]$ to the set $D \times F \cap S$ (5.27). In the proposed general network, the $L \times 1$ link flow vector X has the degree of freedom of the number of input links for the M OD pairs, i.e.,

$\sum_{m=1}^{M} N_m$, and $L - \sum_{m=1}^{M} N_m$ elements in X can be determined by the input link flow vector X_{I} and the turning rates between two adjacent links. Therefore, (5.27) can be rewritten by

$$[X_{\mathrm{I}}^*, G^*] = \min_{[X,G]} t^{\mathrm{T}}(X,G)X = \min_{[X_{\mathrm{I}},G]} t_{\mathrm{I}}^{\mathrm{T}}(X_{\mathrm{I}},G)X_{\mathrm{I}} \tag{5.31}$$

where $t_{\mathrm{I}}(X_{\mathrm{I}},G)$ is the $\sum_{m=1}^{M} N_m$ dimensional travel cost vector projected from $t(X,G)$ and can be expressed by

$$\begin{aligned} t_{\mathrm{I}}(X_{\mathrm{I}},G) &= P'^{\mathbf{T}} t(X,G) \\ &= P'^{\mathbf{T}}(C + b(X_{\mathrm{I}},G)) \\ &= P'^{\mathbf{T}} C + P'^{\mathbf{T}} b(X_{\mathrm{I}},G) \\ &= C_{\mathrm{I}} + b_{\mathrm{I}}(X_{\mathrm{I}},G) \end{aligned} \tag{5.32}$$

where P' is the turning rate matrix given by (5.19), C_{I} and $b_{\mathrm{I}}(X_{\mathrm{I}},G)$ are the projected free-travel cost vector and bottleneck delay vector, respectively. Then, $X = P'X_{\mathrm{I}}$ and thus problem (5.27) becomes finding a $[X_{\mathrm{I}}^*, G^*]$ such that $\left[-t_{\mathrm{I}}^{\mathrm{T}}(X_{\mathrm{I}}^*, G^*), 0\right]$ is normal at $[X_{\mathrm{I}}^*, G^*]$ to the set $D \times F \cap S$. Thus, the input flow swap vector for M OD pairs $\Delta X_{m,\mathrm{I}} = \begin{bmatrix} \Delta X_{1,1,m} & \cdots & \Delta X_{1,N_m,m} \end{bmatrix}^{\mathrm{T}}$ can be determined by

$$\Delta X_{m,\mathrm{I}} = \sum_{\{(r:s):r<s\}} w \left\{ \begin{array}{c} X_{1,r,m} \left[\begin{array}{c} t_{r,m,\mathrm{I}}(X_{\mathrm{I}},G) - \\ t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G) \end{array} \right]_+ \\ \times \Delta_{rs} + \Delta_{sr} \times \\ X_{1,s,m} \left[\begin{array}{c} t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G) - \\ t_{r,m,\mathrm{I}}(X_{\mathrm{I}},G) \end{array} \right]_+ \end{array} \right\} = \sum_{\{(r:s):r<s\}} w \left\{ \begin{array}{c} X_{1,r,m} \left[\begin{array}{c} C_{r,m,\mathrm{I}} - C_{s,m,\mathrm{I}} + \\ b_{r,m,\mathrm{I}}(X_{\mathrm{I}},G) - \\ b_{s,m,\mathrm{I}}(X_{\mathrm{I}},G) \end{array} \right]_+ \\ \times \Delta_{rs} + \Delta_{sr} \times \\ X_{1,s,m} \left[\begin{array}{c} C_{s,m,\mathrm{I}} - C_{r,m,\mathrm{I}} + \\ b_{s,m,\mathrm{I}}(X_{\mathrm{I}},G) - \\ b_{r,m,\mathrm{I}}(X_{\mathrm{I}},G) \end{array} \right]_+ \end{array} \right\} \tag{5.33}$$

where $X_{1,r,m}$ and $X_{1,s,m}$ are respectively the input flow on the r-th and s-th route for the m-th OD pair, $C_{r,m,\mathrm{I}}$, $C_{s,m,\mathrm{I}}$, $b_{r,m,\mathrm{I}}$ and $b_{s,m,\mathrm{I}}$ are respectively the free-travel travel costs and bottleneck delay on the r-th and s-th input links for the m-th OD pair. Then the link flow swap vector for

the m-th user can be determined by

$$\Delta X_m = A_m P' \begin{bmatrix} \Delta X_{1,\mathrm{I}} \\ \Delta X_{2,\mathrm{I}} \\ \vdots \\ \Delta X_{M,\mathrm{I}} \end{bmatrix} \tag{5.34}$$

where A_m is the m-th scaling matrix.

To give the green-time proportion adjustment process in the general network, I firstly give $\Delta G_{\mathbf{SP}}$ for the single OD pair case [49]:

$$\Delta G_{\mathrm{SP}} = \sum_{\{(r:s):r<s\}} w \left\{ G_r \begin{bmatrix} [sb(X,G)]_s - \\ [sb(X,G)]_r \end{bmatrix}_+ \Delta_{rs} + G_s \begin{bmatrix} [sb(X,G)]_r - \\ [sb(X,G)]_s \end{bmatrix}_+ \Delta_{sr} \right\} \tag{5.35}$$

where G_r and G_s are green-time proportion vectors on the r-th and s-th route, respectively, $[sb(X,G)]_s$ and $[sb(X,G)]_r$ are stage pressure on the s-th and r-th routes, respectively. Equation (5.35) shows that by P0 control policy, green-time proportion swaps from a lower stage-pressure route to a higher stage-pressure route, and the green-time proportion swap vector is proportional to the product of green-time proportion on the lower stage-pressure route and stage pressure difference between the two routes, ΔG_{SP} is determined by summing the green-time proportion swaps between any pair of routes.

Note that Smith only considered that one signal per route that can be adjusted [49]. In the proposed general network, each route of an OD pair traverses multiple traffic signals, and thus adjustment of each signal can affect equilibrium flow and bottleneck delay on that route. It is hard to directly give the green-time proportion swap vector for each route of each OD pair by (5.35) because the red-green splits of any signals at intersections have impacts on the routes for different OD pairs. Recall that P0 control policy aims at equalizing the stage pressure $P_{n,m} = \sum_{l=1}^{N_{n,m}} s_{l,n,m} b_{l,n,m}$ on each route, i.e., maximizing $\sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{l=1}^{N_n} s_{l,n,m} b_{l,n,m} G_{l,n,m}$. From the feasible set defined by (5.24), adjustment of green-time proportions of traffic signal at different intersections is independent of each other while the green-time proportions for the links incident to the same intersection must sum to one. Thus, the green-time proportion adjustment process in the general network can be realized by simultaneously adjusting the green-time proportion vector for traffic signal at each intersection so that the stage pressure is normal at the adjusted green-time proportion vector to the set $\sum_{m,m'} G_{m,I_{nm,n'm'}} = 1$ for traffic

signal at each intersection:

$$\Delta G_{i,\mathrm{GP}} = \sum_{\{(r:s):r<s\}} w \left\{ G_{r,i} \left[\begin{array}{c} [P(b(X_{\mathrm{I}},G))]_{s,i} - \\ [P(b(X_{\mathrm{I}},G))]_{r,i} \end{array} \right]_{+} \Delta_{rs} + G_{s,i} \left[\begin{array}{c} [P(b(X_{\mathrm{I}},G))]_{r,i} - \\ [P(b(X_{\mathrm{I}},G))]_{s,i} \end{array} \right]_{+} \Delta_{sr} \right\} \tag{5.36}$$

where $i = 1,2,\cdots I$, r and s are the two links incident to the i-th intersection and they may be affiliated with different OD pairs, $[P(b(X_{\mathrm{I}},G))]_{s,i}$ and $[P(b(X_{\mathrm{I}},G))]_{r,i}$ are respectively the stage pressure on the s-th and r-th links incident to the i-th intersection, I is the number of intersections in the general network. With the novel control policy proposed sub-chapter 5.2.3, the stage pressure vector $P(b(X_{\mathrm{I}},G))$ should be changed to $k \bullet (C + b(X_{\mathrm{I}},G))$, the rest are the same as those with P0 control policy.

Let us define ΔX_m and ΔG_m as the link flow swap vector and green-time proportion swap vector for the m-th OD pair obtained by (5.33), (5.34) and (5.36), respectively. Then, the dynamical system can be obtained by

$$\begin{array}{c} [X_m, G_m](t+1) = [X_m, G_m](t) + \lambda(t)\,\Delta[X_m, G_m](t) \\ m = 1 \cdots M \end{array} \tag{5.37}$$

where $\lambda(t)$ is a positive number that represents the step length at the t-th time instant and it is supposed to follow the step length rules proposed in [46].

5.3.2 The Condition for Existence, Uniqueness and Stability of Equilibrium

Let us define the boundary of the set $D \times F \cap S$ as B and express the projected travel cost vector $t_{\mathrm{I}}(X_{\mathrm{I}},G)$ by $t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G)$ for $n_m = 1,2,\cdots,N_m; m = 1,2,\cdots,M$. Then, the following lemma and corollary can be obtained:

Lemma 5.8. *The Wardrop equilibrium of the dynamical system given by (5.37) exists when $t_{n_m,m,I}(X_I,G)$ and $P(b(X_I,G))$ are continuous functions over the set $D \times F \cap S$ for $n_m = 1,2,\cdots,N_m$; $m = 1,2,\cdots,M$ and for any element $Y = [X_I,G]$ on B, it satisfies the condition that $A_{ux,m}\Delta X_{m,I} \le 0$, $m = 1,2,\cdots,M$ and $A_{ug,m}\Delta G_m \le 0$, $m = 1,2,\cdots,M$ for all elements in Y reaching their upper boundaries; while $A_{lx,m}\Delta X_{m,I} \ge 0, m = 1,2,\cdots,M$ and $A_{lg,m}\Delta G_m \ge 0, m = 1,2,\cdots,M$ for all elements in Y reaching their lower boundaries, where $A_{ux,m}$, $A_{ug,m}$, $A_{lx,m}$ and $A_{lg,m}$ are the scaling matrices for the elements in Y reaching their upper and lower boundaries, respectively.*

Proof. Let us rewrite (5.37) as

$$[X_m, G_m](t+1) = f([X_m, G_m](t), \lambda(t)) \qquad (5.38)$$
$$m = 1, 2, \cdots, M$$

Let us define $X_{\mathrm{u},m,\mathrm{I}}$, $G_{\mathrm{u},m}$, $X_{\mathrm{l},m,\mathrm{I}}$ and $G_{\mathrm{l},m}$ as the link flow and green-time proportion vectors of the m-th OD pair containing all elements in an arbitrary vector $Y = [X_{\mathrm{I}}, G]$ on B, which reach their upper and lower boundary, respectively. The rest elements in Y are defined as Y_{r}. Then, from (5.37), the updated $X_{\mathrm{u},m,\mathrm{I}}$ and $G_{\mathrm{u},m}$ will become smaller if $A_{\mathrm{ux},m}\Delta X_{m,\mathrm{I}} \leq 0$ and $A_{\mathrm{ug},m}\Delta G_m \leq 0$; while the updated $X_{\mathrm{l},m,\mathrm{I}}$ and $G_{\mathrm{l},m}$ will become larger if $A_{\mathrm{lx},m}\Delta X_{m,\mathrm{I}} \geq 0$ and $A_{\mathrm{lg},m}\Delta G_m \geq 0$. Because $t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}}, G)$ and $P(b(X, G))$ are continuous functions, there exists a positive number λ that lets the updated Y

$$Y = \begin{bmatrix} X_{\mathrm{u},m,\mathrm{I}} \\ G_{\mathrm{u},m} \\ X_{\mathrm{l},m,\mathrm{I}} \\ G_{\mathrm{l},m} \\ Y_{\mathrm{r}} \end{bmatrix} + \lambda \begin{bmatrix} A_{\mathrm{ux},m}\Delta X_{m,\mathrm{I}} \\ A_{\mathrm{ug},m}\Delta G_m \\ A_{\mathrm{lx},m}\Delta X_{m,\mathrm{I}} \\ A_{\mathrm{lg},m}\Delta G_m \\ \Delta Y_{\mathrm{r}} \end{bmatrix} \qquad (5.39)$$

be located in $D \times F \cap S$. For any vectors on $D \times F \cap S - B$, there exists a positive number λ that lets the updated Y be still located in $D \times F \cap S$ because of the continuity of $t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}}, G)$ and $P(b(X_{\mathbf{I}}, G))$. That means $f : D \times F \cap S \to D \times F \cap S$. The set $D \times F \cap S$ is a compact and convex set. Then, from Brouwer's fixed point theorem, there exists a fixed point $[X_m^*, G_m^*]$ on $D \times F \cap S$ that lets

$$[X_m^*, G_m^*] = f([X_m^*, G_m^*]), m = 1, 2, \cdots, M \qquad (5.40)$$

which means that $\Delta[X_m^*, G_m^*] = 0$, and thus X_m^* is the Wardrop equilibrium. □

Remark 5.9. Note that lemma 5.8 gives the sufficient condition for existence of Wardrop equilibrium, rather than necessary condition. If $t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}}, G)$ and $P(b(X_{\mathrm{I}}, G))$ are not monotone increasing functions for any link flows, there are possibilities that the Wardrop equilibrium exists although the condition is not fulfilled. In that case, the dynamical system becomes unstable in certain regions. The stability will be discussed in the following context.

Corollary 5.10. *If $t_{n_m,m,I}(X_I, G)$ and $P(b(X_I, G))$ are monotone functions of X_I for any link flow and the condition in lemma 5.8 is fulfilled, then X_I^* is unique.*

Proof. Suppose that there is an equilibrium $[X_{\mathrm{I}}^*, G^*]$ in $D \times F \cap S$, which let the travel cost on each input link and the stage pressure on each incident link be equal to each other, that

is, $t_{r,m,\mathrm{I}}(X_{\mathrm{I}}^*, G^*) = t_{s,m,\mathrm{I}}(X_{\mathrm{I}}^*, G^*)$ for any r-th and s-th input links of the m-th OD pair, and $[P(b(X_{\mathrm{I}}, G))]_{s,i} = [P(b(X_{\mathrm{I}}, G))]_{r,i}$ for any r-th and s-th links incident to the i-th intersection. For an arbitrary $[X_{\mathrm{I}}, G] \neq [X_{\mathrm{I}}^*, G^*]$, it has $t_{r,m,\mathrm{I}}(X_{\mathrm{I}}, G) \neq t_{s,m,\mathrm{I}}(X_{\mathrm{I}}, G)$ and $[P(b(X_{\mathrm{I}}, G))]_{s,i} \neq [P(b(X_{\mathrm{I}}, G))]_{r,i}$ because of monotonicity of $t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}}, G)$ and $P(b(X_{\mathrm{I}}, G))$. Then corollary 5.10 is proved. □

Corollary 5.10 gives the condition for uniqueness of the equilibrium for a given G, i.e., a given feasible set. In the proposed dynamical system, G and X change over time and thus there exists a set of equilibria rather than a unique equilibrium even though $t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}}, G)$ and $P(b(X_{\mathrm{I}}, G))$ are monotone functions. Recall that equation (5.31) aims at finding a solution minimizing the sum of product of link flow and link cost. Each equilibrium in the set of equilibria gives a link-flow and link-cost vector, and thus gives different product. Then, the following lemma can be obtained:

Lemma 5.11. *The set of equilibria E is a compact, non-convex set. Any point $[X_I^*, G^*]$ in E satisfies*

$$\frac{\partial^{(K)} t_{k,m,I}(X_I^*, G^*)}{\partial X_I \partial G} = \frac{\partial^{(K)} t_{\ell,m,I}(X_I^*, G^*)}{\partial X_I \partial G} \qquad k \neq \ell, m = 1, 2, \cdots, M \tag{5.41}$$

where K is the number of variables.

Proof. From the definition of Wardrop equilibrium, the vector $\left[-t^{\mathrm{T}}(X_{\mathrm{I}}^*, G^*), 0\right]$ is normal at $[X_{\mathrm{I}}^*, G^*]$ to the set $D \times F \cap S$. Recall that $D \times F \cap S$ is a subset of an affine space. Thus, for each G^*, it has

$$t_{k,m,\mathrm{I}}(X_{\mathrm{I}}^*, G^*) = t_{\ell,m,\mathrm{I}}(X_{\mathrm{I}}^*, G^*) \qquad k \neq \ell, m = 1, 2, \cdots, M \tag{5.42}$$

Because $t_{k,m,\mathrm{I}}(X_{\mathrm{I}}, G)$ is a continuous monotone function on $D \times F \cap S$ for a given G for $k = 1, 2, \cdots, N_m$, and the adjustment of G is a continuous process, the solution space for (5.42) is continuous and a subspace of $D \times F \cap S$, and thus it is compact. Suppose that $\left[X_{1,\mathrm{I}}^*, G_1^*\right]$ and $\left[X_{2,\mathrm{I}}^*, G_2^*\right]$ are any two equilibria in the solution space, then, an arbitrary point on the line segment between the two points can be expressed by $[X_{\mathrm{s,I}}, G_{\mathrm{s}}] = \theta_1 \left[X_{1,\mathrm{I}}^*, G_1^*\right] + (1 - \theta_1) \left[X_{2,\mathrm{I}}^*, G_2^*\right]$ for $0 \leq \theta_1 \leq 1$. It is obvious that $t_{k,m,\mathrm{I}}(X_{\mathrm{s,I}}, G_{\mathrm{s}}) \neq t_{\ell,m,\mathrm{I}}\left(X_{\mathrm{s,I}}^*, G_{\mathrm{s}}^*\right)$ for $k \neq \ell, m = 1, 2, \cdots, M$ because of non-linearity of $t_{k,m,\mathrm{I}}(X_{\mathrm{I}}, G)$. Thus, the solution space is a non-convex set.

The two equilibria $\left[X_{1,\mathrm{I}}^*, G_1^*\right]$ and $\left[X_{2,\mathrm{I}}^*, G_2^*\right]$ satisfy

$$\int_{X_{1,\mathrm{I}}^*}^{X_{2,\mathrm{I}}^*}\int_{G_1^*}^{G_2^*}\left(\frac{\partial^{(K)}t_{k,m,\mathrm{I}}(X_\mathrm{I},G)}{\partial X_\mathrm{I}\partial G}-\frac{\partial^{(K)}t_{\ell,m,\mathrm{I}}(X_\mathrm{I},G)}{\partial X_\mathrm{I}\partial G}\right)\mathrm{d}X_\mathrm{I}\mathrm{d}G=0 \tag{5.43}$$
$$k\neq\ell, m=1,2,\cdots,M$$

where $\int_{X_{1,\mathrm{I}}^*}^{X_{2,\mathrm{I}}^*}\int_{G_1^*}^{G_2^*}$ represents multiple integrals from $\left[X_{1,\mathrm{I}}^*, G_1^*\right]$ to $\left[X_{2,\mathrm{I}}^*, G_2^*\right]$. Let us define $\left[X_{2,\mathrm{I}}^*, G_2^*\right]=\left[X_{1,\mathrm{I}}^*, G_1^*\right]+[\Delta X_\mathrm{I}, \Delta G_1]$, when $[\Delta X_\mathrm{I}, \Delta G_1]$ is close to zero, (5.43) becomes

$$\lim_{[\Delta X_\mathrm{I},\Delta G_1]\to 0}\int_{X_{1,\mathrm{I}}^*}^{X_{1,\mathrm{I}}^*+\Delta X_\mathrm{I}}\int_{G_1^*}^{G_1^*+\Delta G_1}\left(\frac{\partial^{(K)}t_{k,m,\mathrm{I}}(X_\mathrm{I},G)}{\partial X_\mathrm{I}\partial G}-\frac{\partial^{(K)}t_{\ell,m,\mathrm{I}}(X_\mathrm{I},G)}{\partial X_\mathrm{I}\partial G}\right)\mathrm{d}X_\mathrm{I}\mathrm{d}G=0$$
$$k\neq\ell, m=1,2,\cdots,M$$

then lemma 5.11 is proved. □

From the definition in [47], the dynamic system given by (5.37) is stable if and only if, for any initial vector $[X_{0,\mathrm{I}}, G_0]\in D\times F\cap S$, the solution of (5.37) with the initial vector $[X_{0,\mathrm{I}}, G_0]$ converges as $t\to\infty$, to the set of equilibria E. To prove the stability of the dynamical system, let us define a function V by

$$V([X_\mathrm{I}, G])=\sum_{m=1}^{M}V_{X,m}([X_\mathrm{I}, G])+\sum_{i=1}^{I}V_{G,i}([X_\mathrm{I}, G]) \tag{5.44}$$

where

$$V_{X,m}([X_\mathrm{I}, G])=\sum_{r,s=1,r\neq s}^{N_m}X_{1,r,m}\left(t_{r,m,\mathrm{I}}(X_\mathrm{I},G)-t_{s,m,\mathrm{I}}(X_\mathrm{I},G)\right)_+^2=C_{\mathrm{d},m,\mathrm{I}}^{\mathrm{T}}(X_\mathrm{I},G)X_{m,\mathrm{I}}$$

$$V_{G,i}([X_\mathrm{I}, G])=\sum_{r,s=1,r\neq s}^{N_i}G_{r,i}\left([P(b(X,G))]_{s,i}-[P(b(X,G))]_{r,i}\right)_+^2=P_{\mathrm{d},i}^{\mathrm{T}}(X,G)G_i$$

where N_m and N_i are respectively the number of routes for the m-th OD pair and the number of links incident to the i-th intersection, $X_{m,\mathrm{I}}=\begin{bmatrix}X_{1,1,m} & X_{1,2,m} & \cdots & X_{1,N_m,m}\end{bmatrix}^{\mathrm{T}}$, $G_i=\begin{bmatrix}G_{1,i} & G_{2,i} & \cdots & G_{N_i,i}\end{bmatrix}^{\mathrm{T}}$,

$$C_{\mathrm{d},m,\mathrm{I}}(X_{\mathrm{I}},G)=\left[\begin{array}{ccc}\sum_{s=2}^{N_m}\left(\begin{array}{c}t_{1,m,\mathrm{I}}(X_{\mathrm{I}},G)-\\ t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{array}\right)_+^2 & \cdots & \sum_{s=1,s\neq N_m}^{N_m}\left(\begin{array}{c}t_{N_m,m,\mathrm{I}}(X_{\mathrm{I}},G)-\\ t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{array}\right)_+^2\end{array}\right]^{\mathrm{T}} \tag{5.45}$$

$$P_{\mathrm{d},i}(X,G)=\left[\begin{array}{ccc}\sum_{s=2}^{N_i}\left(\begin{array}{c}[P(b(X_{\mathrm{I}},G))]_{s,i}-\\ [P(b(X_{\mathrm{I}},G))]_{1,i}\end{array}\right)_+^2 & \cdots & \sum_{s=1,s\neq N_i}^{N_i}\left(\begin{array}{c}[P(b(X_{\mathrm{I}},G))]_{s,i}-\\ [P(b(X_{\mathrm{I}},G))]_{N_i,i}\end{array}\right)_+^2\end{array}\right]^{\mathrm{T}} \tag{5.46}$$

It is obvious that $V([X_{\mathrm{I}},G])$=0 if and only if $[X_{\mathrm{I}},G]$ is an equilibrium and $V([X_{\mathrm{I}},G])>0$ for all $[X_{\mathrm{I}},G]\in D\times F\cap S\backslash E$, where $A\backslash B$ represents the relative complement of A with respect to B. Now we investigate monotonicity of $V([X_{\mathrm{I}},G])$ with t. The time derivative of $V([X_{\mathrm{I}},G])$ can be expressed by

$$\begin{aligned}\frac{dV([X_{\mathrm{I}},G])}{dt}&=\frac{\partial V([X_{\mathrm{I}},G])}{\partial[X_{\mathrm{I}},G]}\frac{d[X_{\mathrm{I}},G]}{dt}=\frac{\partial V([X_{\mathrm{I}},G])}{\partial X_{\mathrm{I}}}\frac{dX_{\mathrm{I}}}{dt}+\frac{\partial V([X_{\mathrm{I}},G])}{\partial G}\frac{dG}{dt}\\&=\sum_{m=1}^{M}\left(\sum_{n=1}^{M}\frac{\partial V_{X,n}([X_{\mathrm{I}},G])}{\partial X_{m,\mathrm{I}}}+\sum_{i=1}^{I}\frac{\partial V_{G,i}([X_{\mathrm{I}},G])}{\partial X_{m,\mathrm{I}}}\right)\frac{dX_{m,\mathrm{I}}}{dt}\\&+\sum_{i=1}^{I}\left(\sum_{j=1}^{I}\frac{\partial V_{G,j}([X_{\mathrm{I}},G])}{\partial G_i}+\sum_{n=1}^{M}\frac{\partial V_{X,n}([X_{\mathrm{I}},G])}{\partial G_i}\right)\frac{dG_i}{dt}\\&=\sum_{m=1}^{M}\left(C_{\mathrm{d},m,\mathrm{I}}^{\mathrm{T}}(X_{\mathrm{I}},G)+\sum_{n=1}^{M}X_{n,\mathrm{I}}^{\mathrm{T}}\frac{\partial C_{\mathrm{d},n,\mathrm{I}}(X_{\mathrm{I}},G)}{\partial X_{m,\mathrm{I}}}+\sum_{i=1}^{I}G_i^{\mathrm{T}}\frac{\partial P_{\mathrm{d},i}(X_{\mathrm{I}},G)}{\partial X_{m,\mathrm{I}}}\right)\frac{dX_{m,\mathrm{I}}}{dt}\\&+\sum_{i=1}^{I}\left(P_{\mathrm{d},i}^{\mathrm{T}}(X_{\mathrm{I}},G)+\sum_{n=1}^{M}X_{n,\mathrm{I}}^{\mathrm{T}}\frac{\partial C_{\mathrm{d},n,\mathrm{I}}^{\mathrm{T}}(X_{\mathrm{I}},G)}{\partial G_i}+\sum_{j=1}^{I}G_j^{\mathrm{T}}\frac{\partial P_{\mathrm{d},j}(X_{\mathrm{I}},G)}{\partial G_i}\right)\frac{dG_i}{dt}\end{aligned} \tag{5.47}$$

By (5.45) and (5.46), $X_{n,\mathrm{I}}^{\mathrm{T}}\frac{\partial C_{\mathrm{d},n,\mathrm{I}}(X_{\mathrm{I}},G)}{\partial X_{m,\mathrm{I}}}$, $G_i^{\mathrm{T}}\frac{\partial P_{\mathrm{d},i}(X_{\mathrm{I}},G)}{\partial X_{m,\mathrm{I}}}$, $X_{n,\mathrm{I}}^{\mathrm{T}}\frac{\partial C_{\mathrm{d},n,\mathrm{I}}^{\mathrm{T}}(X_{\mathrm{I}},G)}{\partial G_i}$ and $G_j^{\mathrm{T}}\frac{\partial P_{\mathrm{d},j}(X_{\mathrm{I}},G)}{\partial G_i}$ can be determined by

$$X_{n,\mathrm{I}}^{\mathrm{T}}\frac{\partial C_{\mathrm{d},n,\mathrm{I}}(X_{\mathrm{I}},G)}{\partial X_{m,\mathrm{I}}}=\left[\begin{array}{cccc}CX_{\mathrm{d},1} & CX_{\mathrm{d},2} & \cdots & CX_{\mathrm{d},N_m}\end{array}\right]=-\frac{2}{w}\Delta X_{n,\mathrm{I}}^{\mathbf{T}}J_{\mathrm{cx},nm}$$

$$CX_{\mathrm{d},n_m} = 2\sum_{\ell=1}^{N_n} X_{1,\ell,n} \sum_{s=1,s\neq\ell}^{N_n} \left(\; t_{\ell,n,\mathrm{I}}\left(X_\mathrm{I},G\right) - t_{s,n,\mathrm{I}}\left(X_\mathrm{I},G\right) \; \right)_+ \left(\; \frac{\partial t_{\ell,n,\mathrm{I}}(X_\mathrm{I},G)}{\partial X_{1,n_m,m}} - \frac{\partial t_{s,n,\mathrm{I}}(X_\mathrm{I},G)}{\partial X_{1,n_m,m}} \; \right)$$

$$G_i^\mathrm{T} \frac{\partial P_{\mathrm{d},i}\left(X_\mathrm{I},G\right)}{\partial X_{m,\mathrm{I}}} = \left[\begin{array}{cccc} PX_{\mathrm{d},1} & PX_{\mathrm{d},2} & \cdots & PX_{\mathrm{d},N_m} \end{array} \right] = \frac{2}{w} \Delta G_{i,\mathrm{GP}}^\mathrm{T} J_{\mathrm{px},im}$$

$$PX_{\mathrm{d},n_m} = 2\sum_{\ell=1}^{N_i} G_{\ell,i} \sum_{s=1,s\neq\ell}^{N_i} \left(\; \left[P\left(b\left(X_\mathrm{I},G\right)\right)\right]_{s,i} - \left[P\left(b\left(X_\mathrm{I},G\right)\right)\right]_{\ell,i} \; \right)_+ \left(\begin{array}{c} P'\left(b_{s,i}\left(X_\mathrm{I},G\right)\right) \frac{\partial b_{s,i}(X_\mathrm{I},G)}{\partial X_{1,n_m,m}} \\ -P'\left(b_{\ell,i}\left(X_\mathrm{I},G\right)\right) \frac{\partial b_{\ell,i}(X_\mathrm{I},G)}{\partial X_{1,n_m,m}} \end{array} \right)$$

$$X_{m,\mathrm{I}}^\mathrm{T} \frac{\partial C_{\mathrm{d},m,\mathrm{I}}^\mathrm{T}\left(X_\mathrm{I},G\right)}{\partial G_i} = \left[\begin{array}{cccc} CG_{\mathrm{d},1} & CG_{\mathrm{d},2} & \cdots & CG_{\mathrm{d},N_i} \end{array} \right] = -\frac{2}{w} \Delta X_{m,\mathrm{I}}^{\mathbf{T}} J_{\mathrm{cg},mi}$$

$$CG_{\mathrm{d},n_i} = 2\sum_{\ell=1}^{N_m} X_{1,\ell,m} \sum_{s=1,s\neq\ell}^{N_m} \left(\; t_{\ell,m,\mathrm{I}}\left(X_\mathrm{I},G\right) - t_{s,m,\mathrm{I}}\left(X_\mathrm{I},G\right) \; \right)_+ \left(\; \frac{\partial t_{\ell,m,\mathrm{I}}(X_\mathrm{I},G)}{\partial G_{n_i,i}} - \frac{\partial t_{s,m,\mathrm{I}}(X_\mathrm{I},G)}{\partial G_{n_i,i}} \; \right)$$

$$G_j^\mathrm{T} \frac{\partial P_{\mathrm{d},j}\left(X_\mathrm{I},G\right)}{\partial G_i} = \left[\begin{array}{cccc} PG_{\mathrm{d},1} & PG_{\mathrm{d},2} & \cdots & PG_{\mathrm{d},N_i} \end{array} \right] = \frac{2}{w} \Delta G_{i,\mathrm{GP}}^\mathrm{T} J_{\mathrm{pg},ji}$$

$$PG_{\mathrm{d},n_i} = 2\sum_{\ell=1}^{N_j} G_{\ell,j} \sum_{s=1,s\neq\ell}^{N_j} \left(\; \left[P\left(b\left(X_\mathrm{I},G\right)\right)\right]_{s,j} - \left[P\left(b\left(X_\mathrm{I},G\right)\right)\right]_{\ell,j} \; \right)_+ \left(\begin{array}{c} P'\left(b_{s,j}\left(X_\mathrm{I},G\right)\right) \frac{\partial b_{s,j}(X_\mathrm{I},G)}{\partial G_{n_j,i}} \\ -P'\left(b_{\ell,j}\left(X_\mathrm{I},G\right)\right) \frac{\partial b_{\ell,j}(X_\mathrm{I},G)}{\partial G_{n_j,i}} \end{array} \right) \tag{5.48}$$

where $J_{\mathrm{cx},nm}$, $J_{\mathrm{px},im}$, $J_{\mathrm{cg},mi}$ and $J_{\mathrm{pg},ji}$ are the Jacobian matrices of $t_{\mathrm{I},n}\left(X_\mathrm{I},G\right)$, $P_i\left(b\left(X_\mathrm{I},G\right)\right)$, $t_{\mathrm{I},m}\left(X_\mathrm{I},G\right)$ and $P_j\left(b\left(X_\mathrm{I},G\right)\right)$ evaluated at $X_{m,\mathrm{I}}$, $X_{m,\mathrm{I}}$, G_i and G_i, respectively. Thus,

$$\begin{aligned} \left(\sum_{n=1}^{M} \frac{\partial V_{X,n}\left(\left[X_\mathrm{I},G\right]\right)}{\partial X_{m,\mathrm{I}}} + \sum_{i=1}^{I} \frac{\partial V_{G,i}\left(\left[X_\mathrm{I},G\right]\right)}{\partial X_{m,\mathrm{I}}} \right) \frac{dX_{m,\mathrm{I}}}{dt} &= -\frac{2}{w} \sum_{n=1}^{M} \Delta X_{n,\mathrm{I}}^{\mathbf{T}} J_{\mathrm{cx},nm} \Delta X_{m,\mathrm{I}} + C_{\mathrm{d},m,\mathrm{I}}^\mathrm{T}\left(X_\mathrm{I},G\right) \Delta X_{m,\mathrm{I}} \\ + \frac{2}{w} \sum_{i=1}^{I} \Delta G_{i,\mathrm{GP}}^\mathrm{T} J_{\mathrm{px},im} \Delta X_{m,\mathrm{I}} &= C_{\mathrm{d},m,\mathrm{I}}^\mathrm{T}\left(X_\mathrm{I},G\right) \Delta X_{m,\mathrm{I}} - \frac{2}{w} \Delta X^{\mathbf{T}} \sum_{n=1}^{M} A_{n,\mathrm{I}}^\mathrm{T} J_{\mathrm{cx},nm} A_{m,\mathrm{I}} \Delta X \\ &\quad + \frac{2}{w} \Delta G^\mathrm{T} \sum_{i=1}^{I} B_i^\mathrm{T} J_{\mathrm{px},im} A_{m,\mathrm{I}} \Delta X \end{aligned} \tag{5.49}$$

and

$$\left(\sum_{j=1}^{I}\frac{\partial V_{G,j}\left([X_{\mathrm{I}},G]\right)}{\partial G_i}+\sum_{n=1}^{M}\frac{\partial V_{X,n}\left([X_{\mathrm{I}},G]\right)}{\partial G_i}\right)\frac{dG_i}{dt}=-\frac{2}{w}\sum_{n=1}^{M}\Delta X_{n,\mathrm{I}}^{\mathbf{T}}J_{\mathrm{cg},ni}\Delta G_{i,\mathrm{GP}}+P_{\mathrm{d},i}^{\mathrm{T}}\left(X,G\right)\Delta G_{i,\mathrm{GP}}$$
$$+\frac{2}{w}\sum_{j=1}^{I}\Delta G_{j,\mathrm{GP}}^{\mathrm{T}}J_{\mathrm{pg},ji}\Delta G_{i,\mathrm{GP}}=P_{\mathrm{d},i}^{\mathrm{T}}\left(X,G\right)\Delta G_{i,\mathrm{GP}}-\frac{2}{w}\Delta X^{\mathbf{T}}\sum_{n=1}^{M}A_{n,\mathrm{I}}^{\mathrm{T}}J_{\mathrm{cg},ni}B_i\Delta G$$
$$+\frac{2}{w}\Delta G^{\mathrm{T}}\sum_{j=1}^{I}B_j^{\mathrm{T}}J_{\mathrm{pg},ji}B_i\Delta G \quad (5.50)$$

where $A_{m,\mathrm{I}}$ and B_i are the scaling matrices for input links of the m-th OD pair and the i-th intersection, respectively. Combined with (5.49) and (5.50), (5.48) can be rewritten by

$$\frac{dV\left([X_{\mathrm{I}},G]\right)}{dt}=\sum_{m=1}^{M}C_{\mathrm{d},m,\mathrm{I}}^{\mathrm{T}}\left(X_{\mathrm{I}},G\right)\Delta X_{m,\mathrm{I}}+\sum_{i=1}^{I}P_{\mathrm{d},i}^{\mathrm{T}}\left(X_{\mathrm{I}},G\right)\Delta G_{i,\mathrm{GP}}-\frac{2}{w}\Delta X^{\mathbf{T}}\sum_{m=1}^{M}\sum_{n=1}^{M}A_{n,\mathrm{I}}^{\mathrm{T}}J_{\mathrm{cx},nm}A_{m,\mathrm{I}}\Delta X$$
$$+\frac{2}{w}\Delta G^{\mathrm{T}}\sum_{i=1}^{I}\sum_{j=1}^{I}B_j^{\mathrm{T}}J_{\mathrm{pg},ji}B_i\Delta G+\frac{2}{w}\Delta X^{\mathbf{T}}\sum_{n=1}^{M}\sum_{i=1}^{I}A_{n,\mathrm{I}}^{\mathrm{T}}\left(J_{\mathrm{px},in}^{\mathbf{T}}-J_{\mathrm{cg},ni}\right)B_i\Delta G \quad (5.51)$$

The stage pressure function $P_i\left(b\left(X_{\mathrm{I}},G\right)\right)$ is a non-increasing monotone function of G, $J_{\mathrm{pg}}=\sum_{i=1}^{I}\sum_{j=1}^{I}B_j^{\mathrm{T}}J_{\mathrm{pg},ji}B_i$ is a diagonal matrix, and thus it is a negative semi-definite matrix. The travel cost function $t_{n,\mathrm{I}}\left(X_{\mathrm{I}},G\right)$ is non-decreasing monotone function of X_{I}. However, it is clear that $\sum_{m=1}^{M}\sum_{n=1}^{M}A_{n,\mathrm{I}}^{\mathrm{T}}J_{\mathrm{cx},nm}A_{m,\mathrm{I}}$ is non-symmetric and thus it is cumbersome to state that $\sum_{m=1}^{M}\sum_{n=1}^{M}A_{n,\mathrm{I}}^{\mathrm{T}}J_{\mathrm{cx},nm}A_{m,\mathrm{I}}$ is a positive semi-definite matrix. The following theorem is proposed to determine the matrix definiteness.

Theorem 5.12. *A square matrix M is positive (or negative) semi-definite over a given compact space S, i.e., $x^TMx\geq 0$ $\left(or\, x^TMx\leq 0\right)$ for any $x\in S$ iff (if and only if) M is a full-rank matrix and for any element $x_B\in B_S$, where B_S is the boundary of S and satisfies $x_B^TMx_B\geq 0$ $\left(or\, x_B^TMx_B\leq 0\right)$.*

Proof. Let us define $y=x^{\mathrm{T}}Mx$ for $x\in S$. Then, the local minimum (or maximum) point x_{m} can be determined by

$$\frac{\partial y}{\partial x}=x^{\mathrm{T}}M=0$$

When M is a full-rank matrix, $x_{\mathrm{m}}=0$ becomes the global minimum (or maximum). Then, theorem 5.12 can be straightly obtained. □

From (5.32), (5.34) and (5.36), ΔX fulfills $1^{\mathbf{T}}\Delta X_{m,\mathrm{I}} = 0$, $-T_m \leq \Delta X_{l,m,\mathrm{I}} \leq T_m$. Let us define $S_{\Delta X}$ and $S_{\Delta G}$ as the compact space for ΔX and ΔG, respectively, $B_{\Delta X}$ and $B_{\Delta G}$ as their boundaries, respectively. Then, with theorem 5.12, the following corollary can be obtained:

Corollary 5.13. *The matrix* $\sum_{m=1}^{M}\sum_{n=1}^{M} A_{n,I}^{T} J_{cx,nm} A_{m,I}$ *is positive semi-definite over* $S_{\Delta X}$.

Proof. For simplicity of expression, let us express $\sum_{m=1}^{M}\sum_{n=1}^{M} A_{n,\mathrm{I}}^{\mathrm{T}} J_{\mathrm{cx},nm} A_{m,\mathrm{I}}$ by J_{cx}. Then, $\Delta X^{\mathbf{T}} J_{\mathrm{cx}} \Delta X = \Delta X_{\mathbf{I}}^{\mathbf{T}} J_{\mathrm{cx,I}} \Delta X_{\mathbf{I}}$, where $J_{\mathrm{cx,I}}$ is the $\sum_{m=1}^{M} N_m \times \sum_{m=1}^{M} N_m$ Jacobian matrix. From (5.32) and (5.34), it can be known that $J_{\mathrm{cx,I}}$ is a full-rank matrix, and the element on the k-th row and ℓ-th column can be expressed by

$$J_{k\ell,\mathrm{cx,I}} = \frac{\partial t_{k,\mathrm{I}}(X_{\mathrm{I}}, G)}{\partial X_{\ell,\mathrm{I}}} = \frac{\partial \left(P_k'^{\mathrm{T}} C + P_k'^{\mathrm{T}} b\left(P' X_{\mathrm{I}}, G\right)\right)}{\partial X_{\ell,\mathrm{I}}} = P_k'^{\mathrm{T}} \frac{\partial b\left(P' X_{\mathrm{I}}, G\right)}{\partial X_{\ell,\mathrm{I}}} \tag{5.52}$$

The space $S_{\Delta X}$ is a subset of affine space. The vertices on $B_{\Delta X}$ are $\sum_{m=1}^{M} N_m \times 1$ vectors $\begin{bmatrix} 0 & -T_m & 0 & T_m & 0 \end{bmatrix}^{\mathbf{T}}$, where the length of first 0, second 0 and third 0 are respectively $\sum_{m'=1}^{m-1} N_{m'} + k_{1,m}$, $k_{2,m}$ and $\sum_{m'=m+1}^{M} N_{m'} + k_{3,m}$, and $N_m = 2 + k_{1,m} + k_{2,m} + k_{3,m}$. There are $\sum_{m=1}^{M} \begin{pmatrix} 2 \\ N_m \end{pmatrix}$ vertices on $B_{\Delta X}$. Furthermore, for each vertex ΔX_v, it can be shown that $\Delta X_v^{\mathbf{T}} J_{\mathrm{cx,I}} \Delta X_v \geq 0$ from (5.52) and the fact that $t_{n,\mathrm{I}}(X_{\mathrm{I}}, G)$ is a non-decreasing monotone function of X_{I}. Then for each $x_B \in B_{\Delta X}$, it has $x_B^{\mathrm{T}} M x_B \geq 0$, and thus from theorem 5.12, corollary 5.13 is proved.

Corollary 5.13 shows that the third term in (5.51) is negative, which leads to the following corollary that shows the summation of last three terms in (5.51) to be also negative. Let us express $\sum_{n=1}^{M}\sum_{i=1}^{I} A_{n,\mathrm{I}}^{\mathrm{T}} \left(J_{\mathrm{px},in}^{\mathbf{T}} - J_{\mathrm{cg},ni} \right) B_i$ by J_{xg}, then, the following corollary can be obtained: □

Corollary 5.14. *The derivative* $dV\left([X_I, G]\right)/dt$ *fulfills the following inequality*

$$\frac{dV\left([X_I, G]\right)}{dt} \leq \sum_{m=1}^{M} C_{d,m,I}^{T}(X_I, G)\,\Delta X_{m,I} + \sum_{i=1}^{I} P_{d,i}^{T}(X_I, G)\Delta G_{i,GP} \tag{5.53}$$

Proof. Equation (5.51) can be rewritten by

$$\frac{dV([X_{\mathrm{I}},G])}{dt}=\sum_{m=1}^{M}C_{\mathrm{d},m,\mathrm{I}}^{\mathrm{T}}(X_{\mathrm{I}},G)\,\Delta X_{m,\mathrm{I}}+\sum_{i=1}^{I}P_{\mathrm{d},i}^{\mathrm{T}}(X_{\mathrm{I}},G)\Delta G_{i,\mathrm{GP}}$$
$$-\frac{2}{w}\left(\Delta X^{\mathbf{T}}J_{\mathrm{cx}}^{'}\Delta X+\Delta G^{\mathrm{T}}J_{\mathrm{pg}}^{'}\Delta G-\Delta X^{\mathbf{T}}J_{\mathrm{xg}}\Delta G\right) \tag{5.54}$$

where $J_{\mathrm{pg}}^{'}=-J_{\mathrm{pg}}$ and $J_{\mathrm{cx}}^{'}=\frac{1}{2}\left(J_{\mathrm{cx}}+J_{\mathrm{cx}}^{\mathbf{T}}\right)$. Recall that J_{pg} is a negative semi-definite matrix and J_{cx} is a non-symmetric positive semi-definite matrix over $S_{\Delta X}$. Thus, $J_{\mathrm{pg}}^{'}$ and $J_{\mathrm{cx}}^{'}$ are positive semi-definite matrices. From Cauchy-Schwarz inequality, it has

$$|\Delta X^{\mathbf{T}}J_{\mathrm{xg}}\Delta G|\leq\sqrt{\Delta X^{\mathbf{T}}H_1^{\mathrm{T}}H_1\Delta X}\sqrt{\Delta G^{\mathbf{T}}H_2^{\mathrm{T}}H_2\Delta G} \tag{5.55}$$

where $J_{\mathrm{xg}}=H_1^{\mathrm{T}}H_2$. Then

$$\left(\Delta X^{\mathbf{T}}J_{\mathrm{cx}}^{'}\Delta X+\Delta G^{\mathrm{T}}J_{\mathrm{pg}}^{'}\Delta G\right)^2=\left(\Delta X^{\mathbf{T}}J_{\mathrm{cx}}^{'}\Delta X\right)^2+\left(\Delta G^{\mathrm{T}}J_{\mathrm{pg}}^{'}\Delta G\right)^2+2\Delta X^{\mathbf{T}}J_{\mathrm{cx}}^{'}\Delta X\Delta G^{\mathrm{T}}J_{\mathrm{pg}}^{'}\Delta G \tag{5.56}$$

let $H_2=J_{\mathrm{pg}}^{'\frac{1}{2}}$, which is a diagonal matrix. Furthermore, $H_1^{\mathbf{T}}$ in (5.55) becomes $H_1^{\mathrm{T}}=J_{\mathrm{xg}}J_{\mathrm{pg}}^{'-\frac{1}{2}}$. Then, (5.55) can be transformed to

$$\left(\Delta X^{\mathbf{T}}J_{\mathrm{xg}}\Delta G\right)^2\leq\Delta X^{\mathbf{T}}J_{\mathrm{xg}}J_{\mathrm{pg}}^{'-1}J_{\mathrm{xg}}^{\mathrm{T}}\Delta X\Delta G^{\mathbf{T}}J_{\mathrm{pg}}^{'}\Delta G \tag{5.57}$$

and it has

$$2\Delta X^{\mathbf{T}}J_{\mathrm{cx}}^{'}\Delta X\Delta G^{\mathrm{T}}J_{\mathrm{pg}}^{'}\Delta G-\Delta X^{\mathbf{T}}J_{\mathrm{xg}}J_{\mathrm{pg}}^{'-1}J_{\mathrm{xg}}^{\mathrm{T}}\Delta X\Delta G^{\mathbf{T}}J_{\mathrm{pg}}^{'}\Delta G=$$
$$Tr\left(\Delta G\Delta X^{\mathbf{T}}\left(2J_{\mathrm{cx}}^{'}\Delta X\Delta G^{\mathrm{T}}J_{\mathrm{pg}}^{'}-J_{\mathrm{xg}}J_{\mathrm{pg}}^{'-1}J_{\mathrm{xg}}^{\mathrm{T}}\Delta X\Delta G^{\mathbf{T}}J_{\mathrm{pg}}^{'}\right)\right)=$$
$$Tr\left(\Delta G\Delta X^{\mathbf{T}}\left(2J_{\mathrm{cx}}^{'}-J_{\mathrm{xg}}J_{\mathrm{pg}}^{'-1}J_{\mathrm{xg}}^{\mathrm{T}}\right)\Delta X\Delta G^{\mathbf{T}}J_{\mathrm{pg}}^{'}\right) \tag{5.58}$$

The matrix $\Delta X\Delta G^{\mathbf{T}}$ in (5.58) is with the rank of one, and from the definition of travel cost function and stage pressure vector, $J_{\mathrm{pg}}^{'}\left(2J_{\mathrm{cx}}^{'}-J_{\mathrm{xg}}J_{\mathrm{pg}}^{'-1}J_{\mathrm{xg}}^{\mathrm{T}}\right)$ can be shown to be a positive semi-definite matrix, and thus (5.58) is a non-negative number. From (5.55) to (5.58), the last term in (5.54) is a non-positive number and thus corollary 5.14 is proved. □

Then, the following lemma can be obtained:

Lemma 5.15. *The derivative* $\frac{dV([X_I,G])}{dt}<0$ *for all* $[X_I,G]\in D\times F\cap S\backslash E$

Proof. From the definition of the input flow swap vector (5.32) and the green-time proportion swap vector (5.36), (5.53) can be transformed to

$$\begin{aligned}\frac{dV([X_{\mathrm{I}},G])}{dt} \leq w\sum_{m=1}^{M}\sum_{n_m=1}^{N_m}\sum_{s=1,s\neq n_m}^{N_m}\begin{pmatrix} t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G) \\ -t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{pmatrix}_+^2\Bigg(\sum_{s=1,s\neq n_m}^{N_m}(-X_{1,n_m,m})\begin{pmatrix} t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G)- \\ t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{pmatrix}_+ \\ +X_{1,s,m}\begin{pmatrix} t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)- \\ t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{pmatrix}_+\Bigg)+w\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\sum_{s=1,s\neq n_i}^{N_i}\begin{pmatrix}[P(b(X_{\mathrm{I}},G))]_{s,i}- \\ [P(b(X_{\mathrm{I}},G))]_{n_i,i}\end{pmatrix}_+^2 \\ \times\Bigg(\sum_{s=1,s\neq n_i}^{N_i}(-G_{n_i,i})\begin{pmatrix}[P(b(X_{\mathrm{I}},G))]_{s,i} \\ -[P(b(X_{\mathrm{I}},G))]_{n_i,i}\end{pmatrix}_+ +G_{s,i}\begin{pmatrix}[P(b(X_{\mathrm{I}},G))]_{n_i,i} \\ -[P(b(X_{\mathrm{I}},G))]_{s,i}\end{pmatrix}_+\Bigg)\end{aligned}$$

Because of

$$\begin{aligned}0 &\leq \sum_{n_m=1}^{N_m}\sum_{s=1,s\neq n_m}^{N_m}\begin{pmatrix} t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G) \\ -t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{pmatrix}_+^2\sum_{s=1,s\neq n_m}^{N_m}X_{1,s,m}\begin{pmatrix} t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)- \\ t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{pmatrix}_+ \\ &\leq \sum_{n_m=1}^{N_m}\sum_{s=1,s\neq n_m}^{N_m}\begin{pmatrix} t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G) \\ -t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{pmatrix}_+^2\sum_{s=1,s\neq n_m}^{N_m}X_{1,n_m,m}\begin{pmatrix} t_{n_m,m,\mathrm{I}}(X_{\mathrm{I}},G)- \\ t_{s,m,\mathrm{I}}(X_{\mathrm{I}},G)\end{pmatrix}_+\end{aligned}$$

and similarity for the stage pressure function, Lemma 5.15 is proved. □

Lemma 5.15 and the fact that $V([X_{\mathrm{I}},G])$=0 if and only if $[X_{\mathrm{I}},G]$ is an equilibrium and $V([X_{\mathrm{I}},G])>0$ for all $[X_{\mathrm{I}},G]\in D\times F\cap S\backslash E$ indicate that given any initial vector $\left[X_{\mathrm{I}}^0,G^0\right]\in D\times F\cap S$, the dynamical system given by (5.37) can converge to an equilibrium on E as time goes to infinity by using the novel control policy, or the P0 control policy if the two conditions proposed in the subsection **??** are satisfied.

5.4 Numerical Results

The proposed dynamical system given by (5.37) is tested on an one-OD two-route network illustrated by Fig. 1.1 and a two-OD two-route network illustrated by Fig. 1.2.

In order to validate the superiority of my proposed novel control policy over the P0 control policy, It starts from an arbitrary feasible point fulfilling the condition given by (5.23), (5.24) and (5.25), set the stage pressure vector to $P(b)=k\bullet(C+b)$ and $P(b)=s\bullet b$, respectively, and observe if the dynamical system can converge to a feasible point as time goes to infinity. In order to validate the condition of existence of equilibrium point proposed by Lemma 5.8, it

starts from an arbitrary feasible point for the cases that the boundary condition in Lemma 5.8 is satisfied or not, and compare the output of the dynamical system.

For the one-OD two-route network, the length of two routes and flow of two routes are set to be 0.5km, 1km and 100 vehicles/h, 150 vehicles/h, respectively. The free speed of both routes, the total demand and the duration of traffic light cycle are set to be 50km/h, 100 vehicles/h and 120 seconds, respectively. It starts from an arbitrary point $XV_0 = [\ 58.33 \quad 41.67\ ]$, $GV_0 = \begin{bmatrix} 0.67 & 0.33 \end{bmatrix}$ and sets the step length by the rules proposed in [46]. The output of the dynamical system is illustrated by Fig. 5.2 and Fig. 5.3, which show the time evolution of traffic flow and green-time proportion on two routes, respectively, when the novel control policy is used. The dynamical system is stable as the equilibrium point can be approached after approximately 1000 iterations. Fig. 5.4 and Fig. 5.5 illustrate the output of the dynamical system by using the P0 control policy. It can be seen that traffic flow and green-time proportion cannot approach a feasible equilibrium point within the first 180 iterations, and after the 180th iteration, the solution of system is no more feasible. Fig. 5.6 and 5.7 show that even start from a feasible point, no feasible equilibrium can be found if the condition of existence of equilibrium proposed by Lemma 5.8 is not satisfied.

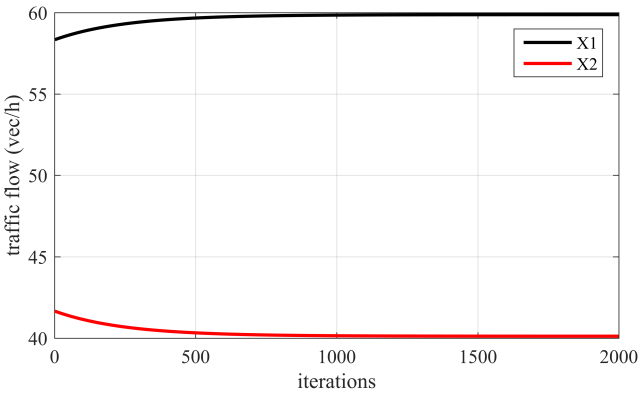

Figure 5.2: Time evolution of traffic flow using novel control policy

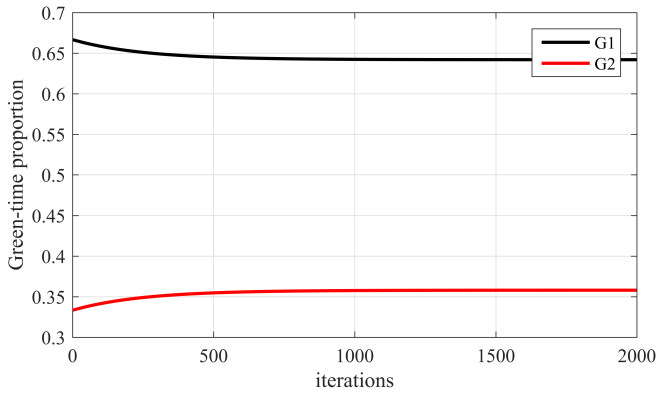

Figure 5.3: Time evolution of green-time proportion using novel control policy

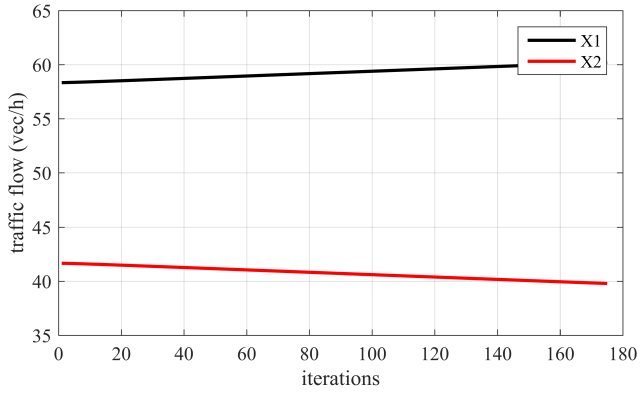

Figure 5.4: Time evolution of traffic flow using P0 control policy

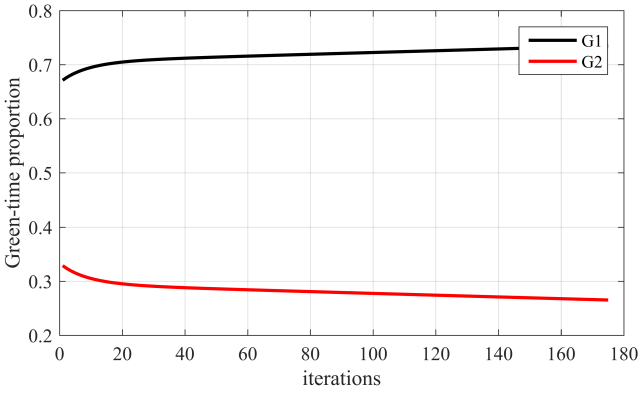

Figure 5.5: Time evolution of green-time proportion using P0 control policy

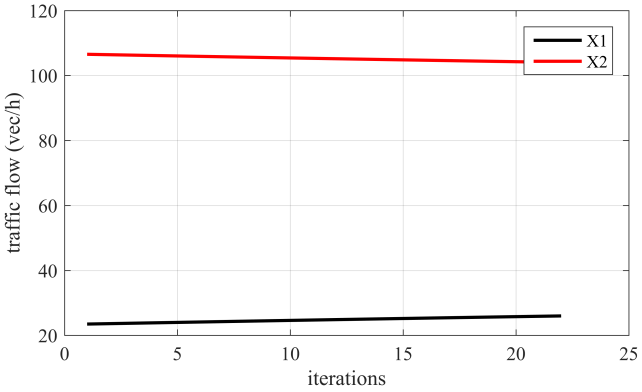

Figure 5.6: Time evolution of traffic flow using novel control policy; the condition of existence of equilibrium is NOT fulfilled

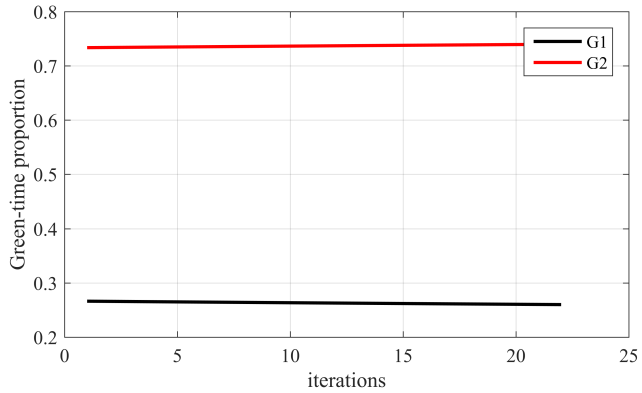

Figure 5.7: Time evolution of green-time proportion using novel control policy; the condition of existence of equilibrium is NOT fulfilled

For the two-OD two-route network, the length and saturation flow of two routes of OD1 and OD2 are set to be 1km, 2km, 1km, 2km and 100 vehicles/h, 350 vehicles/h and 100 vehicles/h, 300 vehicles/h, respectively. The turning matrix is set to be

$$P' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.8 & 0 & 0.2 & 0 \\ 0.64 & 0 & 0.16 & 0.2 \\ 0.04 & 0.8 & 0.16 & 0 \\ 0.64 & 0.64 & 0.136 & 0.16 \\ 0.2 & 0 & 0.8 & 0 \\ 0.16 & 0.2 & 0.64 & 0 \\ 0.16 & 0 & 0.04 & 0.8 \\ 0.136 & 0.16 & 0.064 & 0.64 \end{bmatrix}$$

The free speed of routes of two OD pairs are set to be 50km/h, the total demand of two OD pairs are set to be 186.75km/h and 129.43km/h, respectively. Fig. 5.8 and Fig. 5.9 illustrate the time evolution of traffic flow and green-time proportion by using the novel control policy on the two-OD two-route network. Under the condition of equilibrium existence, starting from an arbitrary feasible point, the dynamical system can approach a feasible equilibrium point after approximately 500 iterations. Fig. 5.10 and Fig. 5.11 show that the traffic flow and green-time proportion become unfeasible after approximately 100th iteration, that is, no feasible equilibrium point can be approached even with a feasible initial point if the condition of equilibrium existence is not fulfilled.

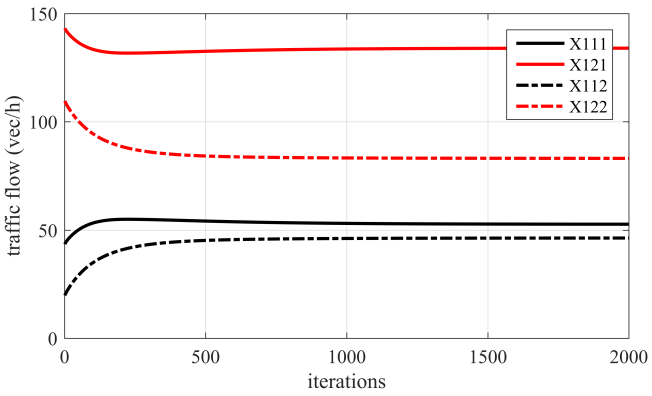

Figure 5.8: Time evolution of traffic flow using novel control policy on two-OD two-route network; the condition of existence of equilibrium is fulfilled

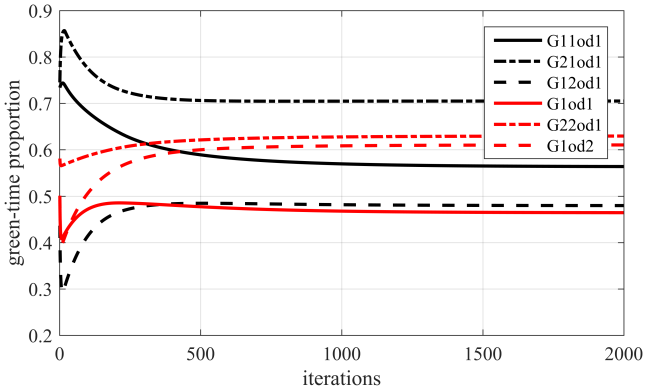

Figure 5.9: Time evolution of green-time proportion using novel control policy on two-OD two-route network; the condition of existence of equilibrium is fulfilled

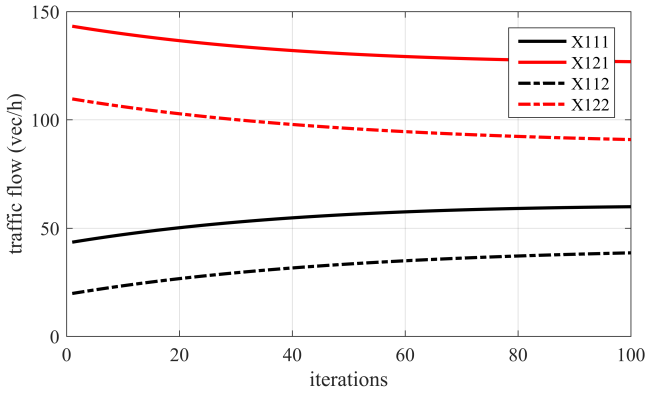

Figure 5.10: Time evolution of traffic flow using novel control policy on two-OD two-route network; the condition of existence of equilibrium is NOT fulfilled

5.5 Summary

In this chapter, a hybrid traffic assignment model incorporating the flow swap process, green-time proportion swap process and flow divergence was proposed for a general network with multiple OD pairs and multiple routes. I gave two conditions for achieving general network capacity maximization by using P0 control policy. Then, a novel control policy is proposed and its superiority over P0 control policy is proved: by using the proposed control policy, the condition of capacity maximization via intentionally constructing bottleneck delays can be relieved. I gave the condition for existence of Wardrop equilibrium for the dynamical system, which is determined by the sign of flow swap and green-time proportion swap for each element on the boundary of the set $D \times F \cap S$. The condition for uniqueness of Wardrop equilibrium was also given, that is, for a given green-time proportion vector, the travel cost of function and stage pressure function must be monotone. I also gave the characteristics of the set of

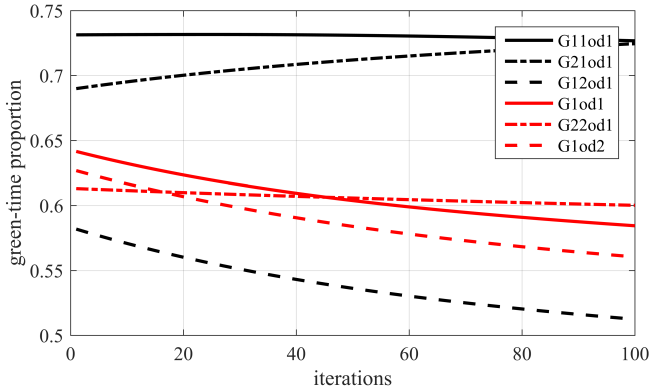

Figure 5.11: Time evolution of green-time proportion using novel control policy on two-OD two-route network; the condition of existence of equilibrium is NOT fulfilled

equilibria when green-time proportion vector changes over time, which is useful to determine the set of equilibria. Finally, Lyapunov stability analysis was utilized to prove stability of the dynamical system.

The theoretical results derived in this chapter and the proposed control policy can be applied in real scenario to increase network capacity and reduce travel time in an urban network. Specifically, when travel costs on some links go high due to incident, network capacity can be maintained at an equilibrium by applying the proposed control policy.

Chapter 6

Conclusion

This thesis is motivated by the demand to improve the traffic safety, efficiency, and alleviate traffic congestion in transportation networks. The improvement of traffic safety, efficiency and alleviation of traffic congestion strongly rely on traffic data analysis. Missing data problem, where some subsets of traffic data become missing due to certain reasons such as sensor malfunction or lossy data transmission, has negative impact on accurate results of traffic data analysis. Furthermore, existing research efforts mostly deal with developing certain data-driven or model-driven methods for data estimation. There is a lack of study on the achievable estimation accuracy and the condition to achieve accurate estimation results. Traffic data analysis provides an understanding of traffic flow on traffic network and thus is a basis of developing strategy to maximize road network capacity. Almost all existing research work focus on seeking an optimal control policy to optimize throughput by considering only one OD pair in a simple network, without consideration of the flow divergence among different OD pairs. Furthermore, no existing work investigated the conditions for the existence and uniqueness of equilibrium and stability of the dynamical systems when flow swap process, green-time proportion swap process and flow divergence among different OD pairs are simultaneously considered.

To fill the gap, this thesis focuses on answering the following questions: 1) how to improve the performance of traffic data estimation without increase of computational complexity by incorporating the topological information? 2) what is the lower bound of traffic data estimation, does the optimal estimator exist? 3) is there an optimal control policy, which can maximize traffic network capacity in a general urban network?

To improve the performance of traffic data estimation, an Optimum Closed Cut (OCC) based spatio-temporal imputation technique is proposed, which is implemented in two stages: a) employing graph theory to search OCC in the road network, for which the traffic on roads intersected by the closed cut has the maximum correlation with that on the target road while

minimizing the number of intersected roads; b) estimating the missing data on the target road using OCC based Kriging estimator, incorporating both the road topological information and flow conservation law to improve the estimation accuracy. Experimental results using traffic data provided by Sydney RMS (Road Maritime Service) indicate that the OCC searching algorithm can effectively capture the optimum set of neighboring sensors. OCC based estimator can provide more accurate estimation results compared to NHA (Nearest Historical Average) and correlative kNN (k Nearest Neighbors) methods. The road topological information and flow conservation law can be explored to further improved the estimation performance while reducing the number of sensors involved in the data estimation, hence improving the computational efficiency.

To answer the second question, I investigate the fundamental limits of missing traffic data estimation accuracy in sensor-equipped traffic networks, using the spatial-temporal random effects (STRE) model. I derive Squared Flow Error Bound (SFEB) for the cases of the Fisher matrix being a singular and non-singular matrix, respectively. I show the sufficient and necessary condition of the existence of an unbiased estimator is that the number of missing points is less than or equal to the rank of the Fisher matrix. For the case that no unbiased estimator can be found, I derive an inequality for the SFEB and show that SFEB is readily determined by the co-variance matrix of the unknown (missing) parameter vector, flow correlation between the unknown and the available data, and the sensor locations. I develop an optimal spatio-temporal Kriging estimator, which is efficient in both cases where the causal relationship among available data points exists or does not exist. The theoretical analysis is verified using real traffic data. The proposed theoretical results can be used as a benchmark to evaluate the performance of missing data estimators as the performance of any missing data estimators is lower bounded by the proposed SFEB. Corollary 4.2 can be utilized to optimize the sensor locations and the number of deployed sensors in the network. An application example is to deploy the sensors on the locations so that FIM is a full-rank matrix. Lemma 4.7 can be utilized to reduce computational complexity with a large size of the observed parameter vector. As an example, the missing points can be estimated iteratively with optimal selection of available data according to Lemma 4.7 at each iteration, rather than using the optimal Kriging estimator.

The last question deals with designing an optimal traffic control policy in a general network. I propose a hybrid dynamical system, which incorporates flow swap process, green-time proportion swap process and flow divergence for a general network. Firstly, I gave two conditions for achieving general network capacity maximization by using P0 control policy. Then, a novel control policy is proposed and its superiority over P0 control policy is proved. Furthermore, I give the condition for existence of Wardrop equilibrium for the dynamical system, which is determined by the sign of flow swap and green-time proportion swap for each ele-

ment on the boundary of the set $D \times F \cap S$. Finally, Lyapunov stability analysis was utilized to prove stability of the dynamical system.

In summary, OCC based Kriging estimator proposed in this thesis provides a strategy to better utilize the topological information and spatial correlation between the missing data points and the measured data points. It provides an optimum trade-off between estimation accuracy and computational complexity. The investigation of fundamental limits of data estimation provides a benchmark for missing traffic data estimator, the derived sufficient and necessary condition can be used to determine the existence of an unbiased estimator. The proposed optimal traffic control policy in a general network provides a new signal control strategy with consideration of traffic assignment model.

In this thesis, optimal traffic control policy is only investigated for the case with inelastic travel demand, vertical queue, and no spatial queue and blocking back are considered. Therefore, elastic travel demand, spatial queue and blocking back should be considered in the future work. Besides, this thesis only investigated the traditional traffic network, that is, no connected and autonomous vehicles. In the future work, the research can be extended to next generation traffic network with connected and autonomous vehicles.

References

[1] Noraini Sarina Abdullah and Ting Kien Hua. Using ford-fulkerson algorithm and max flow- min cut theorem to minimize traffic congestion in kota kinabalu, sabah. 2017.

[2] H. Asai, K. Fukuda, and H. Esaki. Traffic causality graphs: Profiling network applications through temporal and spatial causality of flows. In *Teletraffic Congress (ITC), 2011 23rd International*, pages 95–102, Sept 2011.

[3] M. T. Asif, N. Mitrovic, J. Dauwels, and P. Jaillet. Matrix and tensor based methods for missing data estimation in large traffic networks. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):1816–1825, July 2016. ISSN 1524-9050. doi: 10.1109/TITS.2015.2507259.

[4] M.J. Beckmann, C.B. McGuire, and C.B. Winsten. *Studies in the Economics of Transportation*. Yale University Press, 1956. https://books.google.com.au/books?id=3CVPAAAAMAAJ.

[5] Michael G.H. Bell, Fumitaka Kurauchi, Supun Perera, and Walter Wong. Investigating transport network vulnerability by capacity weighted spectral analysis. *Transportation Research Part B: Methodological*, 99:251 – 266, 2017. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2017.03.002. http://www.sciencedirect.com/science/article/pii/S0191261516307159.

[6] Population Reference Bureau. 2016 world population datasheet, inform empower advance. 2016. Available online: http://www.prb.org/pdf16/prb-wpds2016-web-2016.pdf (accessed on 11 October 2017).

[7] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. del Castillo. Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems*, 13(3):1430–1441, Sept 2012. ISSN 1524-9050. doi: 10.1109/TITS.2012.2189006.

[8] Pinlong Cai, Yunpeng Wang, Guangquan Lu, Peng Chen, Chuan Ding, and Jianping Sun. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep

forecasting. *Transportation Research Part C: Emerging Technologies*, 62:21 – 34, 2016. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2015.11.002.

[9] E. Castillo, P. Jimenez, J. M. Menendez, and A. J. Conejo. The observability problem in traffic models: Algebraic and topological methods. *IEEE Transactions on Intelligent Transportation Systems*, 9(2):275–287, June 2008. ISSN 1524-9050. doi: 10.1109/TITS.2008.922929.

[10] G. Chang, Y. Zhang, and D. Yao. Missing data imputation for traffic flow based on improved local least squares. *Tsinghua Science and Technology*, 17(3):304–309, June 2012. doi: 10.1109/TST.2012.6216760.

[11] Anthony Chen and Panatda Kasikitwiwat. Modeling capacity flexibility of transportation networks. *Transportation Research Part A: Policy and Practice*, 45(2): 105 – 117, 2011. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2010.11.003. http://www.sciencedirect.com/science/article/pii/S0965856410001576.

[12] Chao Chen, Jaimyoung Kwon, John Rice, Alexander Skabardonis, and Pravin Varaiya. Detecting errors and imputing missing data for single loop surveillance systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1855:160–167, 2003. doi: 10.3141/1855-20. http://dx.doi.org/10.3141/1855-20.

[13] Suh-Wen Chiou. Maximizing reserve capacity for a signalized road network design problem. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 1090–1095, Sept 2006. doi: 10.1109/ITSC.2006.1707367.

[14] European Commission. Roadmap to a single european transport area - towards a competitive and resource efficient transport system. 2011.

[15] J.H. Conklin. *Data Imputation Strategies for Transportation Management Systems*. Research report (University of Virginia. Center for Transportation Studies). University of Virginia, 2003. https://books.google.com.au/books?id=yTsJIAAACAAJ.

[16] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, 2011. http://www.bibsonomy.org/bibtex/27eb00d2c02d4494513681956dae0825b/roman.minguez.

[17] N. Cressie and C.K. Wikle. *Statistics for Spatio-Temporal Data*. CourseSmart Series. Wiley, 2011. ISBN 9780471692744. https://books.google.com.au/books?id=-kOC6D0DiNYC.

[18] Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2007.00633.x. http://dx.doi.org/10.1111/j.1467-9868.2007.00633.x.

[19] Noel Cressie, Tao Shi, and Emily L. Kang. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745, 2010. doi: 10.1198/jcgs.2010.09051. http://dx.doi.org/10.1198/jcgs.2010.09051.

[20] Reinhard Diestel. *Graph Theory (Graduate Texts in Mathematics)*. Springer, August 2005. ISBN 3540261826. http://www.bibsonomy.org/bibtex/28e6027f87536254e3d0cadc76403c9b5/msn.

[21] H. Dong, L. Jia, X. Sun, C. Li, and Y. Qin. Road traffic flow prediction with a time-oriented arima model. In *2009 Fifth International Joint Conference on INC, IMS and IDC*, pages 1649–1652, Aug 2009. doi: 10.1109/NCM.2009.224.

[22] Alireza Ermagun, Snigdhansu Chatterjee, and David Levinson. Using temporal detrending to observe the spatial correlation of traffic. *PLOS ONE*, 12(5):1–21, 05 2017. doi: 10.1371/journal.pone.0176853. https://doi.org/10.1371/journal.pone.0176853.

[23] Liping Fu and Xuhui Yang. Design and implementation of bus holding control strategies with real-time information. 2002.

[24] Texas AM Transportation Institute. Technical report 2015 urban mobility scorecard, inrix. Available online: https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-scorecard-2015.pdf (accessed on 11 October 2017).

[25] Yiannis Kamarianakis and Poulicos Prastacos. *Spatial time-series modeling: A review of the proposed methodologies*. 2005.

[26] Emily L. Kang and Noel Cressie. Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association*, 106(495):972–983, 2011. doi: 10.1198/jasa.2011.tm09680. http://dx.doi.org/10.1198/jasa.2011.tm09680.

[27] M.G. Karlaftis and E.I. Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387 – 399, 2011. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2010.10.004.

[28] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. ISBN 0-13-345711-7.

[29] AndrÃ© Klein and Guy MÃ©lard. Computation of the fisher information matrix for time series models. *Journal of Computational and Applied Mathematics*, 64(1):57 – 68, 1995. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(95)00006-2. http://www.sciencedirect.com/science/article/pii/0377042795000062.

[30] Abdullah Kurkcu, Ender Faruk Morgul, and Kaan Ozbay. Extended implementation method for virtual sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2528:27–37, 2015. doi: 10.3141/2528-04. https://doi.org/10.3141/2528-04.

[31] Tung Le, PÃ©ter KovÃ¡cs, Neil Walton, Hai L. Vu, Lachlan L.H. Andrew, and Serge S.P. Hoogendoorn. Decentralized signal control for urban road networks. *Transportation Research Part C: Emerging Technologies*, 58:431 – 450, 2015. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2014.11.009. Special Issue: Advanced Road Traffic Control.

[32] Li Li, Yuebiao Li, and Zhiheng Li. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, 34:108 – 120, 2013. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2013.05.008.

[33] Chen Liang, Wu Lenan, and R. Piche. Posterior cramer-rao lower bound for mobile tracking in mixed los/nlos conditions. In *2009 17th European Signal Processing Conference*, pages 90–94, Aug 2009.

[34] M. J. Lighthill and G. B. Whitham. On kinematic waves. ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 229(1178):317–345, 1955. ISSN 0080-4630. doi: 10.1098/rspa.1955.0089. http://rspa.royalsocietypublishing.org/content/229/1178/317.

[35] Henry X. Liu, Xiaozheng He, and Will Recker. Estimation of the time-dependency of values of travel time and its reliability from loop detector data. *Transportation Research Part B: Methodological*, 41(4):448 – 461, 2007. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2006.07.002. http://www.sciencedirect.com/science/article/pii/S0191261506000890.

[36] Richard Mounce and Malachy Carey. Route swapping in dynamic traffic networks. *Transportation Research Part B: Methodological*, 45(1):102 –

111, 2011. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2010.05.005. http://www.sciencedirect.com/science/article/pii/S0191261510000809.

[37] ManWo Ng. Synergistic sensor location for link flow inference without path enumeration: A node-based approach. *Transportation Research Part B: Methodological*, 46(6): 781 – 788, 2012. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2012.02.001.

[38] Yu (Marco) Nie. A class of bush-based algorithms for the traffic assignment problem. *Transportation Research Part B: Methodological*, 44(1):73 – 89, 2010. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2009.06.005. http://www.sciencedirect.com/science/article/pii/S0191261509000769.

[39] L. Qu, L. Li, Y. Zhang, and J. Hu. Ppca based missing data imputation for traffic flow volume: A systematical approach. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):512–522, Sept 2009. ISSN 1524-9050. doi: 10.1109/TITS.2009.2026312.

[40] M. Rashid and R. M. Dansereau. Cramer-rao lower bound derivation and performance analysis for space-based sar gmti. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6031–6043, Nov 2017. ISSN 0196-2892. doi: 10.1109/TGRS.2017.2719401.

[41] Matthew S. Ryan and Graham R. Nudd. The viterbi algorithm. Technical report, Coventry, UK, UK, 1993.

[42] Y. Sheffi. *Urban Transportation Networks: Equilibrium Analysis With Mathematical Programming Methods*. Prentice-Hall, 1984. ISBN 9780139397295. https://books.google.com.au/books?id=zx1PAAAAMAAJ.

[43] Y. Shen and M. Z. Win. Fundamental limits of wideband localization 2014; part i: A general framework. *IEEE Transactions on Information Theory*, 56(10):4956–4980, Oct 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2060110.

[44] Y. Shen, H. Wymeersch, and M. Z. Win. Fundamental limits of wideband localization x2014; part ii: Cooperative networks. *IEEE Transactions on Information Theory*, 56 (10):4981–5000, Oct 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2059720.

[45] Brian Smith, Professor, Chen Wu, Lu Smith, Lu, and Smith. Traffic flow prediction for urban network using spatio-temporal random. 2011.

[46] M. J. Smith. A descent algorithm for solving monotone variational inequalities and monotone complementarity problems. *Journal of Optimization Theory and Applications*, 44(3):485–496, Nov 1984. ISSN 1573-2878. doi: 10.1007/BF00935463. https://doi.org/10.1007/BF00935463.

[47] Michael J. Smith. The stability of a dynamic model of traffic assignment an application of a method of lyapunov. *Transportation Science*, 18(3):245–252, 1984. doi: 10.1287/trsc.18.3.245. https://doi.org/10.1287/trsc.18.3.245.

[48] Mike Smith. Dynamics of route choice and signal control in capacitated networks. *Journal of Choice Modelling*, 4(3):30 – 51, 2011. ISSN 1755-5345. doi: https://doi.org/10.1016/S1755-5345(13)70041-1. http://www.sciencedirect.com/science/article/pii/S1755534513700411.

[49] Mike Smith. Traffic signal control and route choice: A new assignment and control model which designs signal timings. *Transportation Research Part C: Emerging Technologies*, 58:451 – 473, 2015. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2015.02.002. Special Issue: Advanced Road Traffic Control.

[50] M.J. Smith. The existence, uniqueness and stability of traffic equilibria. *Transportation Research Part B: Methodological*, 13(4):295 – 304, 1979. ISSN 0191-2615. doi: https://doi.org/10.1016/0191-2615(79)90022-5. http://www.sciencedirect.com/science/article/pii/0191261579900225.

[51] M.J. Smith, R. Liu, and R. Mounce. Traffic control and route choice: Capacity maximisation and stability. *Transportation Research Part B: Methodological*, 81:863 – 885, 2015. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2015.07.002. ISTTT 21 for the year 2015.

[52] F. Soriguera and F. RobustÃ©. Estimation of traffic stream space mean speed from time aggregations of double loop detector data. *Transportation Research Part C: Emerging Technologies*, 19(1):115 – 129, 2011. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2010.04.004. http://www.sciencedirect.com/science/article/pii/S0968090X10000604.

[53] Mechthild Stoer and Frank Wagner. A simple min-cut algorithm. *J. ACM*, 44(4):585–591, July 1997. ISSN 0004-5411. doi: 10.1145/263867.263872. http://doi.acm.org/10.1145/263867.263872.

[54] P. Stoica and T. L. Marzetta. Parameter estimation problems with singular information matrices. *IEEE Transactions on Signal Processing*, 49(1):87–90, Jan 2001. ISSN 1053-587X. doi: 10.1109/78.890346.

[55] Agachai Sumalee and Hung Wai Ho. Smarter and more connected: Future intelligent transportation system. *IATSS Research*, 42(2):67 – 71, 2018. ISSN 0386-1112. doi: https://doi.org/10.1016/j.iatssr.2018.05.005. http://www.sciencedirect.com/science/article/pii/S0386111218300396.

[56] S. Tak, S. Woo, and H. Yeo. Data-driven imputation method for traffic data in sectional units of road links. *IEEE Transactions on Intelligent Transportation Systems*, 17(6): 1762–1771, June 2016. ISSN 1524-9050. doi: 10.1109/TITS.2016.2530312.

[57] Huachun Tan, Guangdong Feng, Jianshuai Feng, Wuhong Wang, Yu-Jin Zhang, and Feng Li. A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies*, 28:15 – 27, 2013. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2012.12.007. Euro Transportation: selected paper from the {EWGT} Meeting, Padova, September 2009.

[58] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multi hop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, Dec 1992. ISSN 0018-9286. doi: 10.1109/9.182479.

[59] United Nations Population Fund (UNFPA). Technical report state of world population 2011: People and possibilities in a world of 7 billion. United Nations Population Fund: New York, NY, USA, 2011.

[60] J.W.C. van Lint, S.P. Hoogendoorn, and H.J. van Zuylen. Accurate freeway travel time prediction with state space neural networks under missing data. *Transportation Research Part C: Emerging Technologies*, 13(5):347 – 369, 2005. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2005.03.001.

[61] H.L. Van Trees, K.L. Bell, and Z. Tian. *Detection Estimation and Modulation Theory, Part I: Detection, Estimation, and Filtering Theory*. Detection Estimation and Modulation Theory. Wiley, 2013. ISBN 9781118539705. https://books.google.com.au/books?id=dnvaxqHDkbQC.

[62] R.A. Vincent, A.I. Mitchell, D.I. Robertson, Transport, and Road Research Laboratory. *User Guide to TRANSYT Version 8*. TRRL laboratory report. Transport and Road Research Laboratory, 1980. https://books.google.com.au/books?id=UuGDGAAACAAJ.

[63] Francesco Viti, Marco Rinaldi, Francesco Corman, and Chris M.J. TampÃšre. Assessing partial observability in network sensor location problems. *Transportation Research Part B: Methodological*, 70:65 – 89, 2014. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2014.08.002.

[64] Eleni I. Vlahogianni, John C. Golias, and Matthew G. Karlaftis. Short term traffic forecasting: Overview of objectives and methods. *Transport Reviews*, 24(5):533–557, 2004. doi: 10.1080/0144164042000195072. http://dx.doi.org/10.1080/0144164042000195072.

[65] Yibing Wang, Markos Papageorgiou, and Albert Messmer. Real-time freeway traffic state estimation based on extended kalman filter: A case study. *Transportation Science*, 41(2):167–181, 2007. doi: 10.1287/trsc.1070.0194. http://pubsonline.informs.org/doi/abs/10.1287/trsc.1070.0194.

[66] F.V. Webster. *Traffic Signal Settings*. Road research technical paper. H.M. Stationery Office, 1958. https://books.google.com.au/books?id=c9QOQ4jXK5cC.

[67] Billy M. Williams and Lester A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, 129(6):664–672, 2003. doi: 10.1061/(ASCE)0733-947X(2003)129:6(664). http://dx.doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664).

[68] S.C. Wong and Hai Yang. Reserve capacity of a signal-controlled road network. *Transportation Research Part B: Methodological*, 31(5):397 – 402, 1997. ISSN 0191-2615. doi: https://doi.org/10.1016/S0191-2615(97)00002-7. http://www.sciencedirect.com/science/article/pii/S0191261597000027.

[69] Xinkai Wu and Henry X. Liu. A shockwave profile model for traffic flow on congested urban arterials. *Transportation Research Part B: Methodological*, 45(10):1768 – 1786, 2011. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2011.07.013. http://www.sciencedirect.com/science/article/pii/S0191261511001135.

[70] Yao-Jan Wu, Feng Chen, Chang-Tien Lu, and Shu Yang. Urban traffic flow prediction using a spatio-temporal random effects model. *Journal of Intelligent Transportation Systems*, 20(3):282–293, 2016. doi: 10.1080/15472450.2015.1072050. http://dx.doi.org/10.1080/15472450.2015.1072050.

[71] Lin Xiao and Hong K. Lo. Combined route choice and adaptive traffic control in a day-to-day dynamical system. *Networks and Spatial Economics*, 15(3):697–717, Sep 2015.

ISSN 1572-9427. doi: 10.1007/s11067-014-9248-4. https://doi.org/10.1007/s11067-014-9248-4.

[72] Zhixian Yan. Traj-arima: A spatial-time series model for network-constrained trajectory. In *Proceedings of the Second International Workshop on Computational Transportation Science*, IWCTS '10, pages 11–16, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0429-0. doi: 10.1145/1899441.1899446. http://doi.acm.org/10.1145/1899441.1899446.

[73] Ching-Fei Yang, You-Huei Ju, Chung-Ying Hsieh, Chia-Ying Lin, Meng-Hsun Tsai, and Hui-Ling Chang. iparking: a real time parking space monitoring and guiding system. *Vehicular Communications*, 9:301 – 305, 2017. ISSN 2214-2096. doi: https://doi.org/10.1016/j.vehcom.2017.04.001. http://www.sciencedirect.com/science/article/pii/S2214209616301930.

[74] Hai Yang, Michael G.H. Bell, and Qiang Meng. Modeling the capacity and level of service of urban transportation networks. *Transportation Research Part B: Methodological*, 34(4):255 – 275, 2000. ISSN 0191-2615. doi: https://doi.org/10.1016/S0191-2615(99)00024-7. http://www.sciencedirect.com/science/article/pii/S0191261599000247.

[75] Su Yang, Shixiong Shi, Xiaobing Hu, and Minjie Wang. Spatiotemporal context awareness for urban traffic modeling and prediction: Sparse representation based variable selection. *PLOS ONE*, 10(10):1–22, 10 2015. doi: 10.1371/journal.pone.0141223. https://doi.org/10.1371/journal.pone.0141223.

[76] Kegen Yu and Eryk Dutkiewicz. Crlb derivation for mobile tracking in nlos propagation environments, 10 2012.

[77] Y. Zhang and Y. Liu. Data imputation using least squares support vector machines in urban arterial streets. *IEEE Signal Processing Letters*, 16(5):414–417, May 2009. ISSN 1070-9908. doi: 10.1109/LSP.2009.2016451.

[78] Y. Zhang and Y. Liu. Missing traffic flow data prediction using least squares support vector machines in urban arterial streets. In *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, pages 76–83, March 2009. doi: 10.1109/CIDM.2009.4938632.

[79] Ming Zhong, Pawan Lingras, and Satish Sharma. Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. *Transportation Research Part*

C: Emerging Technologies, 12(2):139 – 166, 2004. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2004.07.006.

[80] P. Zhou, Y. Zheng, and M. Li. How long to wait? predicting bus arrival time with mobile phone based participatory sensing. *IEEE Transactions on Mobile Computing*, 13(6): 1228–1241, June 2014. ISSN 1536-1233. doi: 10.1109/TMC.2013.136.